POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Implementation of an Agent-Based Recommendation System Using Time-Variant Markov Chains: Investor Personas Forecasting with a Business Case Application

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Lorenzo Grossi**

Student ID: 976087
Advisor: Daniele Marazzina
Co-advisors: Raffaele Zenti
Academic Year: 2023-24

Tutto cambia, niente è perso per sempre

A Pietro

Abstract

This study aims to define a structured simulation process that enables predictive analysis of probabilistic scenarios regarding the behavior of a financial institution's client base and how it may evolve over medium and long-term horizons.
To demonstrate the utility and potential of this method, an operational strategy is developed to address the variations identified in the case study through the forecast, which are expected to be disruptive.
Building on a previously developed agent-based recommendation system, we present three possible 'evolutionary models' of client features, based on Gaussian perturbations of the data. The ultimate goal is to estimate the investor type of the customer in the future.
To quantify the probabilities of each client changing agents (a characterization that defines the investor type), time-varying Markov Chains, estimated via Monte Carlo simulations, are employed. The resulting chains exhibit significant variability over the years, though no substantial differences are observed across the three evolutionary models.
The simulation outcomes indicate a considerable increase in non-investor profiles, and to address these findings, we propose an operational strategy based on recommendations of specific insurance products tailored to these clients.
Given the freedom in developing the evolutionary model and the flexibility of the chains, the usefulness of this thesis extends beyond this case study and the financial sector, demonstrating significant potential in any domain involving recommendation systems and agent-based evolution, ranging from marketing strategies to healthcare, logistic, and many other fields.

Keywords: Financial Forecasting; Realistic Future Scenarios; Agent-Based Recommendation Systems; Investor Personas; Time-Variant Markov Chains; Client Segmentation; Business Case; Feature Variations; Monte Carlo Simulations

Abstract in italiano

Questo studio si propone di definire un processo di simulazione strutturato che consenta l'analisi predittiva di scenari probabilistici relativi al comportamento della base clienti di un istituto finanziario e alla sua evoluzione su orizzonti di medio e lungo termine.
Per dimostrare l'utilità e il potenziale di questo metodo, viene sviluppato una strategia operativa per affrontare le variazioni identificate nel caso di studio attraverso la predizione, che si prevede saranno dirompenti.
Basandoci su un sistema di raccomandazione agent-based precedentemente sviluppato, presentiamo tre possibili "modelli evolutivi" delle caratteristiche del cliente, basati su perturbazioni gaussiane dei dati. L'obiettivo finale è stimare che tipo di investitore il cliente sarà in futuro.
Per quantificare le probabilità che ogni cliente cambi agente (una caratterizzazione che definisce il tipo di investitore), si utilizzano catene di Markov variabili nel tempo, stimate tramite simulazioni Monte Carlo. Le catene risultanti mostrano una significativa variabilità nel corso degli anni, anche se non si osservano differenze sostanziali tra i tre modelli evolutivi.
I risultati delle simulazioni indicano un aumento considerevole dei profili non investitori e, per far fronte a questi risultati, proponiamo una strategia operativa basata sulla raccomandazione di prodotti assicurativi specifici per questi clienti.
Data la libertà nello sviluppo del modello evolutivo e la flessibilità delle catene, l'utilità di questa tesi si estende al di là di questo caso di studio e del settore finanziario, dimostrando un potenziale significativo in qualsiasi dominio che coinvolga i sistemi di raccomandazione e l'evoluzione basata sugli agenti, dalle strategie di marketing alla sanità, alla logistica e a molti altri campi.

Parole chiave: Previsione Finanziaria; Scenari Futuri Realistici; Sistemi di Raccomandazione Basati su Agenti; Profili di Investitori; Catene di Markov a Tempo Variabile; Segmentazione dei Clienti; Pianificazione Aziendale; Variazioni delle Caratteristiche; Simulazioni Monte Carlo

Contents

Abstract **i**

Abstract in italiano **iii**

Contents **v**

Introduction **1**

1 Dataset Exploratory Analysis **3**
1.1 Dataset Presentation . . . 3
1.2 Preliminary Statistical Analysis . . . 6
1.2.1 Correlation Analysis . . . 6
1.2.2 Principal Component Analysis . . . 7
1.2.3 Outliers Detection . . . 8
1.2.4 Univariate Densities (divided by Investor Type) . . . 10

2 Data-Driven Customer Segmentation **13**
2.1 Summary . . . 13
2.2 Mixed Distance for Hierarchical Clustering . . . 14
2.3 Optimization of Cluster Analysis . . . 15
2.4 Analysis of Results . . . 16
2.5 Needs-based recommendation systems . . . 20
2.6 Recommendation algorithm . . . 22
2.7 Results of the Recommendation System . . . 23
2.8 Conclusions . . . 24

3 Models Presentation **27**
3.1 Common Traits across All Models . . . 29
3.1.1 The Algorithm . . . 29
3.1.2 Death Simulation . . . 29

3.1.3 New Client Injection . . . 30
3.1.4 Vanishing Investor Coefficient . . . 31
3.2 First Model: Deterministic . . . 32
3.3 Second Model: Age and Income Variation . . . 32
3.4 Third Model: Variation in Multiple Features . . . 33

4 Models Results **37**
4.1 MonteCarlo Method Implementation . . . 37
4.2 Time-Variant Markov Chain . . . 38
4.3 Effect of the Evanescent Investor Coefficient . . . 38
4.4 First Model Results . . . 40
4.5 Second Model results . . . 42
4.6 Third Model results . . . 44
4.7 Results comparisons . . . 46

5 Strategic Adaptation to Client Distribution Changes **49**
5.1 Current Operational Strategy . . . 50
5.2 Analysis of Variations . . . 51
5.3 Future Strategies Recommendations . . . 52
5.3.1 Cluster 5: Scission and Insurance Needs . . . 54
5.3.2 Cluster 6: Separation and Needs . . . 55
5.3.3 Summary of the Results and Forecasts . . . 56

6 Discussions **59**
6.1 Key Findings . . . 59
6.2 Interpretations . . . 60
6.3 Implications . . . 61
6.4 Limitations . . . 62
6.5 Recommendations . . . 63

Conclusions **65**

Bibliography **67**

List of Figures **71**

List of Tables **73**

List of Algorithms **75**

Acknowledgements **77**

Introduction

The financial world is currently facing a revolutionary period, as various factors are significantly altering the way investments are approached, with no turning back.
Two main reasons contribute to this shift, and their combination will lead to unprecedented scenarios. Specifically:

- A demographic transition resulting in a population that is older than ever before[5], coupled with the increasing need for a client base that requires **quality of life in old age**, demanding protection and fulfillment of their expectations.
- A drastic shift in focus among younger generations toward investment products[2], exploring **offers beyond traditional options** (such as cryptocurrencies and sustainable assets), which generates a significant break from the strategies of past generations.

