NOVA IMS
Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação
Master Program in Statistics and Information Management

LOAN MODIFICATIONS AND RISK OF DEFAULT: A MARKOV CHAINS APPROACH

Filipa Cardoso de Almeida

Dissertation presented as the partial requirement for obtaining a Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

LOAN MODIFICATIONS AND RISK OF DEFAULT:

A MARKOV CHAINS APPROACH

by

Filipa Cardoso de Almeida

Dissertation presented as the partial requirement for obtaining a Master's degree in Statistics and Information Management, Specialization in Risk Analysis and Management

Advisor: Doctor Bruno Miguel Pinto Damásio

March 2020

ACKNOWLEDGMENTS

The realization of this dissertation would not be possible without the help and support of several people, to whom I would like to leave my most sincere gratitude.

First of all, I would like to thank my advisor, Bruno Damásio, for all the support and accompaniment throughout the development of this work, for all the advice and for always taking me further. A special thanks to Carolina Vasconcelos for the help given in the development of this thesis.

I would also like to thank my friends that accompanied me the most in this journey, Bruna, Luis Miguel, Joana, and Bigo, for all the hours we spent doing work together, studying and also decompressing from all these responsibilities.

A huge appreciation to all my friends who always understood my more considerable absence during the development of this thesis, Inês, Sara, Diogo and Sofia, who listened to me, gave me advice and strengthened me so that this phase of my life resulted with the greatest success.

Last but not least, my most enormous gratitude to my family, for always being there and giving me fundamental support so that I never give up. To my mother, Paula, my sister Beatriz, my aunt and uncle, Carla and Raúl, my grandparents, Conceição and António and my boyfriend, Fábio, thank you, without you none of this would have been possible.

This dissertation marks the end of a phase in my life that, without all of you, would not have been possible to achieve. For this, I leave here my most sincere thanks to all.

ABSTRACT

With the housing crisis, credit risk analysis has had an exponentially increasing importance, since it is a key tool for banks' credit risk management, as well as being of great relevance for rigorous regulation. Credit scoring models that rely on logistic regression have been the most widely applied to evaluate credit risk, more specifically to analyze the probability of default of a borrower when a credit contract initiates. However, these methods have some limitations, such as the inability to model the entire probabilistic structure of a process, namely, the life of a mortgage, since they essentially focus on binary outcomes. Thus, there is a weakness regarding the analysis and characterization of the behavior of borrowers over time and, consequently, a disregard of the multiple loan outcomes and the various transitions a borrower may face. Therefore, it hampers the understanding of the recurrence of risk events. A discrete-time Markov chain model is applied in order to overcome these limitations. Several states and transitions are considered with the purpose of perceiving a borrower's behavior and estimating his default risk before and after some modifications are made, along with the determinants of post-modification mortgage outcomes. Mortgages loans are considered in order to take a reasonable timeline towards a proper assessment of different loan performances. In addition to analyzing the impact of modifications, this work aims to identify and evaluate the main risk factors among borrowers that justify transitions to default states and different loan outcomes.

KEYWORDS

Loan Modification; Default; Markov Chains; Self-Organizing Maps; Credit Risk

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
BMU	Best Matching Unit
CART	Classification and Regression Trees
DT	Decision Tree
DTI	Debt to Income
HAMP	Home Affordable Modification Program
HARP	Home Affordable Refinancing Program
HOMC	Higher-Order Markov Chain
KNN	K-Nearest Neighbor
LTV	Loan-to-Value
MARS	Multivariate Adaptive Regression Splines
MCM	Markov Chain Model
MMC	Multivariate Markov Chain
REO	Real Estate Owned
SOM	Self-Organizing Map
SVM	Support Vector Machine
TARP	Troubled Asset Relief Program
TPM	Transition Probability Matrix

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Since the period of the global financial crisis that began in 2007, we have been observing an upward concern about credit risk management, as a deficit loan management was one of the central pillars of this crisis. The global financial crisis has brought attention to several areas where there was a need for the improvement of many features, especially in banking management regulation and credit risk management.

Historically, loan modifications were not a very common solution, something that changed with the financial crisis, since during this period there was a considerable increase in the inability of borrowers to repay their loans – in the period that preceded the financial crisis, in the United States, the percentage of non-performing loans was around 1.6%, while at the peak of this financial crisis (around 2010), this percentage increased to about 7.5% (CEIC, 2011). Therefore, loan modifications became a viable solution for banks not to lose the full amount of mortgages, thus promoting repayments by distressed borrowers. Since it is not a recent reality, the studies on loan modifications' effects are few, making this a significant problem. Furthermore, there continues to exist a large proportion of investigations solely related to the probability of default, essentially a binary analysis, using traditional credit scoring models to attribute classifications on the credit risk level of clients and using linear regression models. However, many issues, as so or even more critical, are fundamental when we discuss credit risk management. These concerns are, namely:

- To understand why a customer moves from one state to another;
- To assess the probability of a customer going from a state, for example, of delinquency to a normal state, overcoming the difficulty of fulfilling their obligations on credit;
- To recognize the best credit conditions for both the bank and the customer, so that the bank guarantees the receipt of all agreed payments and also so that the conditions are not too rigid for the customer;
- To notice, if there are modifications at a given moment of the credit, which are the most effective, so that the bank can adapt to such in future contracts and the assessment of the probability of redefault.

The traditional approaches are limited methodologies regarding the determination of transition probabilities. This drawback happens because most methodologies employed do not consider several states and the possible transitions between them. Then, there is a greater concern in determining the risk of a customer entering into default when contracting a loan, depending solely on the customer's profile, traced through its history, that is always required at the time of the contract. Consequently, by discoursing the existence of such states and their possible transitions, there is no concern focused on determining the risk of redefault or even which states influence the transition to that state. The continuity of a loan contract after modifications, triggered by a default event, is a reality. Thus, it is necessary to develop a study focused on the probability of redefault, since the customer's risk profile changes.

A client may reside in more states over the life of a loan than just default. Furthermore, even after a client enters into default, the loan may last after suffering some modifications. In order to properly assess a client's behavior after loan modifications and thus perform a proper credit risk management, it is necessary to consider the proceeding of loans. Even though various methods, such as linear regression, allow to model several states of categorical variables, these can not capture the entire probabilistic structure of the process in the case where the mean is not linear. Hence, a discrete-time Markov chain is considered. Several states and transitions between those states are taken into account, describing all the possible situations in which a loan can be in, allowing the establishment of credit risk based on behavioral aspects. This way, we are able to incorporate historical information in the assessment of the probability of redefault, capturing the occurrence of all states occupied by a client.

1.2. STUDY OBJECTIVES

The main goal of this study is to evaluate the impact of the modifications in loans in order to determine if those modification are effective, i.e., if they reduce the probability of a borrower to default. This assessment becomes essential as a credit risk management tool because it will allow us to determine whether these changes are, in fact, able to be used to mitigate risk in financial institutions, therefore having a considerable impact in the banking business, as well as a significant influence on the life of borrowers. This ambition requires an analysis of the variables influencing several groups of clients, with similar characteristics. Those groups will be constructed through cluster, using Self-Organizing Maps (SOMs). With this, we will then be able to compare groups of clients, instead of comparing individual clients.

Considering the goal of this work, we will focus on the loans' performance data. We will use a discrete-time Markov chain model (MCM) to estimate the probabilities of transitioning from one state to another, relying on the Transition Probability Matrix (TPM) to calculate the probability of default. Following this rationale, we can also evaluate how the modifications of the various states of a loan influence the probability of default considering the history of clients.

Furthermore, in order to achieve the main objective of this dissertation, we aim to compare the probability of default before the modifications and the probability of default after the modifications. For the first group of loans, we will limit our data to loans that never defaulted and estimate their corresponding TPM. For the second group of loans, we will solely consider loans which conditions were modified and estimate their corresponding TPM. Therefore, we can infer if the modifications are effective. Additionally, we will evaluate the impact that different loan performances have in the probabilities estimated as well as link the different risk factors to the different performances of individuals.

1.3. Study Relevance and Importance

Credit risk models are a critical tool in risk management and credit risk management is one of the major concerns of banks, since loans are one of the main bank's products. Therefore, the development of several studies on the best method of credit risk assessment is reasonable, as well as a growing attempt to find a method that fits better than the existing ones.

Most investigations are mainly concern with credit risk modelling in several types of loans such as consumer loans, credit card loans and mortgages. The most common credit-scoring methodology, the logistic regression model does not allow an evaluation of the client's behavior between non-default and default states. Hence, it is possible to infer that this model only concerns with the transition to a default state from a non-default state. Additionally, multinomial methodologies are other comprehensively applied strategies in which several states are predominantly considered. However, they impose a functional form, only modeling the conditional mean. This limitation started to be widely recognized, so the adoption of other methodologies was found to be necessary. The MCM has proven to be a very efficient solution when it comes to evaluating the behavior of a client over the life of a loan. Some examples of studies using Markov Chains are those of Régis, D. E. & Artes, 2016, to identify credit card risk and Leow, M. & Crook, J. 2014, in which the MCM was used to predict the probability of credit card default to estimate the probability of delinquency and default for credit cards. Betancourt, L. 1999, applied MCM to estimate losses from a portfolio of mortgages, and therefore, estimate the accuracy of Markov chain models on mortgage loan losses. Chamboko, R. & Bravo, J. M 2016, utilized MCM as a tool to model transition probabilities between the various states and estimating the probability of loans transitioning to and from various loan outcomes and acquisition and performance explanatory variables. Malik, M. & Thomas, L.C. 2012, among others, tested MCM to estimate consumer credit ratings and to model retail credit risk.

As we can see, using Markov chains model is advantageous when we want to describe the dynamics of credit risk, since it focusses on transition probabilities between different states. Consequently, this methodology is very valuable to model credit risk, emphasizing the use of transition probabilities to determine the probability of default. However, these studies do not extend to the probability of redefault, which becomes a reality and a source of alarm since loan modifications are a recent solution in credit risk management. For this reason, the development of this study is critical to fulfilling this gap.

The discrete-time Markov chain model circumvents the use of simplified approaches, considering the states and transitions that occur during the lifetime of a loan. In this research, we will only consider long-term loan data, so that sufficient time for analysis is provided. In a simplified approach, just two states and one type of transition are considered – states of default and non-default and the transitions to one another. Several states, such as delinquency, recovery, short sale, prepayment, among others, are considered in the discrete-time Markov model. Subsequently, we can calculate transitions and respective probabilities between all states.

The great advantage of this methodology is that we are able to model the entire probabilistic structure of a process, capturing complex and less noticeable relationships. Moreover, unlike traditional parametric methods, not only nonlinearities at the conditional mean, but also higher moments of the distribution of a process are taken into account.

2. LITERATURE REVIEW

Credit is one of the main products of banks, so it is necessary to implement extremely rigorous and careful management to be closer to the needs of banks, as well as the needs of clients. It is fundamental to model credit risk so that banks are capable of adapting to some eventual modifications that occur over time, especially in the event that clients do not comply with their credit obligations. This need is clearly recognized, and therefore credit risk models have been used for a long time, being a subject of research over the last 50 years. As a result of credit risk models being an important matter of study, several models apply not only to credit risk in mortgages but also to consumer credit risk, credit cards, among others. Consequently, several credit risk models were developed.

As previously stated, loan modifications started to become more common with the 2007 financial crisis. Throughout this crisis, it is possible to identify various socioeconomic aspects that have markedly influenced some behaviors. This global crisis has led to a deterioration of economic conditions, particularly a large increase in unemployment, which is one of the factors that most affects an individual's capacity to cope with his expenses, particularly when borrowing, following income, cashflow and liquidity shocks. There are also other social factors such as divorces or other casualties from the social forum, that lead to health degradations which, in turn, can lead to more serious situations that have a negative consequence when these individuals have commitments with banks. Nevertheless, numerous aspects beyond the conditions of the loans themselves, should also be carefully analyzed. These elements include the rigidity of mortgage contracts, some default trigger events such as high house prices, interest rates and borrowers' credit history, which are essential to consider.

Additionally, there are other concepts that we must include when we analyze the course of loans, such as strategic default, that happens when a borrower chooses to default despite having enough monetary funds to continue to pay the mortgage. Also, the incorporation of several states, like "delinquency" – a pre-default state –, "foreclosure" – the bank takes control of a property, expels the householder and sells the home after the householder is incapable fulfilling his mortgage as specified in the mortgage contract – and "cure" – recovery from delinquency state. This complexity of characteristics worthy of analysis converts into the necessity to find the methods that best fulfill this objective. Accordingly, we can observe the development of several different studies and methodologies capable of supporting such complexity.

Altman (1968) proposed a discriminant analysis to determine combinations of observable characteristics, i.e., the contribution of each explanatory variable, to assess the probability of default. This credit scoring model is widely recognized as the "Z-Score Model," which uses financial ratios to predict corporate bankruptcy by attributing a Z-Score[1] to an obligor. The author concluded that companies with a Z-Score below 1.81 go bankrupt, while companies with Z-Scores above 2.99 do not fall into bankruptcy. For Z-Scores between those two values, it is considered a "zone of ignorance," where we cannot accurately determine whether the company falls into bankruptcy. Following this proposed methodology, many models based on credit scoring appeared with important contributions

[1] a Z-Score is the number of standard deviations from the mean a data point is.

to credit risk analysis. The most well-known models in this area, therefore being the most applied when it comes to credit scoring, are the Regression Models. Within Regression Models, we have Logit Models (Martin, 1977) and Probit Models (Ohlson, 1980). As previously mentioned, these models are quite adequate in credit scoring. However, these methodologies only focus on the probability of default in order to perform credit scoring studies, being simple probabilistic formulas for classification and, therefore, not capable to accurately deal with nonlinear effects of explanatory variables.

In the meantime, other models that fulfill more complex needs emerged, such as machine learning models. An example is the machine learning classification technique named K-Nearest Neighbors (KNN). Chatterjee and Barcun (1970) first applied the nearest neighbor to credit risk evaluation. Years later, Henley and Hand (1997) considered the development of a credit scoring system using KNN methods. These strategies are non-parametric, whose algorithm analyzes patterns of the k-nearest observations that are most identical to a new observation. KNN methods can be applied to classification and regression predictive problems, frequently being employed for its easy interpretation and short calculation time. Other simple and easily understandable models, such as Decision Trees (DT), can also be applied as credit risk models. A decision tree consists of a non-parametric approach of nodes and edges, mapping observations of an individual to make conclusions about the individual's class. It is constructed automatically by a specific training algorithm employed on a given training dataset. DT models are frequently used together with other methods, such as a rule draw, to interpret some complex models, like artificial neural networks (ANN). ANN are computational methods that replicate the human brain's way to process information so that one may identify the client's characteristics that, in the credit scoring area, are related to the default event, enabling to determine which characteristics influence the different types of clients. An example of this combination between DT and ANN is the use of DT to visualize the credit evaluation knowledge extracted from neural network on a credit dataset, by Baesens et al. (2003) and Mues et al. (2006).

Classification and Regression Trees (CART), defined as a decision tree graphic that classifies a dataset into a finite number of classes, is a methodology used in credit risk as well. Furthermore, there are Hybrid Methods (Zhang et al. 2008) that combine one or more methods, as in the case with DT and ANN. Several experiments have demonstrated that using two or more single models can generate more accurate results by overcoming weaknesses and assumptions of a single method and therefore produce a more robust forecasting system. As was recognized, there are great benefits in using hybrid methods. Two typical hybrid methodologies commonly used in credit risk, such as Support Vector Machines (SVM), an optimization method, and machine learning procedure, which was first proposed by Vapnik (1995), has the minimization of the upper bound of the generalization error as its main idea. Freidman (1991) proposed a non-linear parametric regression known as Multivariate Adaptive Regression Splines (MARS).

As mentioned previously, studies on credit risk management models fostered the exploration of the most appropriate model and also an examination of which model best overcomes the limitations of other methodologies. Furthermore, following the financial crisis that began in 2007, we have been observing an increasing number of researchers dedicated to analyzing several factors as determinants for events such as default and foreclosure. This circumstance is also why hybrid models emerged. In the last few years a new type of methodology, Survival Analysis, was developed. Survival models have been pointed out as preferable to other models due to their ability to incorporate variations in the credit over time that affect performance on the loan payment and the ability to forecast the occurrence of events (default, recovery, prepayment, foreclosure). Survival models have been frequently used to model the risk of default (Bellotti and Crook 2013; Noh et al. 2005; Sarlija et al. 2009; Tong et al. 2012), but also to model foreclosure on mortgages (Gerardi et al. 2008) and to model recovery from delinquency to current or normal performance (Ha 2010; Ho Ha and Krishnan 2012; Chamboko and Bravo 2016).

Although all these models have their advantages and disadvantages, it is essential to mention that the vast majority consist of two-state credit risk models. The problem with such models is that they tend to disregard the transition behavior between other states beyond just default and no default. Accordingly, multi-state models emerged, providing an answer to this limitation. Multi-state models (Hougaard, P. 1999) are models for a process, such as describing the life history of an individual, which at any time occupies one of a few possible states, thus allowing the modelling of different events as well as intermediate and successive events. Multi-state models are mostly interpreted as Markov models considering that these models acknowledge several states as well as the transitions between them, therefore allowing the calculation of the probability of transition between events through the earlier discussed TPM. Trench et al. (2003) created a Markovian decision-making method to lead a bank to identify the price of credit card owners in order to improve their profits. Additionally, this methodology was also used in revolving consumer credit accounts, influenced by the consumer's behavior as well as the impact on the economy of that behavior. Furthermore, Malik and Thomas (2012) conceived a Markov chain model based on developmental results to determine the credit risk of consumer loan portfolios.