Starting from these premises, a new imperative emerges for financial institutions: to develop **models that forecast future behaviors** and prepare to address them.
Forecasting is rapidly becoming a significant and challenging task, with many approaches available. Given the intrinsic difficulty of this problem, we will simulate different models and compare their outcomes to assess their similarities and differences.

Among other reasons, models to predict future scenarios are crucial primarily for:

- Financial planning and budgeting, establishing realistic budgets crucial for financial stability and competitiveness.
- Preallocating resources and products, ensuring optimal utilization within capacity and availability when products are needed.
- Developing a detailed plan over time for products suited to each customer's needs, which can make the difference between a client purchasing or not purchasing products from the institution's portfolio.
- Providing stakeholders with an accurate plan of a plausible scenario to gain reliability and trustworthiness, which are essential for a modern financial institution.

Reliable and complete data is crucial for performing accurate simulations and effectively identifying customers.
However, possessing data alone is not sufficient to guarantee satisfactory and accurate results. It is equally important to identify the most suitable methods and make correct and innovative inferences.
In summary, data is the lifeblood of sales forecasting, enabling proactive and informed financial operations management, but it needs to be processed through the correct channels to yield meaningful results.

Regarding the literature, the topic of forecasting financial behavior is gaining interest, but most studies focus on the evolution of financial products over time (particularly using Moving Average Processes and their variations), rather than on the evolution of clients and their needs.
As mentioned earlier, this will be a crucial breakpoint that institutions will face in the near future.
Our approach is distinct and innovative, exploring a path that has not been fully investigated despite its potential and broad applications, extending beyond just financial contexts given its flexibility and adaptability.

We will build on the work of our colleague Veronica Lucchetti[12], who, within the field of agent-based recommendation systems, segmented a dataset of bank clients into clusters or "personas" (that are concrete and identifiable representations of client groups, particularly useful for distinguishing generations with differing interests and needs) and identified their specific requirements and suitable products.
Particularly, with the author's consent, some parts of Chapters 1 and 2 have been extracted from her thesis "Data-Driven Customer Segmentation: A Needs-Based Cluster Analysis for Optimizing Financial Product Recommendations"[12].
Our contribution will involve developing models to track the evolution of the client base through persona classification, monitoring the increases or decreases in group sizes.
We will utilize time-variant Markov Chains to model the transition probabilities of clients between clusters based on their evolving features, estimating these transitions through Monte Carlo simulations.
Subsequently, we will analyze how the demand for financial products within a portfolio shifts to forecast the necessities and quantities that a case-study financial institution needs to preallocate.
Finally, we will provide a concrete application of these methodologies and devise an operational strategy to address the future needs of the institution under examination.

1 | Dataset Exploratory Analysis

This chapter will be devoted to an initial exploration of our data; understanding and ensuring their reliability is a necessary condition in order to build the models presented in the following chapters.
First, we will describe the data to gain an understanding of the terms used to characterize our clients (Section 1.1).
Second, with a particular focus on numerical features, we will assess whether all features are relevant and necessary for the analysis or if some features provide redundant information (Subsection 1.2.1 and Subsection 1.2.2).
To proceed, in Subsection 1.2.3, we will examine the behavior of each client to identify any outliers or anomalies that should be excluded from the analysis.
Finally, Subsection 1.2.4 will present a graphical exploration of the univariate distribution of each numerical variable, segmented by Investor Type, the most significant categorical feature for our subsequent analysis. This will help us identify any recognizable patterns that could be useful and exploitable moving forward.

1.1. Dataset Presentation

The dataset utilized for this project consists of anonymized clients from an Italian Financial Institution, described by both categorical and numerical features. The dataset is sourced from the internet and provides valuable insights into the financial activities of 5,000 unique customers.
As anticipated, it consists of 5,000 data entries encompassing various demographic and socioeconomic attributes of individuals, crucial for understanding various aspects of personal finance and socioeconomic dynamics.
It comprehends variables ranging from basic demographic characteristics such as Age and Gender to numerical financial indicators like Income, Wealth, and Debt, that provide valuable insights into individuals' financial behaviors, decision-making processes, and overall financial well-being.
Categorical variables represent qualitative characteristics, such as Gender or Job cate-

gory, while numerical variables quantify quantitative aspects, such as Income or Wealth amounts, and clearly there will be the necessity to treat them with different techniques.

Below is a comprehensive compilation detailing the variables and their respective characteristics within the dataset:

- **ID**: Numerical identifier assigned to each individual in the dataset.
- **Age**: Age of the individual in years.
- **Gender**: Gender of the individual (Female, Male).
- **Job**: Occupation of the individual. This category distinguishes individuals who are unemployed, are employees, hold managerial positions, are entrepreneurs or are retired.
- **Area**: Geographic region where the individual resides. The geographic regions are categorized as North, Central, and South/Islands. They refer to the Italian territory.
- **CitySize**: Size of the city where the individual resides. The city sizes are classified as small towns, medium-sized towns, and large cities with a population exceeding 200,000 residents.
- **FamilySize**: Number of components in the individual's family.
- **Income**: Normalized Income of the individual.
- **Wealth**: Normalized Wealth of the individual.
- **Debt**: Normalized Debt of the individual.
- **FinEdu**: Normalized Financial Education level of the individual.
- **ESG**: Normalized Environmental, Social, and Governance (ESG) propensity of the individual.
- **Digital**: Normalized Digital propensity of the individual.
- **BankFriend**: Normalized Bank Friendliness of the individual.
- **LifeStyle**: Normalized Lifestyle Index of the individual; higher values indicate a higher lifestyle index.
- **Luxury**: Normalized Luxury spending propensity of the individual.
- **Saving**: Normalized Saving propensity of the individual.

- **Investments**: Type of investment of the individual. This category distinguishes between individuals with no interest in investments, those who primarily make lump sum investments, and others who predominantly engage in capital accumulation.