In summary, multi-state models can assume different states through time. Most commonly, the Markovian assumption is adopted. In the particular case of loans, Markov chains helps to describe the dynamics of credit risk, since they estimate transition probabilities between different states. Nonetheless, Markov Models do not exclusively apply to credit studies. In fact, MCM have the most wide-ranging of applications. One of the most widely known case of Markov chains' application is Google's PageRank (Page, L., et al., 1998). This element shows us that this methodology can also be used in the world of computing and programming (to program algorithms) and computer science – randomized algorithms, machine learning, program verification, performance evaluation (quantification and dimensioning), modeling queuing systems and stochastic control.

As a statistical model, Markov chains have many applications in the real world, with such a wide range ranging from music (Volchenkov, D. & Dawin, J. R., 2012), to linguistics (Markov, 1913), finance (Siu et al., 2005; Fung and Siu, 2012), to the estimation of option prices (Norberg, R., 2003) and financial markets (Maskawa, 2003; Nicolau, 2014; Nicolau and Riedlinger, 2015), economics (Mehran, 1989), economic history (Damásio and Mendonça, 2018), operational research (Asadabadi, 2017; Tsiliyannis, 2018; Cabello, 2017), management (Horvath et al., 2005), forecasting (Damásio and Nicolau, 2013) and sports (Bukiet et al., 1997). They are also used in medicine (Li et al., 2014), biology (Gottschau, 1992; Raftery and Tavaré, 1994; Berchtold, 2001), such as DNA sequences and genetic networks, physics (Gómez et al., 2010; Boccaletti et al., 2014), astronomy and environmental sciences (Turchin, 1986; Sahin and Sen, 2001; Shamshad et al., 2005). Regarding the engineering area, Markov chains have been used in chemical engineering (to predict chemical reactions), physical engineering (to model heat and mass transfers), and aerospace. Indeed, most population models are Markov chains – they are used when we want to know how population changes over time or when we want to estimate the probability that a population, animal or plant, may be extinct.

The extensive use of Markov chains shows the great utility that this methodology has, not only within the applied mathematics area but also in most scientific areas. This versatility proves that a model that was first published in 1906 continues to be a handy and efficient tool in the most varied scientific aspects.

3. METHODOLOGY

In order to achieve the objective of this dissertation, we propose the application of an innovative hybrid methodology in the evaluation of the performance on loans, i.e., a new multidisciplinary combination of two distinct methodologies in this subject. These two methodologies consist of a clustering technique named Self-Organizing Maps (SOMs) and the application of the Markov Chains methodology. The application of the first methodology – SOM – will be the basis for the second and principal methodology – Markov Chains.

Fannie Mae's Single-Family Loan Performance Data is the data source employed in this study. This data comprises both borrower and loan information at inception, as well as performance data on loans. R software will be used to accomplish the main objective of this work.

3.1. SELF-ORGANIZING MAPS

SOMs were introduced by Teuvo Kohonen (1982) and are a class of artificial neural networks that use unsupervised learning[2] neural networks for feature detection in large datasets, identifying individuals with similar characteristics, organizing and gathering them by groups or clusters. This approach contrasts with other artificial neural networks, since they apply competitive learning[3] instead of error-correction learning[4].

A SOM comprises neurons in a grid, which iteratively adapt to the intrinsic shape of our data. The result allows us to visualize data points and identify clusters, being used to produce a low-dimension space of training samples. Therefore, its main objective is to reduce the dimensionality of data, performing a discretized representation of the continuous input space, where there are the initial dataset and the input vectors – lines of the matrix of observation –, named map. The reduction of dimensionality then occurs in the nodes or space where the vectors will be projected.

3.1.1. Theoretical Framework

The SOM algorithm follows five steps. Initially, we have an input space, $X \in \mathcal{R}^n$. At the start of the learning, each node's weight, $\{w_1, w_2, \dots, w_M\}$ is initialized, where w_i is the weight vector associated with each neuron, and M is the total number of neurons. Next, one data point is chosen randomly from the dataset, and then every neuron is examined to calculate which one's weights are more similar – and, therefore, closest – to the input vector. The winning node is known as the Best Matching Unit (BMU)[5]. The BMU is moved closer to the randomly chosen data point– the distance moved by the BMU is determined by a learning rate, which decreases after each iteration. The BMU's neighbors are also

[2] Unsupervised learning means that we only have input data and no output variables, contrasting with supervised learning, where input data and output variables are given.

[3] Competitive learning is a form of unsupervised learning artificial neural networks where, given the input, nodes compete with each other to maximize the output.

[4] Error-correcting learning is a type of supervised learning where we compare the system output with the desired output value and use that error (the difference between the desired and obtained values) to direct the training.

[5] BMU is a technique which calculates the distance from each weight to the sample vector, by running through all weight vectors. The weight with the shortest distance is the winner. The most commonly way used to determine that the distance is the Euclidean distance.

moved closer to that data point, through a neighborhood function, with farther away neighbors moving less. Finally, these steps are repeated for N iterations.

Succinctly, at each time t, present an input x_t, and select the winner, such as

$$\nu(t) = BMU = \arg min_{k\epsilon\Omega} \left\| x_t - w_{k_t} \right\| \tag{3.1}$$

where $\left\| x_t - w_{k_t} \right\|$ is the Euclidean distance.

Weights are adjusted after obtaining the winning neuron until the map converges to increase the similarity with the input vector. The rule to update the weight vector is given by

$$\Delta w_k(t) = \alpha(t)\eta(\nu, k, t)\left[x_t - w_{\nu_t}\right] \tag{3.2}$$

where coefficient $\alpha_t\ is$ the previously mentioned learning rate and $\eta(\nu, k, t)$ if a neighbor function.

The use of the SOM methodology becomes very interesting and useful because it allows us to map the input data, that is, it permits us to allocate customers to a particular group, with each group from the beginning, with each group formed containing customers with similar characteristics. This aspect proves to be quite advantageous not only for this study, since it facilitates the interpretation and evaluation of our data, but it is also a tool with great added value for banks since it allows them to replicate the same procedure in their business with new and ongoing customers. In such a way, it allows them to understand the profile of each customer in advance and thus make an initial forecast of the future behavior of those same customers, comparing them with others in the same group.

Since we have an extensive dataset, we can see that this methodology becomes quite useful in our case. Additionally to what was previously mentioned, the SOM methodology reveals to be a useful approach because it is a numerical and non-parametric method as well as a methodology that does not need *a priori* assumptions about the distribution of data and a method that allows the detection of unexpected characteristics in the data because of its use of unsupervised learning. The application of the SOM methodology makes it is possible not only to reduce the dimensionality, but also to organize the data. That is why its interpretation is simpler. This first methodology will allow the identification of comparable clients, let us say, with the same loan maturity, with the same interest rate or which performed the same statuses. Considering that the result of the application of this methodology is the organization of data in clusters that contain groups of clients with similar characteristics, we will then be able to compare groups of clients, instead of comparing individual clients. Additionally, it also makes it easier to apply Markov chains, since it allows the implementation of a Markov chain on each cluster.

3.2. Markov Chains

3.2.1. Theoretical Framework

3.2.1.1. First Order Markov Chains

The Markov chain is named after the well-known Russian mathematician Andrey A. Markov (1856-1922), distinguished for his work in number theory, analysis, and probability theory. He lengthened the weak law of large numbers and the central limit theorem to a specific series of dependent random variables. Accordingly, he created a special class, denominated Markov processes: random processes in which, given the present, the future is independent of the past. Therefore, a Markov chain is a Markov process defined into a countable state space. This factor means that the probability that the process will be in a given state at a given time t may be deducted from the knowledge of its state at time $t < t_{t-1}$ and does not depend on the history of the system before t.

Consider the stochastic process

$$\{X_t, t = 0, 1, 2, \dots\} \tag{3.3}$$

That takes discrete-time values at any time point t:

$$X_t = j\,, j = 0, 1, 2, \dots \tag{3.4}$$

in which j represents the state of the chain.

Without any loss of generality, to ease the notation we assume M to be finite, as follows:

$$M = \{1, 2, \dots, m\} \tag{3.5}$$

For the discrete-time context, we can conclude the present state X_t is independent of past states, such that:

$$P(X_t = j \mid F_{t-1}) = P(X_t = j \mid X_{t-1} = i) \tag{3.6}$$

where F_{t-1} is the $\sigma - algebra$ generated by the available information until $t - 1$.

Considering that we can calculate the probability of a state transiting to the next state – transition states –, we can then call this a transition probability. Hence, it is possible to construct a transition probability matrix (TPM):

$$\begin{bmatrix} P(X_t = 1|X_{t-1} = 1) & \cdots & P(X_t = m|X_{t-1} = 1) \\ \vdots & \ddots & \vdots \\ P(X_t = 1|X_{t-1} = m) & \cdots & P(X_t = m|X_{t-1} = m) \end{bmatrix} \tag{3.7}$$

This operation is denominated the one-step transition probability matrix of the process. Additionally, we can also calculate the probability that the chain will visit state j after n-steps given the fact that it was in state i at time $t - 1$. Thus, we have the n-step transition matrix, P^n, in which P is the one-step transition probability matrix and P^n is equal to $P \times P$ n times.

One of our objectives is to describe how the process travels from one state to another in time. Then we have:

$$P(X_t = j \mid X_{t-1} = i) \tag{3.8}$$

We are mostly concerned in how Markov chains evolve in time. From that point of view, there are two types of behaviors that are important to highlight: (i) transient behavior, which describes how chain moves from one state to another in finite time steps; and (ii) limiting behavior, which defines the behavior of X_n as $n \rightarrow \infty$. Thus, it is fundamental to define some concepts:

- For every state i in a Markov Chain, let f_i be the probability that beginning in state i, the process will ever re-enter state i.
- State i is said to be recurrent if $f_i = 1$ and transient if $f_i < 1$, i.e., if the probability is different from 1.
- A recurrent state is said to be positive if its mean recurrence time[6] is finite and is aided to be null if its mean recurrence time is infinite. Consequently, an irreducible[7] Markov chain is positive recurrent if all its states are positive recurrent. Positive recurrent irreducible Markov chains are often called ergodic.

A Markov Chain with a finite state space M is said to have a long-run distribution, i.e., a limit distribution if

$$\lim_{n \to \infty} P(X_{t+n} = a | \, F_{t-1}) = \pi_a \tag{3.9}$$

As previously mentioned, a Markov Chain is said to be ergodic if it is positive recurrent and aperiodic. Under these conditions, we have the following equation:

$$\pi P = \pi, \qquad with \sum_{i=1}^{m} \pi_i = 1 \;\; and \;\; \pi_i \geq 0 \tag{3.10}$$

where P is the PTM associated with the Markov Chain. Therefore, for any $n \geq 1$, we have:

$$\pi_i = P(X_t = i) \tag{3.11}$$

3.2.1.2. Absorbing Markov Chains

A Markov chain is absorbing if it has at least one absorbing state, and if from every state it is possible to go to an absorbing state. A state i of a Markov chain is called absorbing if it is impossible to leave it (Grinstead, C. M & Snell, J. L. 1999), such as:

$$P(X_t = i \mid X_{t-1} = i) = P_{ii} = 1 \tag{3.12}$$

When a Markov chain process attains an absorbing state, we must denominate it *absorbed*. By opposition, a state which is not absorbing is called transient, a definition that was previously provided.

Consider an arbitrary absorbing Markov chain. Now reorder the states so that the transient states come first. With t transient states and r absorbing states, the transition matrix P can be written in the following canonical form:

$$P = \begin{pmatrix} Q & R \\ \mathbf{0} & I \end{pmatrix} \tag{3.13}$$

[6] Mean recurrence time is the average time it requires to visit a state i, starting from i.

[7] A Markov chain is said to be irreducible if all states belong to the same class. State i and state j are said to communicate if state i and state j are accessible (starting from state i, it is possible to enter state j in future number of transitions) (Ching, W. & Ng, M., 2016).

where I is an r-by-r identity matrix, $\mathbf{0}$ is an r-by-t zero matrix, R is a nonzero t-by-r matrix, and Q is a t-by-t matrix. The first t states are transient and the last r states are absorbing.

For an absorbing Markov chain, the matrix $I - Q$ has an inverse N matrix, called the *fundamental matrix.* The entry n_{ij} of N gives the expected number of times that the process is in the transient state j if it is started in the transient state i. The decomposition of the transition matrix into the fundamental matrix allows for certain calculations such as the *mean time of absorption*, i.e., the mean number of steps until absorption from each state. Accordingly, the fundamental matrix N can be written as follows:

$$N = (I_t - Q)^{-1} \tag{3.14}$$

where I_t is a t-by-t identity matrix.

Additionally, it is possible to define the time of absorption as follows. Let t_i be the expected number of steps before the chain is absorbed, given that the chain starts in state i. Now let t be the column vector whose ith entry is t_i. Then,

$$t = Nc \tag{3.15}$$

where c is a columns vector whose entries are 1.

Furthermore, it is possible to define the *probability of absorption*[8] by a specific absorbing state when the chain starts in any given transient state. Let b_{ij} be the probability that an absorbing chain will be absorbed in the absorbing state j if it starts in the transient state i. Now let B be the matrix with entries b_{ij}. Then B is a a t-by-t matrix and

$$B = NR \tag{3.16}$$

Where N is the fundamental matrix and R is as in the canonical form.

Now that a brief theoretical framework of absorbing Markov chains was provided, it is possible to verify that, with the application of the proposed hybrid methodology, the estimation of the probability of absorption as well as the estimation of the mean absorption time, we will be able to perform an important comparison between the different types of credit (modified *versus* unmodified), as well as a comparison between the clusters calculated within of each type of credit, regarding two specific states, that we will address later.

The main objective of this work is to evaluate, based on the results obtained from the proposed hybrid methodology, if the loan modifications are, in fact, effective. For that, we will estimate the probability of client defaults considering unmodified loans and the probability of default considering modified loans whereby the terms of the contract are altered. To that end, we will stack individuals and eliminate spurious transitions, that is, transitions between individuals and, therefore, between credits.

[8] Given a transient state i we can define the absorption probability to the recurrent state j as the probability that the first recurrent state that the Markov chain visits (and therefore gets absorbed by its recurrent class) is $j, f_i^* j$ (Spedicato, G.A., Kang, T.S., Yalamanchi, S.B., Yadav, D. & Cord´on, I., 2014) .

Thus, we will be able to estimate the transition probabilities based on the performance of each individual.

By applying a Markov chain on each cluster obtained from the application of the SOM methodology, we will estimate a TPM for each cluster built for unmodified modified credits, considering the respective loan modifications. The objective is to compare the estimated TPMs and evaluate the differences between modified and unmodified credits. Therefore, we are able to evaluate if the modifications are, in fact, effective and what modifications are most efficient.

The use of these two methodologies together has proven to be quite useful in other studies outside the financial scope. A hybrid approach combining a SOM and a Hidden Markov Model (HMM) was previously used to meet the increasing requirements by the properties of DNA, RNA and protein chain molecules (Ferles, C. & Stafylopatis, A. 2013), as well as an application concerning o stroke incidence (Morimoto, H. 2016). Additionally, it was adopted as a hybrid methodology to forecast the influence of climatic variables (Sperandio, M, Bernardon, D. P. & Garcia, V. J. 2010), to test speech recognition (Somervuo, P. 2000) and also to analyze career paths, as a study to evaluate the insertion of graduates and to identify the main career paths categorizations (Massoni, S., Olteanu, M & Rousset, P 2010).

Although the hybrid use of these methodologies has been implemented in other areas, its use in financial and banking areas represents an interdisciplinary innovation. Thus, not only is it presented as a methodology that simplifies the interpretation and processing of data, but also as an innovative approach.

4. DATA

4.1. DATA SET

The primary dataset used in this study is Fannie Mae's Single-Family Loan Performance Data, that provides data on US mortgages purchased from original lenders. The Single-Family Fixed-Rate Mortgage (primary) dataset contains a subset of Fannie Mae's 30-year and less, fully amortizing, full documentation, single-family, conventional fixed-rate mortgages (Fannie Mae, 2019).

We will analyze a total of 149 404 loans acquired by Fannie Mae in 2006, divided into 40 079 modified credits and 109 325 unmodified credits, following their performance until 2015. This timespan is an interesting period to evaluate, since there was an economically and financially critical period that, as previously mentioned, began in 2007. We are then able to track loans that were purchased in the pre-crisis period and evaluate their development through the crisis period and the post-crisis period. Hence, it is interesting to evaluate these 10 years, since it is transversal to several scenarios that occurred during this time.

The financial crisis occurred all over the world, and it is noteworthy to evaluate, especially in the United States. During this period, some remarkable events occurred, such as the collapse of Lehman Brothers and also significant changes in monetary policies. This cataclysm led to historically low-interest rates and the approval of two large-scale debt relief programs – the Home Affordable Refinancing Program (HARP) and the Home Affordable Modification Program (HAMP) – along with the foundation of the Troubled Asset Relief Program (TARP).

The population data is divided into quarters, and for each quarter, we have a division in "Acquisition" and "Performance" datasets. We can assess the full history of the contracts in each quarter which means that it does not represent a three-month observation of the mortgages. The "Acquisition" dataset has the information on the origination of the credit and the "Performance" dataset has the full credit information related to its evolution, having a Loan Identifier (ID) that links the "Acquisition" dataset to the "Performance" dataset. In this way, we ensure that the subsequent performance of a loan can be monitored from the outset, therefore allowing the modelling of the various loan outcomes.