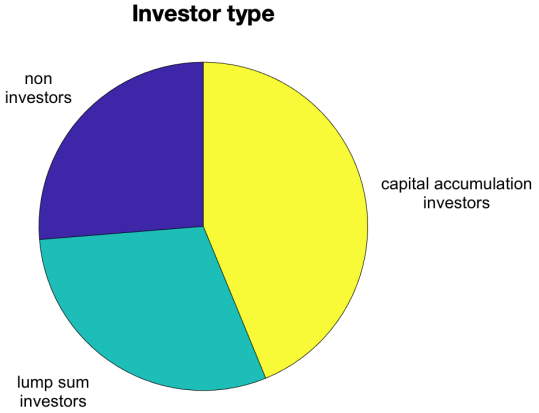

Figure 1.1: *The majority of clients are capital accumulators*

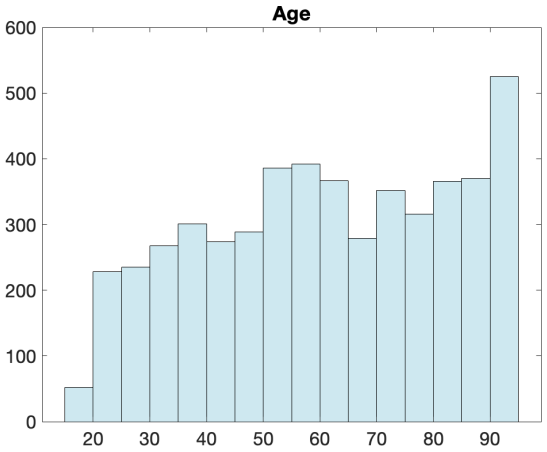

Figure 1.2: *The population is quite old, with an average age of 60 years*

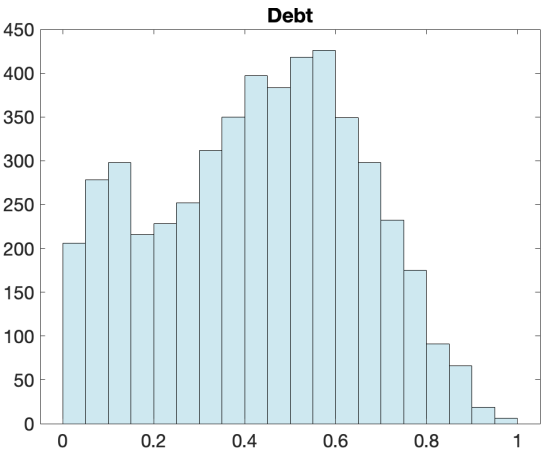

Figure 1.3: *The debt distribution has a major peak and a secondary one*

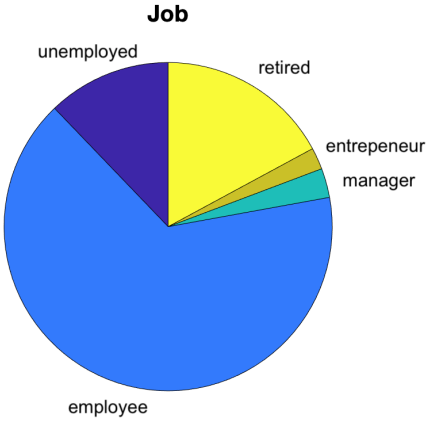

Figure 1.4: *A vast majority of our sample are employees*

Of particular interest during the analysis carried out subsequently will be the **classification of customers based on their investment activity**. The classification according to investments includes distinctions between lump sum investments and capital accumulation investments. They represent two distinct approaches to building and managing investment portfolios. Lump sum investments involve allocating a large sum of money into investments all at once, typically as a single transaction. This approach can be advantageous for investors who have a significant amount of capital available and want to immediately deploy it into the market. On the other hand, capital accumulation investments involve systematically adding smaller amounts of money to investments over

time, often through regular contributions or periodic purchases. This approach allows investors to spread out their investment over time, potentially reducing the impact of market volatility and taking advantage of dollar-cost averaging.
While lump sum investments offer the potential for immediate market exposure and potential returns, capital accumulation investments provide a disciplined and gradual approach to building wealth over the long term. Investors often choose between these two strategies based on their financial goals, risk tolerance, and investment timeline.

Noteworthy findings from the dataset include the revelation that a majority of clients, approximately 73%, have active investments, with a subset of this group, consisting of only 29%, having lump sum investments, while 44% have capital accumulation investments (see Figure 1.1). In contrast, approximately 27% of clients didn't make any investments.
Another remarkable fact is the **advanced age** of some clients (about 10% of them are over 90 years old, the mean age is 60.45 years and the standard deviation is 21.82 years), which can be clearly seen in Figure 1.2 and will surely impact our results, particularly in the estimate of deaths through the mortality rate (described in Subsection 3.1.2, which will be modeled as a function of the clients' age, as intuition suggests).
Finally, from Figure 1.4, we can notice that most of the individuals are employees, followed by about 10-12% of them who are either unemployed or retired.

1.2. Preliminary Statistical Analysis

The dataset has been prepared normalizing the numerical variables (in particular Age and Family Size, being the others numerical features expressed in percentages and so already in the [0,1] range). We decided to include FamilySize in the following quantitative analysis given its intrinsic numerical nature, albeit discrete, as an ordinal variable.
The absence of missing values indicates accuracy in data collection and, a priori, allows us to retain all 5000 observations.

1.2.1. Correlation Analysis

In order to start analyizing the multivariate prospective of our numerical data, and in particular the pairwise relationship of the quantitative features, we focus on the **correlation matrix**[19], reported in Figure 1.5, computed as

$$\mathbf{C} = \frac{1}{n-1}\mathbf{Z}^T\mathbf{Z} \tag{1.1}$$

where $\mathbf{Z}$ is the standardized matrix composed by numerical features and n the number of observations (equal to 5000 in our case).
Some intuitive relations are confirmed by the computation, but the highest linear correlation value is around 0.56: we retain it is not enough to consider to discard one feature, considering also the significant meaning of each of them.