The "Acquisition" data includes static data on both borrower and loan information at the time of origination. This information comprises the Acquisition Data elements, such as the Interest Rate, the Loan Amount, the Number of Borrowers, the Borrower Credit Score, and the Loan Term. The "Performance" data includes the proceeding of loans from the time of the acquisition up until its current status. This dataset is segregated in months and displays the loan performance characteristics, since it considers a dynamic performance over time. The information that follows the behavior of the clients is contained in the Performance Data Elements, such as the Current Loan Delinquency Status, the Zero Balance Code, the Current Interest Rate and the Modification Flag. Further details on these data elements are available in Table 10.1 and Table 10.2 presented in the annexes.

Some variables were modified in order to allows us to apply the previously described SOM methodology and also to be more adequate for our analysis. Later, in section 4.2., we will outline the variables considered in this study as well as the ones that were modified. A description of those modifications and the reasoning behind it will also be provided.

Before we go further in this study, i.e., before we develop the SOM and Markov chain methodologies, it is fundamental to analyze the global data set. This analysis will allow us to understand its composition before segmenting it into modified and unmodified credits as well as to assess the disparity of each credit class – modified or unmodified – in relation to the entire data set. Accordingly, in Table 4.1 we present the descriptive statistics, as well as a complementary analysis of these results.

By analyzing the descriptive statistics, we can observe that about 10% of the individuals were first time home buyers, with the majority of the contracts owned by one borrower (about 53%), even though a considerable percentage of the contracts were owned by two borrowers (about 46%). Additionally, most of the contracts were originated by correspondent lending, which is the process through which a financial institution underwrites mortgage loans using its own capital. Nevertheless, a considerable segment of the contracts was purchased through retail lending, i.e., it is based on lending to individual or retail customers, most often by banks, and institutions focused solely on the credit business.

Furthermore, the interest rate of the contracts under analysis range from 3,000% to 10,950% with a mean of 6,470% and a mode of 6,500%. Regarding the loan amount, we have a range from $7 000 to $802 000 with a mean value of $155 449 and a mode of $100 000. Most of the contracts have a duration of 360 months, which corresponds to 30 years. This scenario is typical since we are considering mortgage loans, which are usually very long-term contracts.

Regarding the risk characteristics of the individuals, we have the LTV ratio, the DTI ratio, and the Borrower Credit Score. The LTV ratio corresponds to the percentage of the property value that the loan covers, which means that if we have an LTV ratio of 70% it indicates that the loan covers 70% of the property appraisal value. Therefore, the higher the amount borrowed, the greater the risk the bank takes, since it means that the bank lends a larger amount of money. In fact, in some situations, derived from the high risk taken by the bank, it may require the borrower to purchase mortgage insurance to offset that risk. In the data set under study, we have an LTV ratio between 1% and 97% with a mean value of 7% and a mode of 80%. Although most banks only allow a loan that corresponds to a maximum of 80% of the property appraisal value, in our case study, we have values of 97% because Fannie Mae had a program for low-income borrowers that allow an LTV of this value. However, it requires mortgage insurance until the ratio falls to 80%.

Regarding the DTI ratio, that is the total of monthly debt payments divided by the gross monthly income, we can infer, by its meaning, that the lower this ratio, the better, representing a lower risk individual. Here we have a range of 1% to 64%. Additionally, the mode presents a value of 40%, which means that the majority of individuals included in this cluster applies almost 40% of their monthly income to pay their debts.

Lastly, the credit score of individuals lies between 378 and 850. Considering that this variable can assume values between 300 and 850, we can conclude that we are in the presence of very different clients in respect of the primary classification of credit risk. Furthermore, we have a mode value of 675. Considering that in this variable scores above 650 indicate a good credit history, we can infer that the majority of the individuals under study have a credit score that can be considered favorable.

Table 4.1 - Descriptive Statistics Total Data Set

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate (%)	3,000	10,950	6,470	6,500	6,500
Original Loan Amount	7 000	802 000	155 449	100 000	135 000
Original Loan Term (months)	96	360	326	360	360
Original Loan to Value (LTV) (%)	1	97	70	80	75
Original Debt to Income (DTI) Ratio (%)	1,00	64,00	38,87	40,00	39,00
Borrower Credit Score at Origination	378	850	700	675	698

	Class	Percentage
Channel / Origination Type	Broker	17,342
	Correspondent	44,004
	Retail	38,654
Number of Borrowers	1	52,898
	2	46,498
	3	0,443
	4	0,153
	5	0,004
	6	0,001
First Time Home Buyer Indicator	No	90,158
	Yes	9,691

4.2. Data Preparation

The first step to consider was a screening of the credits to be considered. As mentioned above, the loans originated in 2006, and, in order to observe a reasonable period, a ten-year analysis interval was considered. Along these lines, we have information about the performance of credits until 2015. This factor means that all credits that had no information available until 2015 were withdrawn. Similarly, all information exceeding the period considered, that is, all information after 2015 was not taken into account for this study. Additionally, we also had some credits with information gaps, namely information breaches greater than one year which, in order to ensure the veracity of this study and also to respect the 10-year period considered for this study, were also removed from the analysis.

After screening the credits to be analyzed, we proceeded to the development of the first methodology – SOM. This methodology has the particularity of only supporting numerical variables. According to this, considering that we have numerical and categorical variables, we needed to carry out some adjustments. As a result, the variables that underwent some amendments were the following: Origination Channel, First Time Home Buyer Indicator, Modification Flag, Origination Date, Modification Date, Maturity Date and Current Delinquency status.

The Origination Channel first presented the values B (Broker), C (Correspondent), and R (Retail). By transforming them into a numeric variable, we now have the following values: 1, which corresponds to “Broker,” 2, which corresponds to “Correspondent” and 3, which corresponds to “Retail.” The First

Time Home Buyer Indicator and the Modification Flag were binary variables, i.e., the possible values were No (N) if the individual was not a first time home buyer or if the credit was not modified and Yes (Y), if the individual was a first time home buyer or if the credit was modified. For the Modification Flag variable, we now have the value 1, which corresponds to "No," and the value 2, which corresponds to "Yes." The First Time Home Buyer Indicator has the particularity of some lack of information. Due to that fact, we might have empty values, represented by the letter U (Unknown). Therefore, by transforming this variable, and because this transformation is performed alphabetically, in this specific case, we have the following values: 1, that corresponds to "No," 2, that corresponds to "Unknown" and 3, that corresponds to "Yes."

Regarding the date variables, there was a need for a different treatment between some of them. For the Modification Date variable, we had a date with the "month/day/year" format, which was transformed into the number of months that occurred between the date of origin of the loan and the time of its modification. For the Origination Date and Maturity Date variables, which also had the same type of format, we preserved the year of origination and the year of maturity.

To conclude the description of all the implemented changes, we have the Delinquency Status variable. Initially, this variable was represented in the number of days the client was delinquent. Since we will apply a Markov Chain methodology, it becomes necessary to have these variables in states that comprise an interval of the days of delinquency. For this, and also because the SOM only allows numeric variables, we chose to modify this variable to states 1, 2, 3, and 4. Additionally, it was noted that, in some cases, this variable assumes a value "X" on the last date of observation. When this happens, the variable Zero Balance Code only presents the values "01" or "06", that corresponds to performing situations, i.e., situations where individuals have a normal or current performance. In order to validate this, it was noted that in all these cases, in the penultimate moment of observation, all individuals had less than 30 days past due, which, once more, means that they were all in performing positions. Therefore, for these cases, we added a state, represented by the number 5, that corresponds to situations where an individual prepaid the loan being in a normal performance position.

It was also noted that in the cases where we did not have information on the Delinquency Status variable (i.e., an N.A. value), the Zero Balance Code variable presents the remaining codes, that is, the codes "02", "03", "09" or "15". These different codes correspond to situations where the individual was in a non-performing position and, for reasons of non-payment of the credit, the bank is forced to reduce the credit to zero. Accordingly, for these cases, we added a state, represented by the number 6. Thus, we have six different states, described in Table 4.2.

Table 4.2 - States of the Markov Chain Methodology

States	Designation	Description
State 1	Current/Normal Performance[9]	0 to 29 days past due
State 2	Delinquency[10]	30 to 59 days past due
State 3	Pre-Default[11]	60 to 89 days past due
State 4	Default[12]	90 to 119 days past due
State 5	Prepayment[13]	Loan is reduced to zero
State 6	Third-Party, Short or Note Sales / REO[14]	Loan is reduced to zero

[9] Current or Normal Performance corresponds to a credit performance situation in accordance with compliance.

[10] Delinquency corresponds to a situation where the borrower has failed to make payments as required in the loan documents. In this case, we consider a period of 30 to 59 consecutive days of payment failures.

[11] Pre-Default is a state that corresponds to a period of time that comprises up to 30 days less than the Default state – precedes the Default state.

[12] Default is similar to the Delinquency state, i.e., it corresponds to a failure to repay the principal and/or interest on a loan or security. In this case, we consider a period of 90 to 119 consecutive days of payment failures.

[13] Prepayment is the terms used for the settlement of a debt or installment loan before its official due date.

[14] These situations are quite similar, with only a few specific details that differentiate them. In its essence, it corresponds to a sale of the property by a financially distressed borrower for less than the outstanding mortgage balance in order to repay the lender with the income obtained from the sale or to situations where the bank takes possession of the property to recover the money lost as a result of late payment on credits.

5. RESULTS AND DISCUSSION

5.1. SOM

This section presents the results obtained from the application of the SOM methodology. This first methodology serves as the basis for the second methodology. It aims to try to understand how clients group together in clusters considering their similar characteristics. In this way, through this hybrid methodology, innovative in matters relating to banking, it will be possible to model the behavior of groups of individuals. The use of this hybrid methodology is necessary, since it contrasts with what has been studied and performed until today. Currently, the application of a methodology of a simple assessment of the risk of each individual at the time of contracting is still quite frequent, which is overcome with this more complex, but more complete, methodology.

The application of the SOM methodology is carried out through the K*ohonen* package in R software. Aside from the construction of clusters, this package allows the visualization of the quality of our developed SOM and the evaluation of the relationships between the variables in our dataset. This evaluation is accomplished by several plots:

- The *training iterations progress* plot, that represents the distance from each node's weight to the samples represented by that node;
- The node counts plot, that grants the visualization of how many samples are mapped to each node on the map. Ideally, the sample distribution should be reasonably uniform;
- The neighbor distance plot, also known as the "U-Matrix," it represents the distance between each node and its neighbors. Areas of low neighbor distance indicate groups of similar nodes. Contrarily, areas with large distances indicate dissimilar nodes;
- Codes or weight vectors plot that allows the identification of patterns in the distribution of samples and variables;
- Heatmaps plot that allows the visualization of the distribution of a single variable across the map. Commonly, a SOM process involves the creation of multiple heatmaps and then the comparison of these heatmaps to identify interesting areas in the map.

In section 5.1.1, we present the results obtained for modified credits and the results for unmodified credits in section 5.1.2.

5.1.1. Modified Credits

Progression of the Learning Progress. As mentioned above, the plots available in the *Kohonen* package are a handy tool to assess the quality of the developed SOM model. Therefore, it makes sense to initiate this evaluation with the assessment of variations along the number of iterations of the model, since it allows us to make some conclusions on the stability of it. The number of iterations is defined in the software routine. However, there should be a certain criterion with the choice of the number of iterations. If the curve that represents the stability of the model is continuously decreasing, more iterations are necessary to consider. In the case of modified credits, 300 iterations were considered, and the Training Progress plot is presented in Figure 5.1.

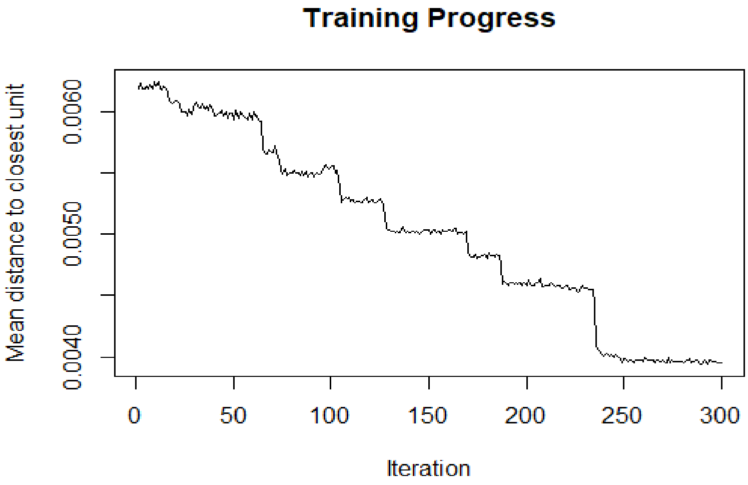

Figure 5.1 - Training Progress of Modified Credits

Through the analysis of Figure 5.1, it is possible to verify that the number of iterations is sufficient, since as the number of iterations increases, the average distance to the nearest cell in the map decreases, and from nearly 250 iterations we reach stability where there is no longer a continuous decrease of that distance. As such, we can proceed with the model in the way it was defined.

Node Counts Plot. After this first analysis, it is interesting to evaluate the number of instances included in each neuron, since this allows us to define whether it is necessary to increase or decrease the size of our map. The size of the map must be reduced if there are too many empty cells and increased if there are areas with very high density. This conclusion should be based on the colors of the chart, as we can see in Figure 5.2.

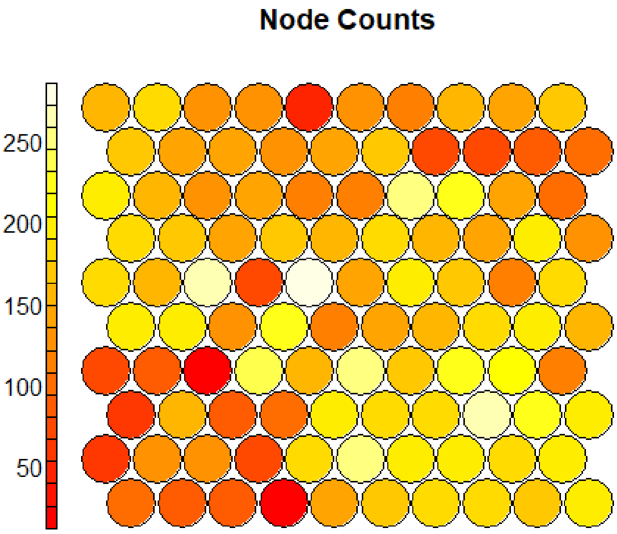

Figure 5.2 - Node Counts of Modified Credits

As mentioned earlier, the distribution should be relatively uniform, which means that, considering the type of graph presented, we should not have large variations in color, i.e., it should be homogeneous.

On the left axis of the plot presented in Figure 5.2 we can observe the scale that allows us to interpret this plot. This scale allows us to assess if nodes tend more to the red color, these contain a smaller number of samples, while the lighter color, i.e., if nodes tend more to the yellow color, it means that these contain a greater number of samples. Evaluating the plot shown in Figure 5.2, we can conclude that there is no major color variation, which makes our distribution relatively uniform, as desired. Additionally, we can notice that there are no empty nodes that would be colored in grey. Additionally, we can observe that there is not a great number of nodes with large values since most contain between 50 and 200 observations. Therefore, there is no need to adjust the size of the map.

Neighbor Distance Plot. As previously mentioned, this plot is often referred to a “U-Matrix”. This nomenclature because it represents a unified distance matrix. Thus, in this plot, we can visualize a Euclidean distance between the code book vectors of neighboring neurons, represented by colors. As in the graphics presented above, we must guide ourselves by the scale displayed on the left side of the chart.

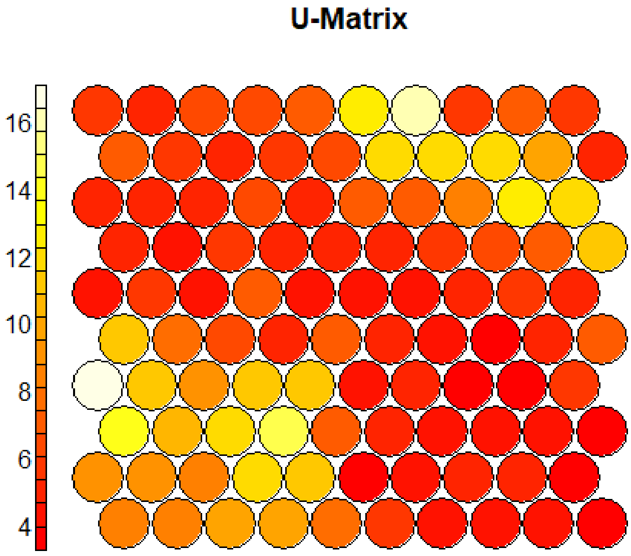

Figure 5.3 - U-Matrix of Modified Credits

In this type of plot, the rationale we must follow is the intensity of color along with the values presented by the scale. That is, the darker the color, the closer the groups of nodes are, which means that they are more similar. Conversely, neurons with lighter colors represent areas with a greater distance between neurons and, consequently, represent more dissimilar individuals. However, we should not forget to look at the values that the scale presents, since, as we can see in Figure 5.3, we can verify that the distances go from 4 to 16, which means we have neurons relatively close to each other. This plot is particularly important, since the construction of clusters is based on the distance between nodes, considering that each cluster is composed of the nearest neurons.

Clustering. Finally, we have the construction of clusters. The clustering process in the SOM methodology is carried out to group individuals with similar characteristics. This way, it is easier to interpret results and also apply the second methodology of this dissertation, the Markov Chains. However, it is necessary to estimate the optimal number of clusters. For this, an examination of the "within-cluster sum of squares" plot is carried out presented in Figure 5.4.

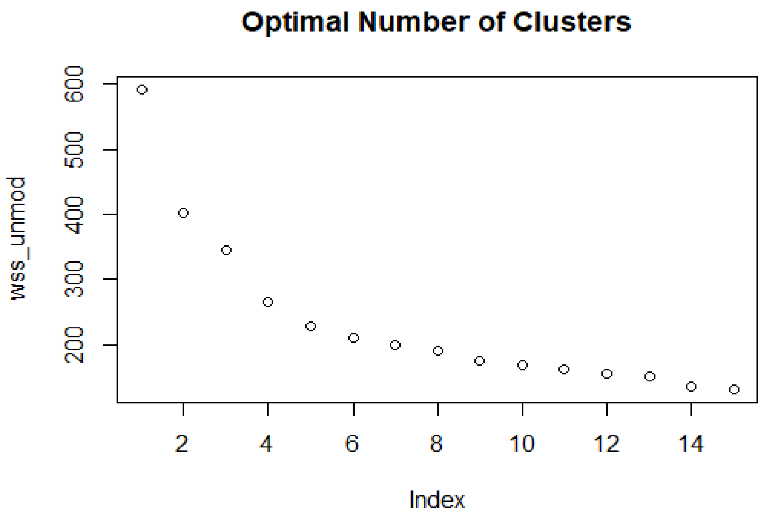

Figure 5.4 - Optimal Number of Clusters of Modified Credits

The rationale to identify the optimal number of clusters is to find the "elbow point" on the plot, that is, the point at which we verify a slight stabilization of the graphic. Although it is not very obvious, we can see in Figure 5.4 that the elbow point is situated in four clusters, so that is what we must consider. Thus, it is concluded that, in the case of modified credits, we will have four distinct clusters. These clusters can be observed in Figure 5.5.

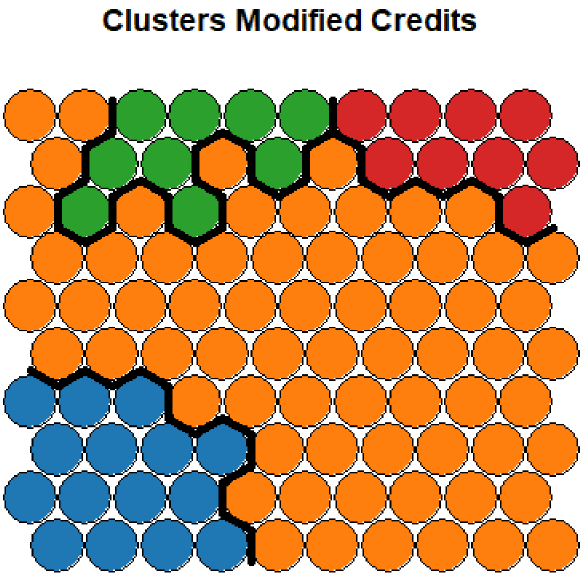

Figure 5.5 - Clusters of Modified Credits

In Figure 5.5, we can observe the four clusters defined earlier. The green cluster contains nine neurons, the red cluster contains nine neurons, the orange cluster contains 66 neurons, and the blue one contains 15 neurons. Consistently with what was described in the evaluation of the Nodes Count plot, there are no empty nodes, which means that we have observations in all neurons.

Before applying the Markov chain methodology, it is crucial to analyze each cluster in order to be able to characterize them and identify some patterns that may exist within each cluster. This step will be accomplished by analyzing the descriptive statistics of each cluster, and since we have four clusters, we will have four different tables, each one containing the descriptive statistics of the borrowers and loans characteristics.

Starting with the first cluster, the information of which is contained in Table 5.1, we can observe that only about 8% of the individuals were first time home buyers, with the majority of the contracts owned by one borrower, even though a considerable percentage of the contracts were owned by two borrowers – about 46%. Additionally, most of the contracts were originated by correspondent lending. Nevertheless, a considerable segment of the contracts was purchased through retail lending.

Furthermore, the interest rate on these contracts ranges from 4,750% to 8,375% with a mean of 6,490% and a mode of 6,5%. Regarding the loan amount, we have a range from $23 000 to $534 000 with a mean value of $213 366 and a mode of $417 000. Due to the reason previously mentioned, most of the contracts have a duration of 360 months (30 years).

More related to the risk characteristics of the individuals, we have the LTV ratio, the DTI ratio and the Borrower Credit Score. Regarding the DTI ratio, we have a range of 3% to 64%. Additionally, the mode presents a value of 44%, which means that the majority of individuals included in this cluster apply 44% of their monthly income to pay their debts. Finally, the credit score of individuals lies between 501 and 814. Considering that this variable can assume values between 300 and 850, we can conclude that we are facing an extensive range. However, if we look at the mode value, we observe a score of 637. Thus, we can see that most individuals in this cluster have a poor credit history and are more susceptible to default situations.

Table 5.1 - Descriptive statistics (Modified Credits – Cluster 1)

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate	4,750%	8,375%	6,490%	6,500%	6,500%
Original Loan Amount	23 000	534 000	213 366	417 000	200 000
Original Loan Term (months)	120	360	348	360	360
Original Loan to Value (LTV)	20%	97%	73%	80%	76%
Original Debt to Income (DTI) Ratio	3,00%	64,00%	42,63%	44,00%	43,00%
Borrower Credit Score at Origination	501	814	676	637	671

	Class	Percentage
Channel / Origination Type	Broker	19,785
	Correspondent	45,547
	Retail	34,668
Number of Borrowers	1	53,198
	2	46,473
	3	0,269
	4	0,060
First Time Home Buyer Indicator	No	91,961
	Yes	8,039

Observations: 3 346

Now evaluating the second cluster, whose information regarding the descriptive statistics is contained in Table 5.2, it is possible to observe that in this group of individuals, similarly to cluster one, only about 8% were first time home buyers, with the majority of contracts owned by one borrower. However, we also have a great percentage of the contracts owned by two borrowers – almost 47%. Additionally, most contracts were originated by correspondent lending – about 46% – and a considerable part was purchased through retail lending – nearly 34%.

With respect to the conditions of the contracts, the interest rate has a range from 4,50% to 8,50%, which is a greater interval than the one in the first cluster, with a mean of 6,497% and a mode of 6,5%. Regarding the loan amounts, we have a range from \$16 000 to \$675 000, with a mean value of \$211 446 and a mode of \$417 000. Most of the contracts have a duration of 30 years (360 months). As previously mentioned, this is a typical situation, and as we will be able to observe, this situation will be verified in every cluster, in both modified and unmodified credits.

More related to the risk characteristics of the individuals, the LTV ratio presents a range from 8% to 97%, which is a greater interval compared to the first cluster. Although the mode value is the same, by having a minimum value lower than the one previously verified, we can infer that we are facing some individuals that represent a slightly lower risk, considering the definition provided for this variable earlier.

Regarding the DTI ratio of the second cluster, we have an interval between 1,00% and 64%. Similar to the situation with the LTV ratio, this also means that we are in the presence of lower-risk individuals, considering that we have a minimum value lower than the one verified in the first cluster, even though the mode value is quite the same.

Finally, the credit score of individuals is between 437 and 825. We can note that the minimum value is lower than the one verified in the first cluster. However, if we look at the mode, we can observe a higher value, which, considering that scores above 650 indicate a good credit history, we are facing individuals that represent a lower risk, validating what was previously mentioned regarding the last two variables.

Table 5.2 - Descriptive Statistics (Modified Credits – Cluster 2)

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate	4,500%	8,500%	6,497%	6,500%	6,500%
Original Loan Amount	16 000	675 000	211 446	417 000	199 000
Original Loan Term (months)	120	360	347	360	360
Original Loan to Value (LTV)	8%	97%	73%	80%	76%
Original Debt to Income (DTI) Ratio	1,00%	64,00%	42,78%	43,00%	43,00%
Borrower Credit Score at Origination	437	825	677	675	673

	Class	Percentage
Channel / Origination Type	Broker	19,850
	Correspondent	46,349
	Retail	33,801
Number of Borrowers	1	52,631
	2	46,927
	3	0,365
	4	0,071
First Time Home Buyer Indicator	No	91,768
	Yes	8,186

Observations: 30 978

Now assessing the third cluster, whose descriptive statistics are summarized in Table 5.3, we can verify that about 7,5% of the individuals were first time home buyers, and about 52% of the contracts are owned by one borrower. However, just like in the two previous situations, we also have a great percentage of contracts owned by two borrowers – about 47%. Additionally, most of these contracts were originated by correspondent lending.

Regarding the contract conditions, the interest rates range from 5,00% to 8,375%, which is a smaller interval than the ones previously verified, even though the mean and mode values are very similar. In the loan amounts we have a range from $17 000 to $645 000, with a mean value of $213 525 and a mode of $417 000. As stated before, the majority of contracts also have a maturity of 30 years.

Regarding the inherent risk with individuals, the LTV ratio presents a range from 16% to 97%. Although the mode value is the same, by having a minimum value greater than the one previously verified, we can infer that we are facing some individuals that represent a slightly higher risk when compared to the second cluster; however they do not representing a risk as high as the one verified with the individuals included in the first cluster.

The DTI ratio of the third cluster spans between 8,00% and 64%. Similar to the situation with the LTV ratio, this also means that we are in the presence of some individuals with higher risk, considering that we have a minimum value higher than the one verified in the first cluster, even though the mode value is the same.

Finally, the borrowers' credit score in this cluster stands between 462 and 817. Compared to the other two clusters, we can note that this interval is the smallest. By having the lowest minimum value so far, considering the definition of this variable, provided earlier, we can infer that we are in the presence of some individuals that represent a higher risk – the worse the credit score, the worse the risk level. However, if we look at the mode, we can also conclude that we have a value higher than 650, which means that most of these individuals have a good credit history, enabling the conclusion that, in general, the third cluster comprises individuals with a lower risk, contrasting with the situation in the first cluster.

Table 5.3 - Descriptive Statistics (Modified Credits – Cluster 3)

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate (%)	5,000	8,375	6,497	6,375	6,500
Original Loan Amount	17 000	645 000	213 525	417 000	200 000
Original Loan Term (months)	96	360	346	360	360
Original Loan to Value (LTV) (%)	16	97	73	80	75
Original Debt to Income (DTI) Ratio (%)	8,00	64,00	42,90	43,00	43,00
Borrower Credit Score at Origination	462	817	676	672	672
	Class			**Percentage**	
Channel / Origination Type	Broker			19,301	
	Correspondent			48,286	
	Retail			32,413	
Number of Borrowers	1			52,213	
	2			47,388	
	3			0,300	
	4			0,100	
First Time Home Buyer Indicator	No			92,346	
	Yes			7,654	

Observations: 3 005

Now evaluating the fourth and last cluster, whose descriptive statistics are summarized in Table 5.4, we can verify that about 7,7% were first time home buyers, with almost 52% of the contracts owned by one borrower. However, just like in the previous situations, a great percentage of contracts are owned by two borrowers, standing very close to the percentage of contracts owned by one borrower – practically 48%. In fact, almost all the contracts are owned by one or two borrowers, with the contracts owned by three or four borrowers representing not even one percent of all contracts. Nevertheless, a considerable segment of the contracts was purchased through correspondent and retail lending, representing, together, about 80% of the contracts.

Concerning the contract conditions, the interest rates vary between 4,990% to 8,125%, the smallest interval of all clusters, with a mean of 6,503% and a mode of 6,5%. Regarding the loan amount, we have a range from $25 000 to $548 000 with a mean value of $211 693 and a mode of $417 000. As previously stated, we again find the majority of contracts with a maturity of 30 years.

More related to the risk characteristics of the individuals, the LTV ratio presents a range from 14% to 97%. Although the mode value is the same, the minimum value is the second highest one, which means that we are in the presence of some individuals that also represent some risk.

The DTI ratio of the fourth cluster comprises between 2,00% and 64%. Additionally, the mode presents a value of 45%, which means that the majority of individuals included in this cluster apply 45% of their monthly income to pay their debts. Furthermore, by having the highest mode value verified so far, we can state that this group of individuals represents a slightly higher risk for this variable. However, the values verified in this variable are not very far from those observed in the remaining clusters, so these individuals do not represent a notably higher risk, which is corroborated by the fact that the average value is the lowest of all clusters.

Finally, the credit score of individuals is comprised between 432 and 817. We can note that the minimum value is the lowest of all clusters. However, if we look at the mode, we can observe that this cluster presents the highest value, which means that this cluster comprises more individuals that represent less risk as far as the quality of credit history.

Table 5.4 - Descriptive statistics (Modified Credits – Cluster 4)

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate	4,990%	8,125%	6,503%	6,500%	6,500%
Original Loan Amount	25 000	548 000	211 693	417 000	198 000
Original Loan Term (months)	120	360	348	360	360
Original Loan to Value (LTV)	14%	97%	73%	80%	75%
Original Debt to Income (DTI) Ratio	2,00%	64,00%	42,31%	45,00%	42,00%
Borrower Credit Score at Origination	432	817	676	676	672

	Class	Percentage
Channel / Origination Type	Broker	19,709
	Correspondent	46,364
	Retail	33,927
Number of Borrowers	1	51,745
	2	47,927
	3	0,218
	4	0,109
First Time Home Buyer Indicator	No	92,218
	Yes	7,745

Observations: 2 750

After an individual analysis of the results obtained relative to the descriptive statistics of each cluster, it is equally important to make an overall evaluation. By comparing the clusters with each other, in contrast to an individual assessment, it becomes possible to draw the different risk profiles that each group of individuals represent for the bank.

By comparing the four clusters, it can be concluded from the outset that regarding the number of borrowers, the origination type, whether or not individuals are first time home buyers, and the loan term, the four clusters present approximately the same results. Thus, we can determine that the clusters differ mainly in the borrowers' characteristics, that is, the LTV, the DTI ratio and the Borrower Credit Score.

Regarding the LTV variable, considering that both the mean and mode are equal throughout each cluster, looking at the minimum values of each group of individuals, we can see that the second cluster has the lowest minimum value. Therefore, we can conclude that this cluster presents a lower risk to the bank regarding the percentage of the property appraisal value granted.

Apropos the DTI ratio, although the mean and mode values differ between each cluster, we can see that the third cluster has the greater mean value, meaning that it contains individuals that pose a higher risk to the bank as they allocate, on average, more of their monthly income with debt compared to the remaining clusters.

Finally, regarding the Borrower Credit Score, by comparing the four clusters, we can observe that both the mean and mode values are approximately the same. However, we can discern that the first cluster has the lowest mode, being the only one to present a figure below 650. Nevertheless, when we look at the values presented by cluster three, we can identify that the minimum value of this variable is the lowest of all. This aspect means that this cluster comprises some individuals with a relatively poor credit history considering that with a value of 462, it is well below 650. Thus, we can conclude that in relation to the borrower credit score, clusters one and three present the greatest risk to the bank, since they have a worse credit history than the other clusters.

Concluding this overall assessment, in consideration of what was stated above, we can verify that the second cluster represents the lowest risk, and the third cluster exemplifies the highest risk.

5.1.2. Unmodified Credits

Progression of the Learning Progress. We now move to the development of the SOM methodology for unmodified credits. Once again, we start the analysis with the graphic that represents the variations existing along the number of iterations. As previously delineated, the purpose of this graphic is to present increasing stability throughout the occurrence of iterations. Only in this way we can guarantee that the number of iterations is sufficient in order to have a good quality model. As in the case of modified credits, we considered 300 iterations. In Figure 5.6 we can observe the Training Progress of the SOM model developed in respect of unmodified credits.

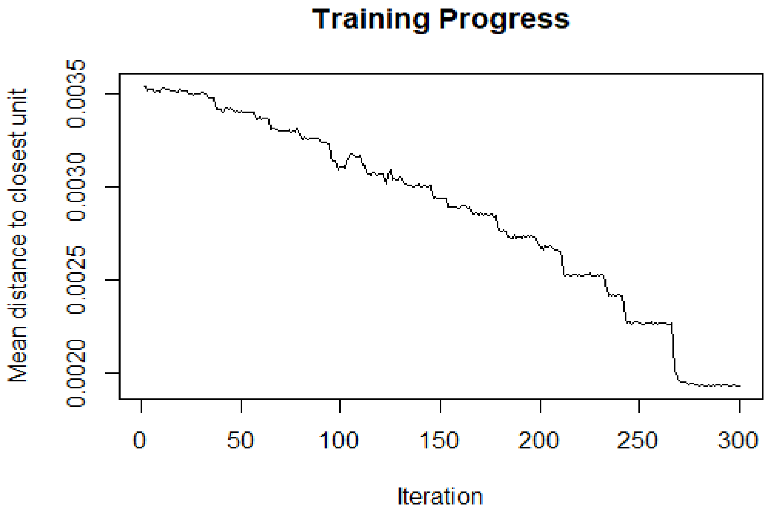

Figure 5.6 - Training Progress of Unmodified Credits

Through the analysis of Figure 5.6, it is possible to confirm that the number of iterations is adequate since, after a sharp decrease, we reach the desired stability shortly after 250 iterations, where there is no longer a continuous decrease of the distance between nearest cells in the map. As such, we can proceed with the model in the way it was defined.

We can point out a slight difference between the model developed for modified credits and the model developed for unmodified credits. In the case of unmodified contracts, more iterations were needed, compared to the case of modified contracts.

Node Counts Plot. As previously determined, the analysis of the Node Counts plot is interesting since it allows us to assess the necessity to change the size of our map, based on the number of instances present in each neuron. We should increase the map size if there are too many observations per node and reduce it if there are empty nodes. In Figure 5.7 we can see the Node Counts graphic related to unmodified credits.