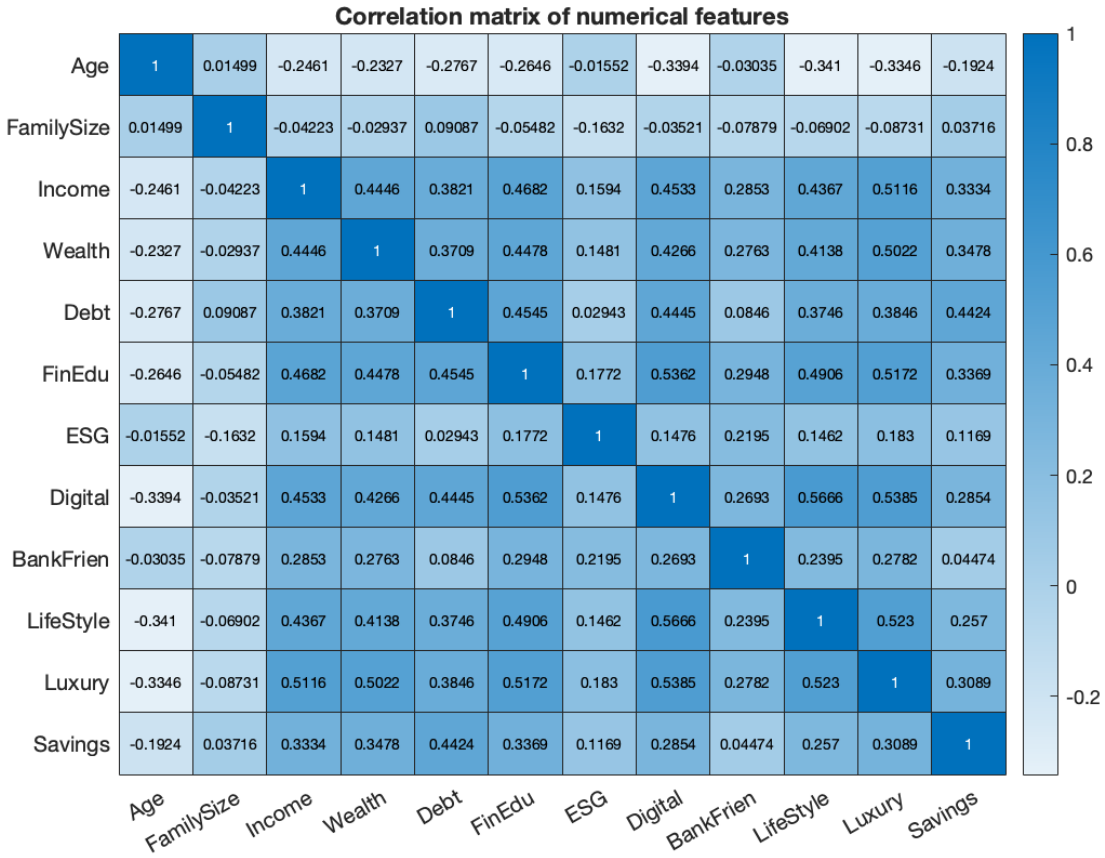

Figure 1.5: *Pairwise correlation between quantitative variables*

1.2.2. Principal Component Analysis

Another analysis performed to understand the significance of our numerical features, and in particular their role, is the **Principal Component Analysis (PCA)**[17], a statistical technique used to simplify the complexity of high-dimensional data while retaining the majority of the data variability.
PCA transforms the original numerical features into a new set of uncorrelated orthogonal variables called principal components. Each principal component is obtained as a linear combination of the original variables, and these components are ordered so that the first few retain most of the variability present in the original dataset; by construction, the

first n components generate the n-dimensional subspace with the maximum variability, representing the most informative part of the data.
Numerically, the first principal component is the eigenvector associated with the largest eigenvalue of the correlation matrix (Computed as 1.1), the second principal component is the eigenvector associated with the second largest eigenvalue, and so on.

Because of this, one of its main uses is to observe the scores (or loadings) of the first principal components in terms of the original variables. If the first principal components, which summarize the majority of the variability, have negligible scores for one of the original features, it means that the significance of that feature is low and its influence is not important.
The scores of the first principal components are reported in Figure 1.6. What emerges, even in this case, is that we can't discard any numerical variable, because all of them make an important contribution to the overall variability, particularly in the first four principal components, which capture about 64% of the total variability.

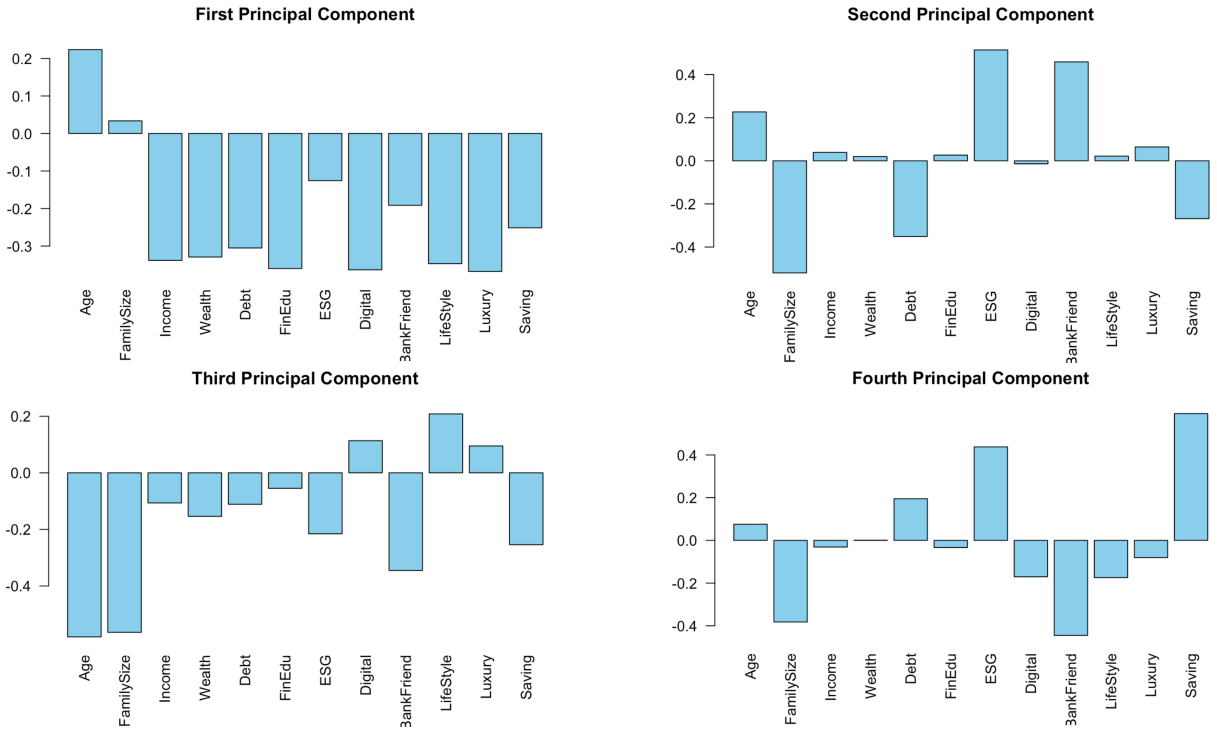

Figure 1.6: *Variables' scores for the first four principal components*

1.2.3. Outliers Detection

Always focusing on numerical variables, a valuable way to understand if a client has anomalous behavior relative to the overall data distribution is through data depth mea-

sures.
First introduced in 1975 by John Tukey[21], **Tukey Depth** (or Halfspace Depth) quantifies how deeply a point lies within a data cloud.
Its strength lies in requiring no distributional assumptions, as the data cloud is a random sample from $\mathbf{F}$, the cumulative distribution function associated with a probability distribution in $\mathbb{R}^d$, $d \geq 1$, and the specific distribution need not be known.
Formally, the Tukey Depth of a point $\mathbf{x}$ with respect to a distribution or a sample of points in $\mathbb{R}^d$ is a measure of depth that reflects the centrality of the point relative to the distribution. More precisely, the Tukey depth of a point $\mathbf{x}$ is defined as the minimum number of points in any half-space containing $\mathbf{x}$.

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a sample of n points in $\mathbb{R}^d$. The Tukey depth of a point $\mathbf{x}$ with respect to X is defined as:

$$\text{depth}(\mathbf{x}; X) = \min_{\mathbf{u} \in \mathbb{S}^{d-1}} |\{i : \mathbf{u} \cdot (\mathbf{x}_i - \mathbf{x}) \geq 0\}| \tag{1.2}$$

where $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$ and $\mathbf{u} \cdot (\mathbf{x}_i - \mathbf{x})$ represents the dot product between $\mathbf{u}$ and $\mathbf{x}_i - \mathbf{x}$. In other words, for each direction $\mathbf{u}$, we calculate the number of points that lie in the half-space defined by the direction $\mathbf{u}$ and the point $\mathbf{x}$. The Tukey depth of $\mathbf{x}$ is the minimum of these counts over all possible directions $\mathbf{u}$.

When focusing on $d = 2$ (two variables at a time), we can utilize Tukey depth to construct a **bagplot**[18], a useful graph for identifying outliers in two-dimensional space.
In a bidimensional plot, outliers are characterized by their depth and isolation from the rest of the points.
Considering 12 numerical variables, we adopt the following approach: we identify outliers in each bagplot formed by pairs of variables and then examine whether any clients are outliers across various pairs of variables.
Most points are outliers in only one pair of dimensions, with a few in two pairs.
Notably, one individual, an unemployed non-investor identified by ID 2218, is an outlier in three different pairs of variables (Figure 1.7). In all cases involving Financial Education, this individual exhibits notably low levels (0.064), despite high levels of Wealth, Lifestyle, Luxury, and Digital Propensity (all above 0.88).
Despite these peculiarities, we decide not to discard the point due to the overall reliability of these data and the fact that outliers across three pairs of variables represent a small percentage.

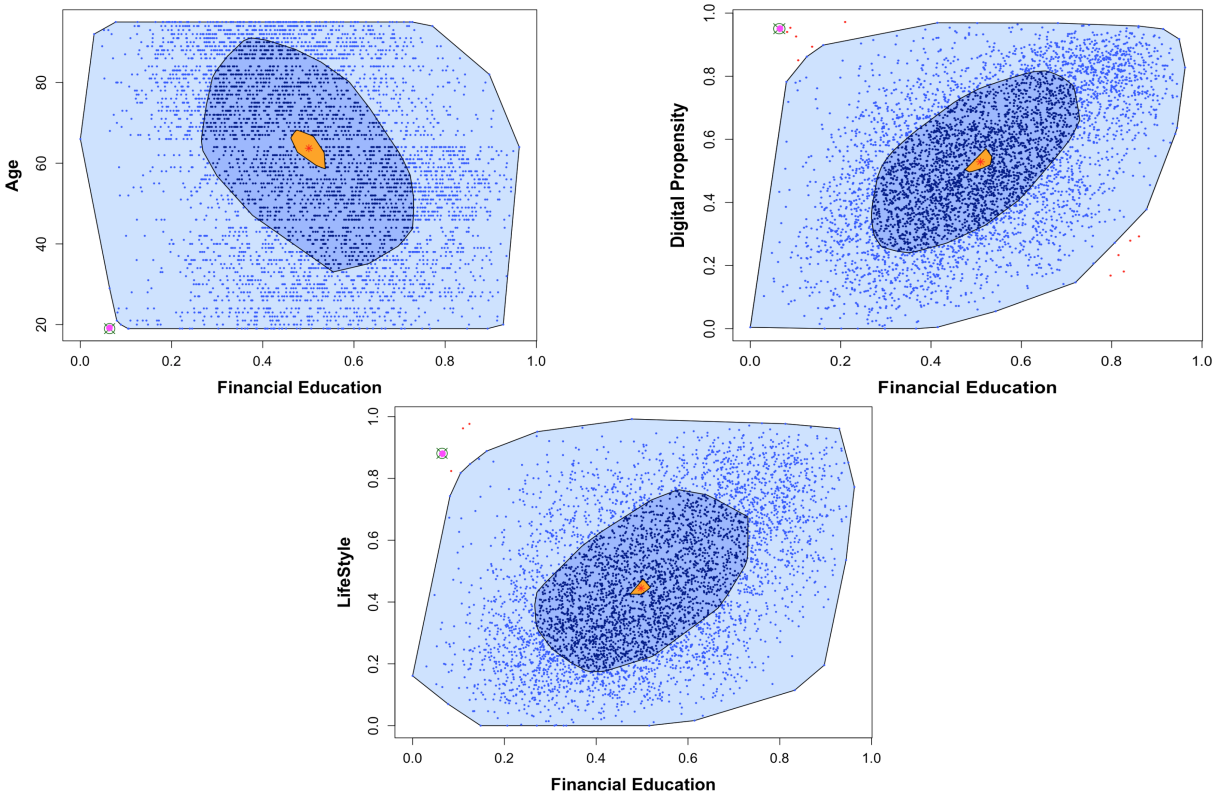

Figure 1.7: *Bagplot of three pairs of variables highlighting client 2218 as an outlier in the bivariate distributions, all related to Financial Education*

1.2.4. Univariate Densities (divided by Investor Type)

Another interesting visual analysis is to observe if any pattern emerges in the **univariate densities** of the numerical variables when divided by type of investment (i.e., considering different groups of investors).