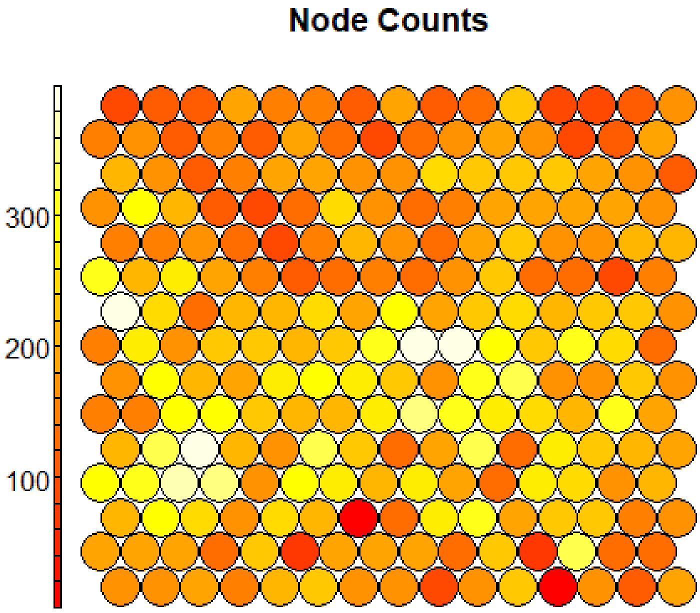

Figure 5.7 - Node Counts of Unmodified Credits

Firstly, it is important to mention that, as we can observe in Figure 5.7, the maps of unmodified credits are larger than those of modified credits. This detail is because we have more unmodified credits than modified credits. Thus, we have more observations in this case than in the previous case. If we considered the same size of modified credits in this case, we would be in a situation where we would have too many observations per node and considering the objective of this plot, it would always be necessary to increase its size.

After observing Figure 5.7, we can verify that we have no empty neurons. According to what was previously described, we can conclude that compared to the modified credits, there is a greater number of observations in each neuron, since in this case, we have a scale that goes up to nearly 350 observations per neuron. This outcome is what we expected since, although the map size is larger than the case of modified credits, the fact is we have a greater number of observations in the case of unmodified credits. Nevertheless, we must not forget that this graphic should present a certain homogeneity or uniformity. By observing the plot, we can conclude that it fulfills that objective.

Neighbor Distance Plot. As expressed earlier, in this type of plot, the rationale we must follow is the intensity of color along with the values presented by the scale, concluding that the darker the neuron is, the closer the neuron is to the nearest neuron, indicating similar groups of nodes. The opposite also occurs. The Euclidean distance between neurons is presented in Figure 5.8.

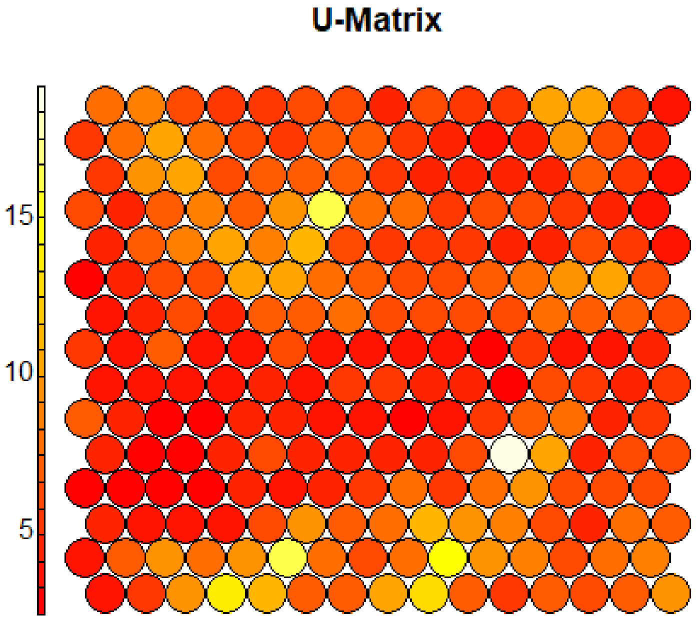

Figure 5.8 - U-Matrix of Unmodified Credits

Similar to the modified credits, we have nodes with approximately the same color. This detail means that we have neurons close to each other and, therefore, similar groups of nodes. Additionally, we have a reasonably similar scale compared to the one observed in the modified credits. As previously mentioned, this means that we have neurons that are close to each other. The nodes that tend more to the red color are closer to each other, while the ones that tend more to the yellow color are farther from each other. The main difference, in this case, is that we have more neurons that are closer to each other compared to the U-Matrix of modified credits.

Clustering. To conclude the SOM methodology, we have the construction of clusters related to unmodified credits. The process is the same as previously performed, so it is necessary to estimate the optimal number of clusters to consider, carried out by the examination of the "within sum of squares" plot, presented in Figure 5.9.

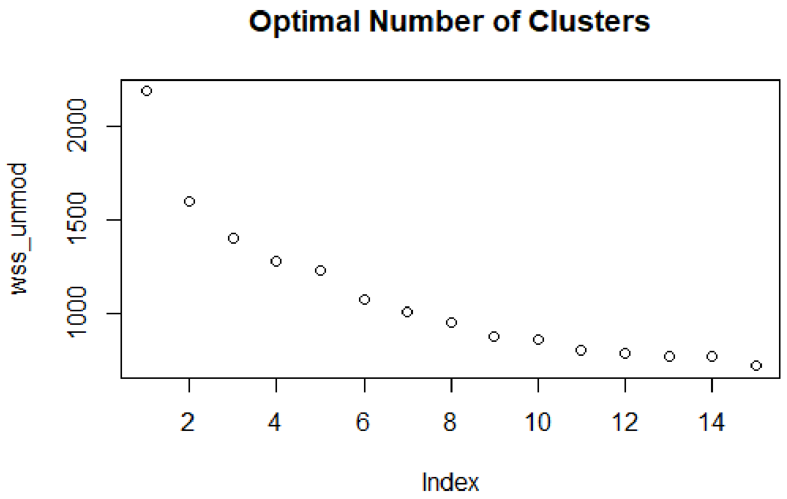

Figure 5.9 - Optimal Number of Clusters of Unmodified Credits

Following the rationale that helps us to interpret this graphic, as previously stated, we can observe that we have the “elbow point” on 4 clusters. Hence, it is at this point that we were able to identify an attenuation of the curve in the graphic above. Therefore, we must consider four distinct clusters in the case of unmodified credits. These clusters can be observed in Figure 5.10.

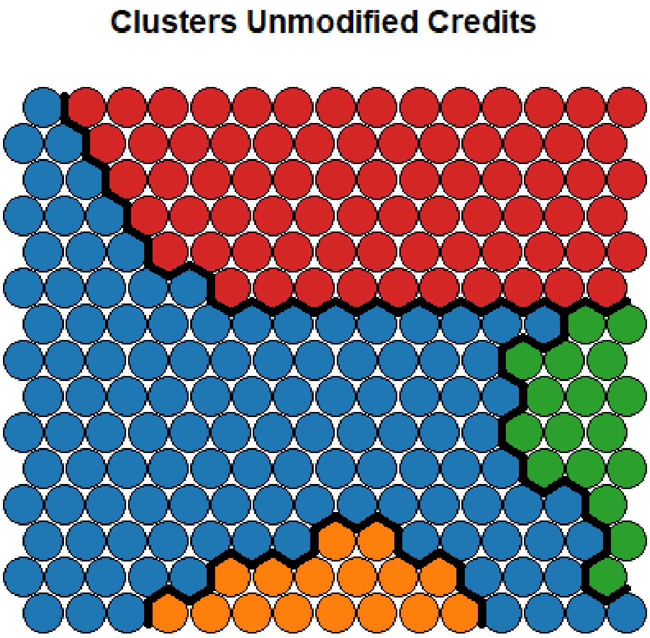

Figure 5.10 - Clusters of Unmodified Credits

In this plot we can observe the previously defined clusters. The red cluster contains 74 neurons, the blue cluster holds 118 neurons, the green cluster comprises 17 neurons, and the orange cluster comprises 16 neurons.

Now that the first methodology is completed, both for unmodified and modified credits, we are able to apply Markov chains to each cluster. However, just like in the case of modified credits, it is important to analyze each cluster in order to characterize them and then interconnect those characteristics with the results obtained from the application of the Markov chain methodology. Since we have four clusters, we will also have four different tables, each one containing the descriptive statistics of each group of individuals.

Starting by evaluating the first cluster, whose descriptive statistics are summarized in Table 5.5, we can verify that, only about 10% were first time home buyers, with the majority of contracts owned by one borrower – about 53%. However, just like in the previous situations, we also have a great percentage of contracts owned by two borrowers – about 42%. In fact, almost all the contracts are owned by one or two borrowers, with contracts owned by three, four, five or six borrowers representing less than one percent of all contracts.

Looking at the origination type, we can observe that the majority of these individuals acquired their credits through correspondent lending. However, we also have almost the same percentage of individuals that acquired their credits through retail lending, with only nearly a 3% difference. Regarding the conditions of the contracts, the origination interest rate ranged from 3,0% to 10,950%, averaging at 6,460%. The monetary property value has a mean value of $134 768, with a mode of $100 000, ranging from $8 000 to $800 000.

More related to the risk characteristics of individuals, the LTV ratio presents an interval between 1% and 97%. It is possible to verify that we have the lowest minimum value so far compared to the results from modified credits, which means that, in this cluster, we are facing individuals that represent a lower risk to the bank, regarding the percentage of the property appraisal value granted.

The DTI ratio of the fourth cluster is comprised between 1,00% and 64%, with a mean of 37,47%, which means the individuals included in this cluster apply 37,47% of their monthly income, on average, to pay their debts.

Finally, the credit score of individuals stands between 378 and 850. The first thing to notice is that we have a considerable interval. Additionally, we also have a low minimum value, and as we will be able to perceive, it is the minimum value among the clusters. However, if we look at the mode, we can observe a credit score higher than 650, which following the rationale on the interpretation of this variable, we can state that the majority of these individuals have a good credit history.

Table 5.5 - Descriptive statistics (Unmodified Credits – Cluster 1)

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate	3,000%	10,950%	6,460%	6,500%	6,500%
Original Loan Amount	8 000	800 000	134 768	100 000	116 000
Original Loan Term (months)	108	360	318	360	360
Original Loan to Value (LTV)	1%	97%	68%	80%	74%
Original Debt to Income (DTI) Ratio	1,00%	64,00%	37,47%	40,00%	38,00%
Borrower Credit Score at Origination	378	850	709	700	708
	Class			**Percentage**	
Channel / Origination Type	Broker			16,378	
	Correspondent			43,229	
	Retail			40,394	
Number of Borrowers	1			53,167	
	2			46,206	
	3			0,439	
	4			0,178	
	5			0,006	
	6			0,002	
First Time Home Buyer Indicator	No			89,565	
	Yes			10,235	

Observations: 64 619

Now evaluating the second cluster, whose descriptive statistics are summarized in Table 5.6, we can verify that only around 10% were first time home buyers, with 53,5% of the contracts owned by one borrower. However, similar to what we observed in the previous situations, we also have a great percentage of contracts owned by two borrowers, performing, together, 99,3% of contracts. This facet means that the contracts owned by three or four borrowers represent less than one percent of all contracts.

Looking at the origination type, we can see that there is close proximity between the percentage of credits acquired through correspondent lending and retail lending, since the percentages are almost 43% and approximately 41%, respectively. We can then conclude that, in this cluster, there is a major division between these two channel types, with the origination type broker only representing circa 16% of the contracts.

Regarding the contract conditions, the origination interest rate ranged from 4,375% to 8,750%, with a mode value of 6,50%. The monetary property value has a mode value of $100 000, ranging from $7 000 to $525 000.

More related to the risk characteristics of the individuals, the LTV ratio presents an interval between 6% and 97%. It is possible to verify that we have a higher minimum valuer compared to the results from the first cluster, which means that, in this cluster, we have some individuals that represent a slightly higher risk to the bank. However, if we look at the mean and mode values, we can observe that these values are equal to the ones previously verified.

Regarding the DTI ratio of the second cluster, we have a minimum value of 2,00% and a maximum value of 64%, with a mean of 37,47% and a mode of 36%, which means that most of the individuals included in this cluster apply 36% of their monthly income to pay their debts.

Finally, the credit score of individuals stands between 504 and 832. Even though we have a low minimum value, it is higher than the value presented in the first cluster. Additionally, if we look at the mode, we can observe a credit score with a value well over 650. Therefore, following the rationale on the interpretation of this variable, we can state that most of these individuals have a great credit history and should not display a significant risk to the bank.

Table 5.6 - Descriptive statistics (Unmodified Credits – Cluster 2)

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate	4,375%	8,750%	6,462%	6,500%	6,500%
Original Loan Amount	7 000	525 000	133 983	100 000	114 000
Original Loan Term (months)	120	360	319	360	360
Original Loan to Value (LTV)	6%	97%	68%	80%	74%
Original Debt to Income (DTI) Ratio	2,00%	64,00%	37,29%	36,00%	37,00%
Borrower Credit Score at Origination	504	832	710	733	709

	Class	Percentage
Origination Channel	Broker	15,846
	Correspondent	42,964
	Retail	41,189
Number of Borrowers	1	53,529
	2	45,764
	3	0,491
	4	0,216
First Time Home Buyer Indicator	No	89,739
	Yes	10,102

Observations: 6 929

In terms of the second cluster evaluation, whose descriptive statistics are summarized in Table 5.7, we can verify that, only almost 11% were first time home buyers, with most of the contracts owned by one borrower. However, identical to what we observed in the foregoing situations, we also have a great percentage of the contracts owned by two borrowers. Actually, almost all the contracts are owned by one or two borrowers, performing, together, about 99% of the contracts.

Looking at the origination type, we can observe that most of these individuals acquired their credits through correspondent lending, with a percentage of almost 44% of the contracts. However, the credits acquired through retail lending have a considerable significance, with a percentage of almost 40% of all contracts. We can then conclude that, in this cluster, there is a significant division between these two channel types, with the origination type broker only representing nearly 16% of the contracts.

Regarding the contract conditions, the origination interest rate ranged from 3,990% to 8,625%, with a mode value of 6,375%. The monetary property value has a mode value of $100 000, similar to the previous clusters, ranging from $12 000 to $645 000.

More related to the risk characteristics of the individuals, the LTV ratio presents an interval between 5% and 97%. Regarding the DTI ratio we have a minimum value of 1,00% and a maximum value of 64%, with a mean of 37,53% and a mode of 38%, which means that most of the individuals included in this cluster apply 38% of their monthly income to pay their debts. By comparing the results of this variable with the values of the previous clusters, it is possible to observe very similar values, which means that with respect to the DTI ratio, we are in the presence of similar individuals.

Finally, the credit score of individuals stands between 450 and 830. Even though we have a low minimum value, if we look at the mode, we can observe a credit score with a value greater than 650. Therefore, following the rationale on the interpretation of this variable, we can state that most of these individuals have a good credit history and should not represent a considerable risk to the bank.

Table 5.7 - Descriptive statistics (Unmodified Credits – Cluster 3)

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate	3,990%	8,625%	6,459%	6,375%	6,500%
Original Loan Amount	12 000	645 000	134 549	100 000	116 000
Original Loan Term (months)	108	360	318	360	360
Original Loan to Value (LTV)	5%	97%	68%	80%	74%
Original Debt to Income (DTI) Ratio	1,00%	64,00%	37,53%	38,00%	38,00%
Borrower Credit Score at Origination	450	830	710	684	710

	Class	Percentage
Channel / Origination Type	Broker	16,227
	Correspondent	43,944
	Retail	39,829
Number of Borrowers	1	53,031
	2	46,289
	3	0,504
	4	0,176
First Time Home Buyer Indicator	No	88,967
	Yes	10,857

Observations: 8 529

Finally, on the evaluation of the fourth and last cluster, whose descriptive statistics are summarized in Table 5.8, we can verify that, only about 10% were first time home buyers, with, again, the majority of the contracts owned by one borrower – 52,5%. Furthermore, similar to the other clusters, we also have a large percentage of contracts owned by two borrowers, with a value very close to that of contracts owned by one borrower – almost 47%. In fact, almost all the contracts are owned by one or two borrowers, with the contracts owned by three, four, five, or six borrowers representing only nearly one percent of all contracts. Even so, we can see that in unmodified credits, the percentage of individuals in this cluster who are first time home buyers is higher than the percentages verified in modified credits, which is transversal to all clusters of unmodified credits.

Looking at the origination type, we can observe that most of these individuals acquired their credits through correspondent lending. However, we also have almost the same percentage of individuals that acquired their credits through retail lending, with only a nearly 2% difference. Therefore, we can state that correspondent and retail lending are the origination channels that characterize more significance in all contracts of the fourth cluster, with a combined percentage of about 83%.

Regarding the conditions of the contracts, the origination interest rate ranged from 3,850% to 8,625%, averaging at 6,461%. The monetary property value has a range from $11 000 to $802 000, averaging at $135 128.

More related to the risk characteristics of the individuals, the LTV ratio presents an interval between 2% and 97%. It is possible to verify that it is the second lowest minimum value compared to the remaining clusters of unmodified credits, which means that, in this cluster, we are facing individuals that represent a low risk to the bank, regarding the percentage of the granted property appraisal value.

The DTI ratio of the fourth cluster is comprised between 1,00% and 64%, with a mean of 37,44%, which means the individuals included in this cluster apply 37,44% of their monthly income on average, to pay their debts.

Finally, the credit score of individuals stands between 455 and 833. Even though we have a low minimum value, if we look at the mode, we can observe a credit score with a value of 700, which is significantly greater than 650. Considering that this variable can only assume a maximum value of 850, we can state that this cluster is one of those that represent a lower risk for the bank in terms of the quality of credit history.