In Figure 1.8 we can see the result: univariate densities do not provide many clues about what might be a discriminant for detecting types of investors. Age is the only variable that shows a slight pattern, distinguishing capital accumulator investors from other types, particularly in the mode, which differs significantly.

From this plot, we also gain an intuition: it seems very unlikely that the multivariate distribution of our data is Gaussian. We can obtain statistical evidence of this by performing a Shapiro-Wilk Test[20] (in its univariate form, the null hypothesis is that the data are normally distributed against the alternative hypothesis that they are not) on one of the covariates, such as Income.

Performing the test on the entire population yields a $p - value < 2.2 \times 10^{-16}$.
Similarly, filtering the population for lump-sum and capital accumulation investors gives

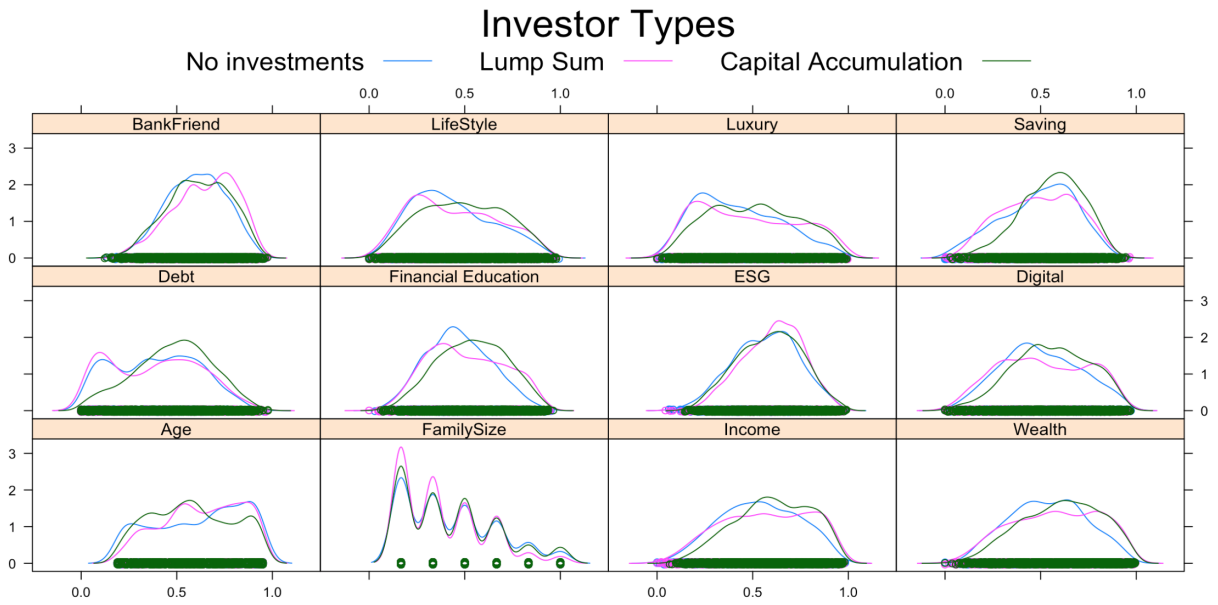

Figure 1.8: *Univariate (normalized) densities divided by type of investors*

a p-value $< 2.2 \times 10^{-16}$, while for non-investors, the test on the income variable results in a p-value $= 7 \times 10^{-9}$.

For all groups, the conclusion is the same: we reject the null hypothesis (that Income is normally distributed, both considering the full population and filtering by type of investors) at any reasonable significance level. Since a necessary condition for a multivariate Gaussian distribution is that each of its components is univariately Gaussian, we conclude that we cannot rely on the assumption of Gaussianity for our numerical data.

2 | Data-Driven Customer Segmentation

The next section will summarize the thesis work by Veronica Lucchetti[12], which will guide the subsequent steps of our analysis.
Specifically, we will highlight the decisions made in collaboration with Veronica that significantly influence the remainder of our work and inform our predictions.
In her thesis, Lucchetti not only presents the obtained results but also performs a valuable empirical analysis to identify variables with minimal impact on the clustering process, which turn out to be Gender and Family Size.
Given that our current segmentation approach relies on these techniques, we assert that these variables play an insignificant role in the segmentation process we are addressing. Therefore, we will reduce complexity by excluding their evolution from consideration.
This reasoning is explained in more detail in Section 2.3.

2.1. Summary

Lucchetti's work focuses on developing an algorithm to recommend optimal banking products tailored to individual customer profiles. This process enables financial institutions to identify profitable investment opportunities, preallocate necessary resources for customers, and engage in personalized consulting by proposing relevant products only when customers are genuinely interested.
The foundation of Lucchetti's approach rests on two primary assumptions: firstly, that a large number of customers can be categorized into a few clusters, each representing personas sharing similar traits and financial needs; and secondly, that customers within each cluster maximize their future profitability by adopting identical products. As such, these products are associated with specific clusters or personas and form the basis for recommendations and suggestions to their members.

The development of the thesis is articulated through the following sections:

- Definition of a method to compute distances between points in the multidimensional space of client summaries, considering both numerical and categorical variables.
- Selection of the most appropriate clustering technique and linkage type for the dataset under examination.
- Determination of the optimal number of clusters using dendrograms and clustering validation indices.
- Profiling of clusters by analyzing variables that describe personas and their unique characteristics, thereby translating numerical data into meaningful real-world personas.
- Development of an algorithm that recommends the most suitable products available to the financial institution for each defined persona. This recommendation is based on a risk index for both products and clients, designed by Lucchetti.

2.2. Mixed Distance for Hierarchical Clustering

After opting to use hierarchical clustering, the need for a suitable distance metric became crucial, particularly given the requirement to handle both categorical and numerical data. The choice of a **mixed distance** approach became evident, and after experimentation, the combination of Hamming distance[8] for categorical variables:

$$\text{Hamming}(x, y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{2.1}$$

where $\delta(x_i, y_i)$ is the Kronecker delta function (which returns 1 if $x_i = y_i$ and 0 otherwise), and Euclidean (or L^2) distance for numerical variables:

$$\text{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.2}$$

proved to be the most suitable for our dataset. This choice was validated through a t-SNE analysis[22] to identify the optimal separation between future clusters. The results of this analysis are depicted in Figure 2.1, illustrating clear separation between groups of points in low-dimensional spaces. The combination of these two distance metrics yielded the clearest distinction among the various combinations tested.