Table 5.8 - Descriptive statistics (Unmodified Credits – Cluster 4)

Variable	Minimum	Maximum	Mean	Mode	Median
Original Interest Rate	3,850%	8,625%	6,461%	6,500%	6,500%
Original Loan Amount	11 000	802 000	135 128	100 000	115 000
Original Loan Term (months)	120	360	318	360	360
Original Loan to Value (LTV)	2%	97%	68%	80%	74%
Original Debt to Income (DTI) Ratio	1,00%	64,00%	37,44%	37,00%	38,00%
Borrower Credit Score at Origination	455	833	709	700	708

	Class	Percentage
Channel / Origination Type	Broker	16,794
	Correspondent	42,656
	Retail	40,550
Number of Borrowers	1	52,544
	2	46,701
	3	0,561
	4	0,185
	5	0,007
	6	0,003
First Time Home Buyer Indicator	No	89,586
	Yes	10,226

Observations: 29 248

As with modified loans, a general evaluation is performed, comparing the clusters with each other. Thus, it is possible to construct different risk profiles for each group of individuals, illustrating the risk that they pose to the bank.

When comparing the four clusters, it can be concluded from the outset that regarding the number of borrowers, the origination type, whether or not individuals are first time home buyers and the loan term, the four clusters present approximately the same results. Thus, similar to the unmodified credits, we can ascertain that the clusters differ mainly in the borrowers' characteristics, specifically the LTV, DTI ratio, and Borrower Credit Score.

Regarding the LTV variable, considering that both the mean and mode are equal throughout each cluster, looking at the minimum values of each group of individuals, we can see that the first cluster has the lowest minimum value. Therefore, we can conclude that this cluster presents a lower risk to the bank regarding the percentage of the granted property appraisal value.

With respect to the DTI ratio, the mean and mode values differ between each cluster. Therefore, considering that the mean values do not differ from each other significantly, by looking at the mode values, we can see that the first cluster shows the greatest value, meaning that this cluster contains more individuals that pose a higher risk to the bank as the majority of them allocate 40% of their monthly income to servicing debt compared to the remaining clusters.

Lastly, regarding the Borrower Credit Score, we can see that the first cluster has the lowest mode and mean values, even though it presents a value above 650. Additionally, it also presents the lowest minimum value. This aspect means that not only does this cluster have more individuals with a poorer credit history, but it also has individuals with the worst credit history of all clusters. In opposition, we can verify that individuals in the second cluster have the best credit history, since not only does it have the highest average and mode values, but also the minimum value is the highest between all clusters.

In order to conclude this overall evaluation, taking into account what was mentioned above, we can conclude that the second cluster represents the least risk to the bank. However, as far as the variable that characterizes the percentage of the granted property appraisal value, this is the most vulnerable cluster. Nonetheless, given the characteristics of the history of individuals, this is considered the least critical cluster. On the other hand, the cluster that displays the highest risk is the first one, considering that, in the characteristics related to the history of individuals, namely, the DTI ratio and the borrower credit score, it is the most vulnerable cluster, even though it does not represent the highest risk in the LTV ratio, *vis à vis* with the remaining clusters.

In considering the main objective of this dissertation, it is necessary not only to perform a comparative evaluation between clusters of the same type of credit but also a comparative analysis between clusters of different types of credit, i.e., to compare the results obtained between modified and unmodified credits. Thus, a comparative analysis of the characteristics that distinguish these two types of credits is performed.

As mentioned earlier, some characteristics, namely, the number of borrowers, whether or not individuals are first time home buyers, the loan term and the origination type, do not differ from cluster to cluster. This dimension not only occurs between clusters of the same credit type but also between clusters of modified and unmodified credit. Consequently, we can see that considering the individuals' data of origin, the modified and unmodified credit clusters differ in the loan amount, the LTV, interest rate, DTI ratio, and Borrower Credit Score.

Regarding the loan amount, we can see that in modified loans there are much higher values, with a mean value around $210 000 and a mode value of $417 000, while in unmodified loans the mean value is about $134 000 and the mode value is $100 000. This trait is justified by the fact that in modified credits, it is necessary to grant a greater part of the value of the property, exhibited by the LTV, which, as mentioned above, illustrates the percentage of the property appraisal value covered by the loan. Although the mode value is the same, in this variable we can observe higher average values in modified credits in about 5%, as well as the minimum value, which is always higher than in unmodified credits.

As far as interest rates are concerned, we can see that they are relatively identical, since the intervals between the minimum and maximum values do not differ significantly, and the average and mode values are also quite similar. Nevertheless, it is possible to identify that the minimum interest rate values of unmodified credits are relatively lower than the values observed in modified credits. This aspect is because, based on an individual's credit history, certain contract conditions may change, as is the case of the interest rate. If an individual has a less favorable credit history, then the interest rate is likely to be higher than it would be if that individual had a better credit history. Thus, as can be confirmed in the Borrower Credit Score, modified loans have worse credit histories, so higher interest rates would be expected in these cases. For the credits understudy, in unmodified credits, not only is the credit score higher than 650 in three of the four clusters, but also in two of them, most individuals have a credit score of 700 or higher, which reveals a very low risk. In the case of modified credits, although the observed values are not unfavorable, it is possible to verify that they are always lower than those identified in unmodified credits. Additionally, one of the four clusters presents a credit score below 650.

Finally, concerning the percentage of monthly income that individuals allocate to their debts, we can note that the clusters whose credits have been modified exhibit higher values, with an average difference of 6% for credits whose conditions have not been modified. This variable not only influences the assignment of the credit score, which, as can be noted, is higher when the DTI ratio is also higher, but also shows a higher risk for the bank, as individuals in modified credits allocate more of their monthly income to servicing debt, which may be due to a lower monthly income but also to more debt.

5.2. Markov Chains

In this section, the results obtained from the application of the Markov chains methodology it will be presented. After estimating the TPMs, we will be able to observe if there are differences between the probability of default of modified and unmodified credits, and thus conclude whether the modifications are effective. Additionally, we will also broach the mean absorption times and the probabilities of absorption of each cluster. In the next tables, we will articulate the transition probabilities between the following states:

$$X_{jt} = \begin{cases} 1, & Normal\ Performance \\ 2, & Delinquency \\ 3, & Pre-Default \\ 4, & Default \\ 5, & Prepayment \\ 6, & Third\ Party, Short\ or\ Note\ Sales/REO \end{cases}$$

5.2.1. Modified Credits

Regarding the modified credits, we will have four TPM, since, in the development of the SOM approach, we obtained four different clusters. Thereby, we present the probabilities calculated by each cluster.

Table 5.9 - Transition Probability Matrix (Modified Credits - Cluster 1)

States	1	2	3	4	5	6
1	0.998794	0.000584	0.000008	0.000000	0.000458	0.000156
2	0.039157	0.946084	0.010843	0.000000	0.000000	0.003916
3	0.064846	0.010239	0.918089	0.003413	0.000000	0.000000
4	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
5	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Table 5.10 - Transition Probability Matrix (Modified Credits - Cluster 2)

States	1	2	3	4	5	6
1	0.998791	0.000605	0.000005	0.000000	0.000449	0.000150
2	0.044352	0.940015	0.009108	0.000000	0.000174	0.006351
3	0.052511	0.004599	0.934074	0.001150	0.000000	0.007666
4	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
5	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Table 5.11 - Transition Probability Matrix (Modified Credits - Cluster 3)

States	1	2	3	5	6
1	0.998785	0.000605	0.000003	0.000400	0.000159
2	0.044828	0.940439	0.008150	0.000131	0.006270
3	0.059761	0.000000	0.932271	0.000000	0.007968
5	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	1.000000

Table 5.12 - Transition Probability Matrix (Modified Credits - Cluster 4)

States	1	2	3	5	6
1	0.998787	0.000646	0.000003	0.000420	0.000148
2	0.041422	0.938783	0.011364	0.000000	0.008431
3	0.041298	0.008850	0.946903	0.000000	0.002950
5	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	1.000000

Before we compare the probabilities obtained from modified and unmodified credits, so we can evaluate if the modifications are actually effective or not, it is important to relate what was deduced through the descriptive statistics with the results obtained from the application of the Markov chains. After evaluating the descriptive statistics one by one, it was concluded that the cluster with the lowest risk for the bank was the second cluster, and the one with the highest risk was the third cluster. Furthermore, we must notice that in clusters three and four, we do not have state 4. This result is due to the fact that, in the period of analysis, these two clusters do not contain this state. This circumstance means that after 2015 this state, as well as transitions to other states, may exist. Nonetheless, the situation does not occur in the period under analysis.

Looking at the estimated results, we can understand that the probabilities related to state recoveries (i.e., retracting one or more states) and prepayments (state 5) are not always higher in the second cluster. Similarly, it is not possible to state that the probabilities related to transitions to following states and to forced credit terminations (state 6) are always lower. It is important to note that the risk that each group of clients represents is assessed according to its original characteristics. This element means that the level of risk initially determined may change after the loan conditions are modified.

Despite all this, we can see that the probability of a mortgage being prepaid – transition to state 5 – is practically always greater in the second cluster, excepting the transition from state 1 to state 5, that is higher in the first cluster, corroborating what was mentioned above regarding the risk that this cluster represents to the bank. By contrast, the probability of a credit being prepaid in cluster three is always the lowest when compared to the remaining cases (excepting in the transition from state 2 to state 5 in the first cluster, since there are no transitions between these two states).

Additionally, we can verify that regarding the probability of a credit being forced to termination – transition to state 6 – the third cluster almost always presents the greatest values. It is possible to see that the probability of transition from state 1 to state 6 is the lowest in the second cluster and it is the highest in the third cluster. Furthermore, the probability of transitioning from state 2 to state 6 differs from what has been observed so far, since the cluster with the lowest value is the first cluster and the one with the highest value is the fourth cluster. In addition, the probability of transitioning from state 3 to state 6 is not the lowest in the second cluster but it is the highest in the third cluster. Finally, we do not have probabilities regarding the transition from state 4 to state 6 since there are no transitions between these two states in any of the clusters.

Concerning the state recoveries, it is important to mention that the probabilities not only reflect the risk that the individuals represent to the bank, but they also incorporate a modification in the risk level when the credits are modified. Hence, it is understandable that the estimated probabilities do not exactly reflect the risk level previously assessed, that is, cluster two does not present the greatest recovery probabilities, and cluster three does not present the lowest recovery probabilities.

In addition to evaluating the obtained results regarding the TPM, it is also fundamental to evaluate the mean absorption times and absorption probabilities, since it allows us to perform a complementary evaluation of the behavior of the individuals inserted in each cluster. Thus, as stated above, we will present these results and their respective evaluation.

Before evaluating the results obtained, it is important to highlight two exceptional situations regarding the modified credits. Clusters 1 and 2 exhibit state 4 as a recurrent state. This status has a similar justification to the one provided to explain why clusters 3 and 4 do not contain state 4. In clusters 1 and 2, there is no transition from state 4 to any other state because of the period of analysis, which extends to 2015. This scenario means that after 2015 this transition may exist, but in the period under analysis this does not occur. Thus, it is stated that state 4 is not absorbing, although in these two clusters, it is considered that way. Regarding clusters 3 and 4, we have a situation in which state 4 does not occur at all (i.e., there are no credits with 90 days or more past due). For this reason, we consider these situations to be exceptional.

In Table 5.13, we can observe the absorption probabilities of each cluster, i.e., the probability of any transient state – 1, 2, and 3 – to be absorbed by any absorbing state – 4 (in the case of the first and second clusters), 5 and 6. The results denoting the mean absorption times, i.e., expected number of steps to move from any of the transient states to any of the recurrent states, are shown in the annexes with their analysis later in this chapter.

Table 5.13 - Absorption Probabilities (Modified Credits)

Cluster 1	States	4	5	6
	1	0.0080327	0.6849833	0.3069840
	2	0.0158921	0.6221768	0.3619311
	3	0.0500124	0.6200505	0.3299371
Cluster 2	States	4	5	6
	1	0.0025075	0.6638867	0.3336058
	2	0.0048569	0.5802096	0.4149335
	3	0.0197779	0.5692732	0.4109489
Cluster 3	States		5	6
	1		0.6276723	0.3723277
	2		0.5534562	0.4465438
	3		0.5538285	0.4461715
Cluster 4	States		5	6
	1		0.6302488	0.3697512
	2		0.5339698	0.4660302
	3		0.5791885	0.4208115

As explained above, the first and second clusters consider state 4 as an absorbing state whereas in the third and fourth clusters this state does not exist. For this reason, we have probabilities of absorption from states 1, 2, and 3 for states 4, 5, and 6 in the first clusters and only absorption probabilities from states 1, 2, and 3 for states 5 and 6 in the last clusters. For this reason, for the purpose of a comparative assessment and because state 4 is not, in fact, an absorbing state, we will not consider the probabilities of absorption for this state in relation to the first two clusters.

As the number of days past due in the credit increases, that is, as we move from state 1 to state 3, we notice a decrease in the probabilities of absorption in state 5 and an increase in the probability of absorption in state 6. This outcome is what we expected, since it is understandable that as the payment arrears of credits increase, the aptitude to prepay a credit decreases, and the possibility of the bank being forced to end the credit in order not to suffer great losses increases.

Starting from state 1, there is a probability of absorption in state 5 of about 0.68 and 0.66 in first and second clusters, respectively, and about 0.63 in the third and fourth clusters. Regarding the absorption in state 6, we can observe that these probabilities exhibit values of about 0.31 and 0.33 for the first and second clusters and around 0.37 for the third and fourth clusters.

Respecting state 2, here, the second row tells us that the absorption probability in state 5 is about 0.62 in the first cluster, 0.58 in the second cluster, 0.55 in the third cluster, and finally, about 0.53 in the fourth cluster. Regarding the absorption in state 6, we can see that there is a probability of 0.36 of absorption in the first cluster, about 0.41 of absorption in the second cluster, about 0.45 of absorption in the third cluster, and 0.47 of absorption in the fourth cluster.

Finally, in respect of state 3, we can observe that, in the first cluster, the probability that the chain will be absorbed in state 5 has a value of 0.62, a value of about 0.60 in the second cluster, a value of roughly

0.55 in the third cluster and a value of circa 0.58 in the fourth cluster. As we can observe, these are, in general, the lowest verified so far. Regarding the absorption in state 6, we can state that starting in state 3, there is a probability of about 0.33 of absorption in the first cluster, almost 0.41 of absorption in the second cluster, nearly 0.45 of absorption in the third cluster and, to conclude, probability about 0.42 of absorption in the fourth cluster.

In conclusion, these results were what we expected, considering the risk assessment previously performed from the descriptive statistics, as well as the evaluation carried out on the results obtained in the TPM. Comparing these last results, we can see that the cluster that presents the lowest absorption probability in state 5, starting in any of the states (1, 2, or 3), is, in general, cluster 3, which was previously identified as the one that represents the greater risk. Additionally, the third cluster is also the one with the highest probability of absorption in state 6, starting in any of the states – with the exception of absorption in state 6 starting in state 2, which is higher in the fourth cluster.

On the other hand, although the second cluster – previously identified as the one that represents the lowest risk for the bank – does not present the highest probability of absorption in state 5 and the lowest probability of absorption in state 6. It shows values close to the first cluster, which, in this case, has the highest probability of absorption in state 5 and the lowest probability of absorption in state 6, having been previously considered, after the second cluster, a cluster with a relatively low risk. Thus, these latest results and their respective evaluations are in line with those performed before.

5.2.2. Unmodified Credits

With respect to the modified credits, we will also have four TPMs, considering the results obtained from the SOM methodology. Thus, we present the probabilities obtained in each cluster.

Table 5.14 - Transition Probability Matrix (Unmodified Credits - Cluster 1)

States	1	2	3	4	5	6
1	0.998104	0.000279	0.000000	0.000000	0.001527	0.000090
2	0.001792	0.965812	0.020985	0.000000	0.000295	0.011116
3	0.000447	0.000596	0.973261	0.005363	0.000298	0.020036
4	0.000000	0.000000	0.000000	0.977839	0.000000	0.022161
5	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Table 5.15 - Transition Probability Matrix (Unmodified Credits - Cluster 2)

States	1	2	3	4	5	6
1	0.998077	0.000297	0.000000	0.000000	0.001541	0.000086
2	0.002828	0.965630	0.020883	0.000000	0.000000	0.010659
3	0.000000	0.000000	0.970381	0.006347	0.000000	0.023272
4	0.000000	0.000000	0.000000	0.947368	0.000000	0.052632
5	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Table 5.16 - Transition Probability Matrix (Unmodified Credits - Cluster 3)

States	1	2	3	4	5	6
1	0.998117	0.000266	0.000002	0.000000	0.001517	0.000099
2	0.001693	0.966193	0.021027	0.000000	0.000328	0.010759
3	0.000724	0.000724	0.972303	0.005974	0.000181	0.020094
4	0.000000	0.000000	0.018072	0.951807	0.000000	0.030120
5	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

Table 5.17 - Transition Probability Matrix (Unmodified Credits - Cluster 4)

States	1	2	3	4	5	6
1	0.998075	0.000293	0.000000	0.000000	0.001538	0.000093
2	0.001514	0.965374	0.020246	0.000000	0.000378	0.012488
3	0.000000	0.000000	0.978337	0.004098	0.000000	0.017564
4	0.000000	0.000000	0.000000	0.967742	0.000000	0.032258
5	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000

After evaluating the descriptive statistics one by one, it was concluded, based on the descriptive statistics, that the cluster with the lowest risk for the bank was the second cluster, and the one with the highest risk for the bank was the first cluster.

Nevertheless, looking at the calculated probabilities, we can understand that the probability of a credit being prepaid is not always greater in the second cluster, as well as the probability of state recoveries. In fact, we can observe that the probability of a credit being prepaid is not always greater in a particular cluster, just like the probability of a credit being forced to termination or the probability of performing a state recovery. Regarding the prepayment of credits, we are in a situation where the probability of transition from state 1 to state 5 is higher in the second cluster, however, when we talk about the probability of transition from state 2 to state 5, we verify a higher value in the fourth cluster. In addition, the probability of transition from state 3 to state 5 is higher in the first cluster. With respect to state 4, there are no transitions from this state to state 5 in any of the clusters.

We can also verify that concerning the probability of a credit being forced to terminate, the second cluster presents the lowest probabilities regarding the transitions from states 1 and 2, subsequently representing a low risk for the bank as far as these two transitions. However, regarding the transitions from states 3 and 4, the second cluster presents the highest probabilities. In this case, we have cluster four with the lowest probability of transition from state 3 to state 6, and cluster one with the lowest probability of transition from state 4 to state 6.

With respect to the state recoveries, firstly, we can verify that recovering from state 2 to state 1 is most probable in the second cluster, which corroborates what was previously stated regarding the risk level of this cluster. Additionally, transitioning from state 3 to state 2 is most probable in the third cluster and least probable in the second and fourth clusters, as well as transitioning from state 3 to state 1 and from state 4 to state 3.

Finally, regarding the transition to the next state – a transition from state 1 to state 2, for example – we can verify that the clusters that represent the lowest risk to the bank are the third and fourth clusters, considering that these two clusters present the lowest probabilities. The probability of transitioning from state 1 to state 2 is lower in the third cluster, while the transitions from state 2 to state 3 and from state 3 to state 4 are less likely in the fourth cluster.

To conclude this evaluation, we can see that the probabilities calculated under the Markov chain methodology do not perfectly reflect the prediction of the risk level based on the statistics of each cluster. However, we can see that the first cluster remains to represent the highest risk to the bank. In addition, the low-risk level is slightly divided by the remaining three clusters, focusing mainly on clusters two and three. Although in the prediction of the risk level of each cluster based on the descriptive statistics the second cluster was considered to represent the lowest risk to the bank, this situation is quite understandable, since it is possible to verify that these two clusters have very similar origin characteristics, with just a few minor differences.

Similarly to what was performed in the modified credits, we will also evaluate the mean absorption times and the absorption probabilities, thus carrying out a complementary evaluation to the results obtained regarding the TPM.

In Table 5.18 we can observe the absorption probabilities of each cluster, i.e., the probability of any transient state – 1, 2, 3 or 4 – to be absorbed by any absorbing state – 5 and 6. As with modified credits, the results of the mean absorption times are presented in the annexes, and their analysis is found later in this chapter.

Table 5.18 - Absorption Probabilities (Unmodified Credits)

Cluster 1	States	5	6
	1	0.81505239	0.1849476
	2	0.06747632	0.9325237
	3	0.02626776	0.9737322
	4	0.00000000	1.0000000
Cluster 2	States	5	6
	1	0.81130872	0.1886913
	2	0.06675325	0.9332468
	3	0.00000000	1.0000000
	4	0.00000000	1.0000000
Cluster 3	States	5	6
	1	0.81554568	0.1844543
	2	0.07063712	0.9293629
	3	0.03231809	0.9676819
	4	0.01211928	0.9878807
Cluster 4	States	5	6
	1	0.80605362	0.1939464
	2	0.04616628	0.9538337
	3	0.00000000	1.0000000
	4	0.00000000	1.0000000

Evaluating these results, we can verify from the outset that, just like in the case of modified credits, the absorption probabilities in state 5 decrease as the number of days past due increase. Similarly, the absorption probabilities in state 6 increase as the number of days past due increase, which is comprehensible, since there is a relationship between the number of days past due in credits and the probability that they will be absorbed, in either of the two absorbing states.

Starting from state 1, we can verify that there is a probability of absorption in state 5 of about 0.81 in all clusters. Regarding the absorption in state 6, we can affirm that values are also very close to each other in all clusters, with a probability about 0.18 of absorption in the first and third clusters and a probability of absorption around 0.19 in the second and fourth cluster.

Respecting state 2, we can observe a significant reduction in the obtained results when comparing the absorption probabilities starting in state 1. In this case, we have probabilities about 0.07 of absorption in the first, second and third clusters and probability 0.05 of absorption in the fourth cluster. Regarding the absorption in state 6, we observe a probability of absorption of about 0.93 in the first, second and third clusters and a probability of absorption of about 0.95 in the fourth cluster.

In what respects the state 3, we have a different situation. In the second and fourth clusters, we do not have absorption probabilities in state 5, which is justified by the fact that we do not have probabilities of transition from state 3 to state 5 in the TPM of these two clusters. However, in the first and third clusters, we can verify a probability of absorption of about 0.03. Regarding the absorption in state 6, we can observe values of 1 in the second and fourth clusters – since we do not have absorption probabilities in state 5 – and probabilities of absorption of about 0.97 in the first and third clusters.

Finally, starting from state 4, we have a situation where we do not have absorption probabilities in state 5 in the first, second and fourth clusters. Nevertheless, in the third cluster, we can observe a probability of absorption of about 0.03. Regarding the absorption in state 6, in the first, second, and fourth clusters, we have probabilities of absorption of 1 – similar to the situation in the probabilities of absorption starting in state 3 – and a value of circa 0.98 in the third cluster.

To conclude these last results, similarly to what was observed in the evaluation of TPM results, the absorption probabilities do not exactly reflect the level of risk previously predicted. However, it is possible to observe that the absorption probabilities do not vary significantly from cluster to cluster. Therefore, we can say that the results are aligned with the risk evaluation initially performed, which is corroborated by the interpretation of the results in the estimated TPM.

After an analysis of the results obtained in each credit category (modified and unmodified), considering the main objective of this dissertation, it is necessary to compare the results obtained between modified and unmodified credits. Only then can we deduce some conclusions, thus respecting the purpose of this dissertation. The first thing to note is that there are no transitions from states 5 and 6 as these states are absorbing. This repercussion was expected considering that these two states correspond to situations where the credit is reduced to zero, i.e., it terminates.

By comparing the results, the first event to note is the difference between the probability of remaining or transitioning to state 1, which corresponds to the normal performance state. It is possible to observe that the results obtained do not significantly differ between modified and unmodified credits, although there are slightly higher values in credits that were modified.

Examining the events related to state 5, which, as mentioned above, corresponds to a situation where an individual anticipated the payment of credit (prepayment), we can conclude that the probability of moving to this state from any other state, is always greater in unmodified credits. This culmination is what we expected, as the modifications are mostly reflected in the maturity of the contract. When a loan is modified, its maturity always tends to increase relative to the initial conditions, coupled with the fact that the fees associated with the loan – predominantly the interest rate – in their total, even though by installment, which is monthly, decrease in order to facilitate the payment of the loan.

Regarding state 6, we can conclude that the probability of transitioning to this state, i.e., the probability that the bank is forced to terminate the contract due to consecutive payment failures, is lower in almost all clusters in mortgages that suffered modifications than in loans which were never modified, reaching a probability of zero in some cases.

Analyzing the remaining states, that is, the states that correspond to the proceeding of contracts, firstly we can notice that the probability of retracting one or more states, interpreted as the recovery from late payment of contracts, is comparatively higher in modified credits than in credits whose conditions have not changed since their inception. Lastly, we can perceive that the probability of moving from state 2 (delinquency) to state 3 (pre-default) and moving from state 3 to state 4 (default) is relatively lower in modified credits, even if regarding state 1 (current/normal performance), the probability of transitioning to state 2 is lower in unmodified credits.

Overall, we can see that the probabilities related to the transition from states that correspond to a normal performance or a short delay in the payment of contracts to states that show a longer delay in the payment of credits are lower in loans that were modified than on unmodified mortgages. Similarly, the probability of a forced termination of a loan is always lower in modified credits than in contracts whose conditions have never changed.

Regarding the mean absorption times, in the appendix, we can first verify that the modified credits take significantly more steps to being reduced to zero, i.e., to reach an absorbing state, compared to the unmodified credits. Additionally, we can observe that we only have mean absorption times for state 4 in unmodified credits due to the reasons previously stated.

In the modified credits, it takes about 1500 steps for the credit to transition from state 1 (current or normal performance) to being reduced to zero. This status, either occurring by being prepaid or by a situation of a third party, short or note sales, except for the second cluster, where it takes almost 1600 steps, contrasting with the unmodified credits, that present a number of steps between about 530 and 545. Regarding the transition from the second state to states 5 and 6, we can observe that, in modified credits it takes approximately 1300 steps in clusters 2 and 4, and almost 1400 steps in clusters 1 and 3 while in unmodified credits it takes a number of steps between approximately 80 and 95. Finally, regarding the transition from state 3 to states 5 and 6 in modified credits, there are more different situations from cluster to cluster, since, in the first cluster, there are around 1390 steps, while in the second cluster this result does not reach 1300, being about 1290 steps. In the third and fourth clusters, the results exhibit a value of slightly more than 1400 steps. On the other hand, in unmodified credits, we can verify that, in the first cluster, it takes about 58 steps to reach an absorbing state, and in the second cluster, it takes about 38 steps. Additionally, we can observe an average of about 62 to go from the third state to states 5 and 6 in the third cluster, and an average of 52 steps on the fourth cluster. Finally, regarding the transition from state 4 to states 5 and 6, we can notice that in the first cluster and third clusters, it takes approximately 45 steps, while in the second cluster this result is lower, presenting a value of an average of 19 steps to reach an absorbing state. In what concerns to the fourth cluster, we have an intermediate value of 31 steps.

Comparing these results, we can see that regarding state 1, modified credits have three times the steps of those verified in unmodified credits. In state 2, we can observe about 15 times more steps in modified credits and about 30 times more steps in state 3 in relation to unmodified credits. This exceptionally significant difference between modified and unmodified credits is not necessarily justified by the fact that in modified loans, it truly takes longer to reach an absorbing state than in unmodified credits. However, it may mean that there are many more transitions for these states in absolute terms in unmodified credits. It is understandable to think that, when we evaluate the number of steps in a universe in which the majority of individuals have not transitioned to any absorbing state, the result will be a much higher number of steps than in a universe where a large part of individuals has transitioned to an absorbing state. That is a result of a weighting of individuals who, in fact, have transitioned and those who have not, which are significantly more.

Concerning the absorption probabilities, the first thing we can notice is that both modified and unmodified credits have the same behavior as the days past due increase – both types of credits exhibit a decrease in the absorption probabilities in state 5 and an increase in the absorption probabilities in state 6. Nevertheless, there are some differences in these results that are quite noticeable. When we talk about absorption probabilities in state 5, we can verify that starting in state 1, the unmodified credits display much higher values than the modified credits. Nevertheless, it is possible to observe that the situation is reversed starting in states 2 and 3: modified credits reveal that the probability that the chain will be absorbed in state 5 is substantially higher than in unmodified credits. Regarding the probability of the chain being absorbed in state 6, we can observe a parallel situation to the one verified in relation to the probability of absorption in state 5. In this case, we can observe that, starting in state 1, the probability of absorption in state 6 is always higher in modified credits than in unmodified credits. On the other hand, when we start in one of the other states, the probability that the chain will be absorbed in state 6 is always higher for unmodified credits.

Before we can affirm that the changes were effective, and because the results in the modified credits did not turn out to be significantly better, it becomes necessary to have statistical evidence on this matter. In this way, homogeneity tests were performed, in order to conclude if the TPM of modified credits and the TPM of unmodified credits are not homogeneous, evidencing, in that case, success in the implementation of the modifications. These results can be observed in Table 5.19.

Table 5.19 - Homogeneity Tests

$\text{Clusters}_{\text{Modified,Unmodified}}$	ChiSq Statistic	d.o.f.	p-value
$\text{Clusters}_{1,1}$	0.1923401	35	1
$\text{Clusters}_{1,2}$	0.2423116	35	1
$\text{Clusters}_{1,3}$	0.2269922	35	1
$\text{Clusters}_{1,4}$	0.2142722	35	1
$\text{Clusters}_{2,1}$	0.1687174	35	1
$\text{Clusters}_{2,2}$	0.2176723	35	1
$\text{Clusters}_{2,3}$	0.2041316	35	1
$\text{Clusters}_{2,4}$	0.1890173	35	1
$\text{Clusters}_{3,1}$	1.446257	35	1
$\text{Clusters}_{3,2}$	1.448682	35	1
$\text{Clusters}_{3,3}$	1.445312	35	1
$\text{Clusters}_{3,4}$	1.448133	35	1
$\text{Clusters}_{4,1}$	1.429183	35	1
$\text{Clusters}_{4,2}$	1.436412	35	1
$\text{Clusters}_{4,3}$	1.427905	35	1
$\text{Clusters}_{4,4}$	1.433984	35	1

These results reveal that we have the same conclusion for the homogeneity test between any cluster. Considering that we have a hypothesis test in which the null hypothesis represents homogeneity and the alternative hypothesis represents no homogeneity among clusters, such as

$$H_0: There\ is\ homogeneity \qquad versus \qquad H_1: There\ is\ no\ homogeneity \tag{5.17}$$

and also considering that we reject the null hypothesis (H_0) for a p-value lower than an alpha value (α), we can conclude that, always having a p-value equal to 1, we do not reject the null hypothesis.

Thus, we can conclude that we cannot affirm that there is no homogeneity between the clusters of modified and unmodified credit. In other words, we can state that there is no statistical evidence that the modifications were indeed effective.

Once the evaluation of the results obtained between each cluster is completed and also after a comparison between modified and unmodified credits, it is essential to identify the risk factors and the alarming characteristics exhibited by the groups of clients based on the high-risk characteristics that were identified in each cluster. This step will enable us to identify a pattern in order for these risk factors to serve as a reference for better risk management.

Initially, we can identify from the outset that the characteristics of the individuals that incorporate the modified credit clusters constitute risk factors. This element is understandable since, as the conditions of the credits are renegotiated and modified, this means that this was the most viable way for both parties in the event of default by the customer. Thus, one of the risk factors that we must identify relates to a characteristic that the individuals represent (in contrast to characteristics relative to contract conditions) – the DTI ratio. Regarding this characteristic, we can verify that a high value in this variable should be considered a risk factor. More specifically, it is noted that when this variable presents a value that starts at 40%, the individuals represent a higher risk. This aspect is possible to notice in modified credits, considering that both the average and mode values of this variable are above 40%, and in unmodified credits, the cluster with the highest risk – cluster one – has a mode value of 40%.

Still, in relation to the characteristics of clients, we can identify that, regarding the credit history of individuals – the Borrower Credit Score – when it presents lower values, the individuals demonstrate a higher tendency to occupy states that are characterized by the infringement of the contractual obligations of loans. However, in this variable, the limits that reveal a higher or lower risk are not so clear, and in most cases, they translate higher or lower risk when associated with other variables concerning the individuals' characteristics.

The third risk factor is regarding the LTV ratio, the variable that reflects the part of the property appraisal value that the loan covers. Although most individuals have an LTV ratio of 80%, considering that this is the upper limit that most banks impose, this variable should be a risk factor when individuals are in groups where the lower limit is higher, as these groups are more likely to move to states with longer days past due.

The fourth and final risk factor is related to the conditions of the contract, namely the loan amount. The first thing to point out is the considerable difference between the loan amounts of modified and unmodified credits. We can verify that the loan amounts of modified credits are significantly higher than the ones observed in unmodified credits. Considering that modified loans generally pose a higher risk – otherwise their conditions would not have been renegotiated – and perceiving that their amounts are notably higher, it is important to identify high loan amounts as a factor that must require special awareness from banks, especially when the loan amount is greater than $200 000. This argument is reinforced by the fact that in unmodified loans the least risky clusters – clysters two and three – are characterized by lower loan amounts.

Thus, we can see that these four risk factors should require special attention from banks. In the case where these variables present particular values, namely a DTI ratio greater than 40%, a low Borrower

Credit Score, a high LTV ratio or significantly high loan amounts, then they should alert banks as well as insurers, when applicable.

6. CONCLUSIONS

This dissertation proposed an innovative hybrid approach: the use of neural networks, Self-Organizing Maps, as a basis to estimate the Markov chains in the context of mortgage loans. This interdisciplinary innovation represents a practical and applicable solution for credit risk management for the banking sector. Therefore, the results obtained in this dissertation are relevant to those involved in the process of risk management and mitigation. The hybrid methodology developed in this study allowed us to understand whether credit modifications are effective. Accordingly, this represents an advantageous solution for banks when they face situations in which their clients have consecutive payment failures of their obligations.

We analyzed a total of 149 404 loans acquired by Fannie Mae in 2006, divided into 40 079 modified credits and 109 325 unmodified credits, following their performance until 2015. In the first methodology applied, the division into the various clusters that were obtained was based on the original information of the contracts. Beyond this, the SOM methodology, allowed us to identify which credits were modified, which was fundamental for the fulfillment of the objective of this dissertation. In addition, it allows banks to predict, at the origin of the contract, based on the behavior of past credits with similar original characteristics, the behavior that new credits will have, and thus be able to guard against such situations, as well as the possibility to assess the likelihood that a particular credit will need to be modified in its procedure.

The application of the Markov chains approach allowed us to evaluate the impact of modifications and, therefore, understand if the interventions were more or less successful in the different clusters. Within all the transitions studied, we can conclude that the transition to more advanced states related to the performance of individuals while credits are active – states 1, 2, 3, and 4 – is less likely in credits that have changed regarding states 2, 3, and 4. On the other hand, that transition is less likely in credits whose contractual conditions are the same from inception regarding state 1. It was also possible to realize that the probability of individuals to transition from states with more late payments to states with less or no late payments – retracting one or more states – is greater when credits are modified than unmodified credits.