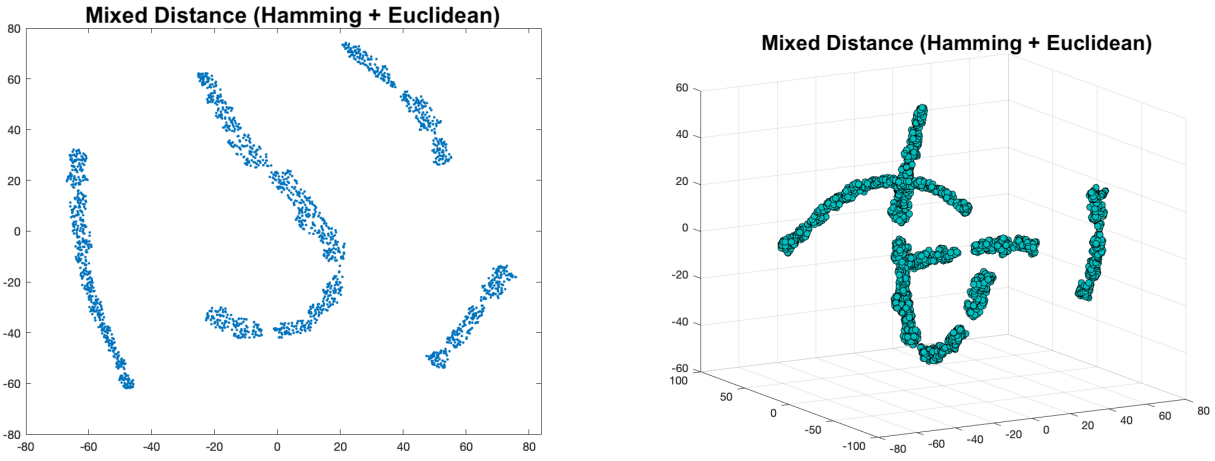

Figure 2.1: *Datapoints distribution in 2-dimesional and 3-dimensional t-SNE reduction*

Finally, significant weights were assigned to dissimilarities in categorical variables, particularly emphasizing the Investor Type. This decision stems from the study's primary goal of matching each cluster with an appropriate investment product, underscoring the importance of clearly distinguishing the preferred investment strategies within each cluster.

2.3. Optimization of Cluster Analysis

The selection of the most suitable linkage technique for our dataset involved comparing the cophenetic coefficients C across different linkages, ultimately leading to the choice of **unweighted average linkage**. Notably, the coefficient for this linkage is $C = 0.8992$, indicating a robust clustering solution that effectively captures the underlying data structure.

Subsequently, the determination of the optimal number of clusters was based on both the dendrogram analysis[11] depicted in Figure 2.2 (to understand the hierarchical relationships between clusters) and the evaluation of several goodness-of-separation indices.
The contemporary evaluation of goodness-of-fit indices[3] confirmed the graphical suggestion from the dendrogram: 6 was identified as the optimal number of clusters for our dataset.

After conducting several trials, it was observed that the variables Gender and Family Size contribute minimally to the clustering process. Consequently, we decided to exclude them from further analysis to prioritize simplicity over negligible effects, following the principle of Occam's Razor.
This decision is supported by empirical evidence demonstrating their insignificance in achieving effective clusterization.

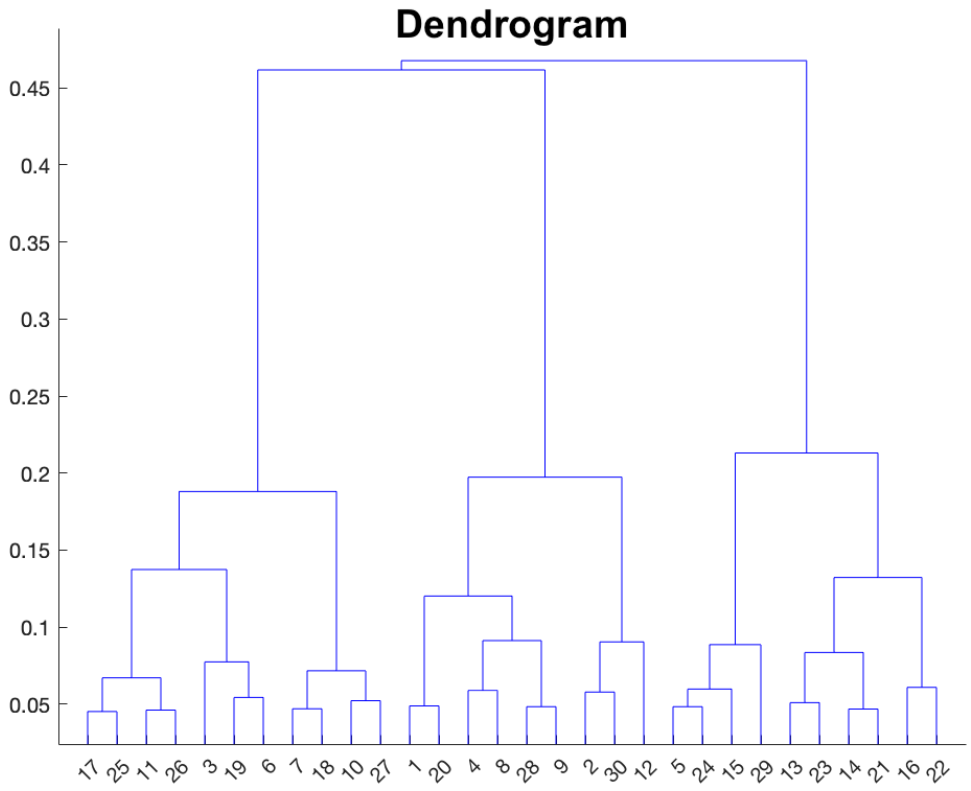

Figure 2.2: *Dendrogram obtained with Average Linkage*

2.4. Analysis of Results

In the following section, individuals in different groups are briefly described based on both categorical variables and their distribution, as well as numerical variables focusing on mean values.

The standout feature is clearly the Investor Type class, which shows distinct separation across clusters; each cluster predominantly consists of clients from a specific Investor Type class. Consequently, we assume that each cluster is primarily associated with this Investor Type class.