Regarding the two absorption states – states 5 and 6 – we have two different conclusions. The probability of transition from any state to state 5 – the probability of a credit being prepaid – is always higher in unmodified credits. However, this case was expected, considering that, as previously mentioned, the modifications are mostly reflected in the maturity of the contract. Regarding state 6, we were able to conclude that the probability of a credit being reduced to zero by the bank due to consecutive payment failures is lower in almost all clusters in modified credits.

Nonetheless, since the results in modified credits were not severely better than in unmodified credits, homogeneity tests were implemented in order to guarantee statistical evidence on this matter. We were able to conclude that, although we can observe some improvements in credit performance after the changes are implemented, we can not affirm that the modifications were effective since there is no statistical evidence in this sense.

Although it was not possible to positively answer the great question of this dissertation, we can understand that modifications might represent a useful tool for the banking sector to protect themselves from situations in which they might lose the capital granted to certain non-compliant customers. However, it is crucial to note that the modifications should not be a solution for all credits in order to facilitate payment by customers solely. This measure should instead be an exceptionally well deliberated and studied solution by financial institutions, to consider in a situation where there is a total incapacity to comply with contractual obligations and in the absence of such renegotiations, the bank's only answer is to enter into due diligence or to be obliged to terminate the credit, forfeiting all income arising therefrom.

Additionally, the application of the hybrid methodology allowed us to identify fundamental risk factors for the banks' risk management. These factors are essentially related to four variables: Borrower Credit Score, DTI Ratio, LTV Ratio, and Loan Amount. It was identified that borrowers that present a low credit score demonstrate a higher tendency to occupy states characterized by non-compliance of contractual obligations. However, the limits that reveal a lower or higher risk are not so clear, therefore being associated with other variables to understand the level of risk. It was also determined that borrowers that exhibit a high DTI ratio reveal a higher level of risk, more precisely when above 40%. The third variable allowed us to determine that borrowers represent a higher level of risk when they are incorporated in groups of clients where the lower limit is higher, since most individuals have an LTV ratio of 80%. The fourth and last variable allowed us to identify that credits with a loan amount higher than $200 000 have a higher probability to occupy states that represent failure of contractual obligations.

We can verify that the hybrid methodology that was developed in this dissertation is a proficient method for banks to map and predict the behavior that certain credits may have, as well as to classify their customers into groups at the time of undertaking credit contracts. Therefore, the realization of this study allowed us not only to understand that the modifications may not be effective but also to understand the behavior of different groups of clients and to identify important risk factors, being an outstanding contribution to the lack of studies that exist on this theme, a characteristic that was highlighted at the beginning of this work. In this way, it represents a breakthrough in terms of research on bank loans, as well as a step forward in terms of new, more complex, and complete methodologies to be applied in the business of banks.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This chapter presents the limitations that occurred during the development of this dissertation, as well as some options for future work that can be carried out following the developed methodology, and also extensions to that methodology.

Starting with the limitations, during the thesis development process, a limitation regarding the data was identified. As we were able to observe in the results obtained from the application of the SOM Markov chains methodologies, there are some cases where there are no significant differences from cluster to cluster. Since the SOM methodology is used to identify different groups of clients, based on their characteristics, this would then be a limitation, since the result obtained does not exhibit groups of clients absolutely different from each other. However, another matter that we were also able to verify is the fact that the application of this methodology allowed us to identify characteristics that distinguish higher and lower risk groups that, in the case of the absence of this methodology, we would not be able to identify. Therefore, despite this marginal limitation, we cannot fail to consider that the application of this methodology was fundamental for the pursuit of the objective of this dissertation.

Regarding further research, we present some ideas that could be of interest to investigate. The first suggestion for future work would be to apply the hybrid methodology developed in this dissertation – SOM methodology as the basis for Markov chains – to other types of credit, such as personal loans, auto credits or revolving credits, in order to implement a proper risk detection and mitigation tool in other types of products in the banking sector.

Another suggestion for future work would be an extension of the work developed in this dissertation. In this case, the objective would be to investigate which modifications are most effective and the impact that each modification has on the various types of credit.

An additional example of possible further research would be an extension to the Markov chains approach, with the application of Higher-Order Markov Chains (HOMC), which was not possible due to our data span and Multivariate Markov Chains (MMC) for the development of the objective of this dissertation.

8. BIBLIOGRAPHY

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4): 589–609.

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. Statistical Methods in Medical Research, 11 (2), 91–115

Asadabadi, M. R. (2017). A customer based supplier selection process that combines quality function deployment, the analytic network process and a Markov chain. European Journal of Operational Research, 263 (3), 1049–1062.

Berchtold, A. (2001). Estimation in the mixture transition distribution model. Journal of Time Series Analysis, 22 (4), 379–397.

Betancourt, Luis. (1999). Using Markov Chains to Estimate Losses from a Portfolio of Mortgages. Review of Quantitative Finance and Accounting, 12 (3), 303-317.

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardenes, J., Romance, M., Sendina-Nadal, I., Wang, Z., Zanin, M. (2014). The structure and dynamics of multilayer networks. Physics Reports, 544 (1), 1–122.

Bukiet, B., Harold, E. R., Palacios, J. L. (1997). A Markov chain approach to baseball. Operations Research, 45 (1), 14–23.

CEIC (2011). Retrieved 31 March 2020 from https://www.ceicdata.com/en/indicator/united-states/non-performing-loans-ratio

Chamboko, R. & Bravo, J. M. (2018). A multi-state approach to modelling intermediate events and multiple mortgage loan outcomes. The Journal of Real Estate Finance and Economics.

Ching, W.-K, Ng., M, K. (2016). Markov Chains: Models, Algorithms and Applications. International Series in Operations Research & Management Science, 1–169.

Collins, J., M, Reid, C. K. & Urban, C. (2017). Sustaining Homeownership After Delinquency: The Effectiveness of Loan Modifications by Race and Ethnicity. Cityscape: A journal of Policy Development and Research, 17 (1), 163-187.

Damásio, B., Mendonça, S. (2018). Modelling insurgent-incumbent dynamics: Vector autoregressions, multivariate markov chains, and the nature of technological competition. Applied Economics Letters, 1–7.

Damásio, B., Nicolau, J. (2013). Combining a regression model with a multivariate Markov chain in a forecasting problem. Statistics & Probability Letters, 90, 108–113.

Fannie Mae (2019). Fannie Mae Single-Family Loan Performance Data. Retrieved 30 September 2019, from http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html.

Ferles, C. & Stafylopatis, A. (2013). Self-Organizing Hidden Markov Model Map (SOHMMM). Neural networks: the official journal of the International Neural Network Society. 48C. 133-147. 10.1016/j.neunet.2013.07.011.

Gómez, S., Arenas, A., Borge-Holthoefer, J., Meloni, S., Moreno, Y. (2010). Discrete-time Markov chain approach to contact-based disease spreading in complex networks. EPL (Europhysics Letters), 89 (3).

Gottschau, A. (1992). Exchangeability in multivariate Markov chain models. Biometrics, 751–763.

Hougaard, P. Lifetime Data Anal (1999) 5: 239. https://doi.org/10.1023/A:1009672031531.

Leow, M. and Crook, J. (2014). Intensity models and transition probabilities for credit card loan delinquencies. European Journal of Operational Research, 236 (2), 685-694. doi:10.1016/j.ejor.2013.12.026.

Li, Y., Zhu, M., Klein, R., Kong, N. (2014). Using a partially observable Markov chain model to assess colonoscopy screening strategies–a cohort study. European Journal of Operational Research, 238 (1), 313–326.

Malik, M., & Thomas, L. C. (2012). Transition matrix models of consumer credit ratings. International Journal of Forecasting, 28 (1), 261-272. doi:10.1016/j.ijforecast.2011.01.007.

Maskawa, J.-i. (2003). Multivariate Markov chain modeling for stock markets. Physica A: Statistical Mechanics and its Applications, 324 (1-2), 317–322.

Massoni, S., Olteanu, M & Rousset, P. "Career-path analysis using drifting Markov models (DMM) and self-organizing maps". MASHS, 2010, Lille, France. ffhal-00443530f

Mehran, F. (1989). Longitudinal analysis of employment and unemployment based on matched rotation samples. Labour, 3 (1), 3–20.

Morimoto, H. (2016). Hidden Markov Models and Self-Organizing Maps Applied to Stroke Incidence. Open Journal of Applied Sciences. doi:06.158-168. 10.4236/ojapps.2016.63017.

Nicolau, J., Riedlinger, F. I. (2015). Estimation and inference in multivariate markov chains. Statistical Papers, 56 (4), 1163–1173.

Page, L., Brin, S., Motwani, R, Winogard, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web.

Quercia, R. G., Lei D. & Janneke R. (2009). Loan Modifications and Redefault Risk: An Examination of Short-Term Impact. Cityscape: A Journal of Policy Development and Research, 11 (3), 171-194.

Raftery, A., E. (1985). A Model for High-Order Markov Chains. Journal of the Royal Statistical Society, 47 (3), 528-539.

Raftery, A., Tavaré, S. (1994). Estimation and modelling repeated patterns in high order markov chains with the mixture transition distribution model. Applied Statistics, 179–199.

Régis, D. E., & Artes, R. (2016). Using multi-state Markov models to identify credit card risk. Production, 26 (2), 330-344. doi:10.1590/0103-6513.160814.

Sahin, A. D., Sen, Z. (2001). First-order Markov chain approach to wind speed modelling. Journal of Wind Engineering and Industrial Aerodynamics, 89 (3-4), 263–269.

Shamshad, A., Bawadi, M., Hussin, W. W., Majid, T., Sanusi, S. (2005). First and second order Markov chain models for synthetic generation of wind speed time series. Energy 30 (5), 693–708.

Siu, T.-K., Ching, W.-K., Fung, S. E., Ng, M. K. (2005). On a multivariate Markov chain model for credit risk measurement. Quantitative Finance, 5 (6), 543–556.

Somervuo, P. (2000). Competing hidden Markov models on the self-organizing map. In Proceedings of the International Joint Conference on Neural Networks, volume 3, pages 169–174, Piscataway, NJ. Helsinki University of Technology.

Spedicato, G.A., Kang T.S., Yalamanchi, S.B., Yadav, D. & Cordón, I., (2014). The markovchain package: A package for easily handling discrete Markov chains in R. Retrieved 19 June 2019 from https://cran.r-project.org/web/packages/markovchain/vignettes/an_introduction_to_markovchain_package.pdf.

Spedicato, G.A., Kang T.S., Yalamanchi S.B., Yadav D. (2019) The markovchain package: A package for easily handling Discrete Markov Chains in R. Retrieved 19 June 2019 from https://cran.r-project.org/web/packages/markovchain/markovchain.pdf.

Sperandio, M, Bernardon, D. P. & Garcia, V. J., "Building forecasting Markov models with Self-Organizing Maps," 45th International Universities Power Engineering Conference UPEC2010, Cardiff, Wales, 2010, pp. 1-5.

Sperandio, M, Bernardon, D. P. & Garcia, V. J., "Building forecasting Markov models with Self-Organizing Maps," 45th International Universities Power Engineering Conference UPEC2010, Cardiff, Wales, 2010, pp. 1-5.

Tracy, J., & Wright, J. (2016). Payment changes and default risk: The impact of refinancing on expected credit losses. Journal of Urban Economics, 93, 60–70.

Tsiliyannis, C. A. (2018). Markov chain modeling and forecasting of product returns in remanufacturing based on stock mean-age. European Journal of Operational Research.

Turchin, P. B. (1986). Modelling the effect of host patch size on mexican bean beetle emigration. Ecology 67 (1), 124–132.

Yadav, D., Kang, T.S. & Spedicato, G.A. (2017). Higher, possibly multivariate, Order Markov Chains in markovchain package. Retrieved 19 June 2019 from https://cran.r-project.org/web/packages/markovchain/vignettes/higher_order_markov_chains.pdf.

Yin, Hujun. (2008). The Self-Organizing Maps: Background, Theories, Extensions and Applications, 725-726. doi:10.1007/978-3-540-78293-3_17

Yu, Lean & Wang, S. & Lai, Kin Keung & Zhou, L. (2008). Bio-inspired credit risk analysis: Computational intelligence with support vector machines. doi:10.1007/978-3-540-77803-

9. APPENDIX

Table 9.1 - Mean Absorption Times (Modified Credits)

Cluster 1	States	1	2	3
		1515.702	1398.271	1386.923
Cluster 2	States	1	2	3
		1490.038	1314.810	1293.732
Cluster 3	States	1	2	3
		1578.866	1397.751	1407.882
Cluster 4	States	1	2	3
		1521.531	1309.561	1420.507

Table 9.2 - Mean Absorption Times (Unmodified Credits)

Cluster 1	States	1	2	3	4
		541.069	92.946	57.563	45.125
Cluster 2	States	1	2	3	4
		534.745	96.080	37.833	19.000
Cluster 3	States	1	2	3	4
		544.725	95.650	62.366	44.137
Cluster 4	States	1	2	3	4
		532.011	82.557	52.027	31.000

10. ANNEXES

Table 10.1 - Acquisition Data Elements (source: Fannie Mae)

Element/Variable	Type	Description	Allowable Values/Calculations
Borrower Credit Score at Origination	Numeric	A numerical value used by the financial services industry to evaluate the quality of borrower credit. When this term is used by Fannie Mae, it is referring to the "classic" FICO score developed by Fair Isaac Corporation.	▪ 300 – 850 ▪ Blank (if Credit Score is < 300 or > 850 or unknown)
Co-Borrower Credit Score at Origination	Numeric	A numerical value used by the financial services industry to evaluate the quality of co-borrower credit. When this term is used by Fannie Mae, it is referring to the "classic" FICO score developed by Fair Isaac Corporation.	▪ 300 – 850 ▪ Blank (if Credit Score is < 300 or > 850, unknown, or is not applicable)
First Time Home Buyer Indicator	Categorical	An indicator that denotes if the borrower or co-borrower qualifies as a first-time homebuyer	▪ Y = Yes ▪ N = No ▪ U = Unknown
Loan Identifier		A unique identifier for the mortgage loan. Variable of acquisition and performance files.	
Number of Borrowers	Numeric	The number of individuals obligated to repay the mortgage loan.	▪ 1 – 10
Original Debt to Income Ratio	Numeric	A ratio calculated at origination derived by dividing the borrower's total monthly obligations (including housing expense) by stable monthly income. This calculation is used to determine the mortgage amount for which a borrower qualifies.	▪ 1% – 64% ▪ Blank (if DTI is = 0, or ≥ 65, unknown, or if the mortgage loan is a HARP refinance
Origination Channel	Categoric	Channel refers to the three options: Retail (R), Correspondent (C), and Broker (B)	▪ R ▪ B ▪ C
Origination Interest Rate	Numeric	The original interest rate on a mortgage loan as identified in the original mortgage loan documents.	▪ Blank = Unknown
Original Loan Term	Numeric	The number of months in which regularly scheduled borrower payments are due under the terms of the related mortgage documents.	▪ 60 – 419

Original Loan-to-Value (LTV)	Numeric	A ratio calculated at the time of origination for a mortgage loan. The Original LTV reflects the loan-to-value ratio of the loan amount secured by a mortgaged property on the origination date of the underlying mortgage loan.	▪ 0% - 97% ▪ Blank (if LTV is > 97% or unknown)
Original Combined Loan-to-Value (CLTV)	Numeric	A ratio calculated at the time of origination for a mortgage loan. The CLTV reflects the loan-to-value ratio inclusive of all loans secured by a mortgaged property on the origination date of the underlying mortgage loan.	▪ 0% - 200% ▪ Blank (if CLTV is > 200% or unknown)
Original Unpaid Principal Balance (UPB)	Numeric	The original amount of mortgage loan as indicated by the mortgage documents.	
Origination Date	Date	The date of the note.	▪ YYYY

Table 10.2 - Performance Data Elements (source: Fannie Mae)

Element/Variable	**Type**	**Description**	**Allowable Values**
Current Actual UPB	Numeric	The current actual outstanding unpaid principal balance of a mortgage loan as it contributes to the current outstanding balance of the Reference Pool.	
Current Interest Rate	Numeric	The rate of interest in effect for the periodic installment due.	
Current Loan Delinquency Status	Categorical	The number of months the obligor is delinquent as determined by the governing mortgage documents.	▪ 1 = Current, or less than 20 days past due ▪ 2 = 30 – 59 days ▪ 3 = 60 – 89 days ▪ 4 = 90 – 119 days ▪ X = Unknown
Maturity Date	Date	The month and year in which a mortgage loan is scheduled to be paid in full as defined in the mortgage loan documents.	▪ YYYY
Modification Flag	Categorical	An indicator that denotes if the mortgage loan has been modified.	▪ Y = Yes ▪ N = No
Modification Date	Date	The number of months occurred since the mortgage loan's origination date and the moment of its modification. Available solely for modified loans.	
Principal Forgiveness Amount	Numeric	A reduction of the UPB owed on a mortgage by a borrower that is formally agreed to by the lender and the borrower, usually in conjunction with a loan modification.	
Zero Balance Code	Categorical	A code indicating the reason the mortgage loan's balance was reduced to zero.	▪ 01 = Prepaid or Matured ▪ 02 = Third Party Sale ▪ 03 = Short Sale ▪ 06 = Repurchased ▪ 09 = Deed-in-Lieu, REO ▪ 15 = Note Sale