Cluster 1: "Retirees from the North"

The first cluster contains 737 clients, and it is an **income investment** cluster. The clients are all residents in medium or big cities mainly in the northern regions of Italy. 45% of them are retiree, as the high age factor might have suggested, but 33% are employees.

Numerical feature	Classification	Mean value
Age	High	81
Income	Below Average	0,44
Wealth	Below Average	0,44
Debt	Low	0,25
Financial Education	Below Average	0,40
ESG Propensity	Above Average	0,58
Digital Propensity	Below Average	0,38
Bank Friendliness	Above Average	0,61
Lifestyle Index	Low	0,28
Luxury Spending	Low	0,27
Saving Propensity	Below Average	0,38

Table 2.1: *"Retirees from the North" numerical features*

Cluster 2, "Young employees from the North"

The second cluster contains 759 observations and is another **income-oriented** cluster, but it consists of young people, mainly employees or still unemployed. Overall, this cluster represents a financially secure and forward-thinking segment, characterized by a focus on wealth accumulation and responsible investment strategies.

Numerical feature	Classification	Mean value
Age	Low	46
Income	Above Average	0.67
Wealth	High	0.71
Debt	Average	0.52
Financial Education	Above Average	0.63
ESG Propensity	Above Average	0.62
Digital Propensity	Above Average	0.67
Bank Friendliness	High	0.71
Lifestyle Index	Above Average	0.59
Luxury Spending	Above Average	0.64
Saving Propensity	Above Average	0.57

Table 2.2: *"Young Employees from the North" numerical features*

Cluster 3, "Affluent Elderly Employees from the North"

The third cluster contains 804 observations and is a **capital accumulation** cluster. Compared to the first cluster, these individuals boast a higher level of financial wealth, which aligns with their capital accumulation tendency.

Numerical feature	Classification	Mean value
Age	High	81
Income	Average	0.55
Wealth	Above Average	0.56
Debt	Below Average	0.44
Financial Education	Average	0.46
ESG Propensity	Average	0.47
Digital Propensity	Average	0.45
Bank Friendliness	Above Average	0.60
Lifestyle Index	Low	0.38
Luxury Spending	Low	0.36
Saving Propensity	Average	0.54

Table 2.3: *"Affluent Elderly Employees from the North" numerical features*

Cluster 4: "Young Employees from the South"

The fourth cluster contains 1386 observations, making it the most populated cluster, and it is a **capital accumulation** cluster. The age of the people is low, but their level of wealth is above average. The cluster encompasses a mix of employed individuals, alongside a smaller proportion of unemployed, mainly from cities in the Center and South.

Numerical feature	Classification	Mean value
Age	Low	44
Income	Above Average	0.64
Wealth	Above Average	0.65
Debt	Average	0.51
Financial Education	Above Average	0.58
ESG Propensity	Above Average	0.61
Digital Propensity	Above Average	0.61
Bank Friendliness	Above Average	0.62
Lifestyle Index	Average	0.55
Luxury Spending	Average	0.55
Saving Propensity	Above Average	0.58

Table 2.4: *"Young Employees from the South" numerical features*

Cluster 5: "Retired Non-Investors"

The fifth cluster contains 770 observations and groups people who are not interested in any type of investment so far. They are mostly retirees from big cities in the North or Center.

Numerical feature	Classification	Mean value
Age	High	79
Income	Average	0.48
Wealth	Average	0.49
Debt	Below Average	0.32
Financial Education	Below Average	0.42
ESG Propensity	Above Average	0.58
Digital Propensity	Below Average	0.42
Bank Friendliness	Above Average	0.60
Lifestyle Index	Below Average	0.33
Luxury Spending	Below Average	0.31
Saving Propensity	Below Average	0.42

Table 2.5: *"Retired Non-Investors" numerical features*

Cluster 6: "Young Wealthy Sensible Savers"

The sixth cluster contains 544 observations and is another non-investment cluster. Seven percent of them are managers or freelancers.

Numerical feature	Classification	Mean value
Age	Low	37
Income	Above Average	0.59
Wealth	Above Average	0.60
Debt	Average	0.48
Financial Education	Average	0.50
ESG Propensity	Above Average	0.57
Digital Propensity	Above Average	0.59
Bank Friendliness	Above Average	0.59
Lifestyle Index	Average	0.51
Luxury Spending	Average	0.51
Saving Propensity	Average	0.55

Table 2.6: *"Young Wealthy Sensible Savers" numerical features*

2.5. Needs-based recommendation systems

The clustering discussed in the previous subsection is based on the premise that individuals within a group share similar characteristics and needs.
The next step involves creating a dataset that encapsulates the distinct needs of each cluster. This will be achieved by considering the values of relevant parameters within each group and augmenting this dataset by introducing a Risk Propensity Index for each client (Section 2.6), which will serve as a measure of the person's inclination towards risk-taking behavior.

Below is a brief list of the products the financial institution can offer to its clients (more details in Lucchetti's thesis[12]). Each product is associated with a Risk Index (RI), which is pivotal in the recommendation algorithm described in the next section, as it will be compared to the Risk Propensity Score of each client.
Another fundamental feature, also crucial in the recommendation system, is the nature of the product (Income or Accumulation), from which the list is divided.

Income Products

- **A Unit Linked Insurance Plan (ULIP)**: A multi-faceted product that offers both insurance coverage and investment exposure in equities or bonds. $RI = 0.3$.
- **Fixed Income Mutual Fund**: An investment vehicle that pools money from multiple investors to invest primarily in fixed-income securities. $RI = 0.12$.
- **Balanced High Dividend Mutual Fund**: An investment fund that aims to provide investors with a balanced approach to wealth accumulation by focusing on both capital appreciation and dividend income. $RI = 0.44$.
- **Fixed Income Segregated Account**: A financial product offered by insurance companies or investment firms that provides investors with a tailored fixed-income investment strategy within a segregated account structure. $RI = 0.13$.

Accumulation Products

- **Balanced Mutual Fund**: A type of investment fund that seeks to provide investors with a diversified portfolio by investing in both equities and fixed-income securities. $RI = 0.55$.
- **Defensive Flexible Allocation Unit-Linked**: A type of life insurance product that combines the benefits of life insurance coverage with investment opportunities. $RI = 0.38$.
- **Aggressive Flexible Allocation Unit-Linked**: A life insurance product that favors higher-risk assets with the potential for higher returns. This aggressive approach seeks to maximize long-term growth potential, albeit with increased volatility and risk exposure. $RI = 0.75$.
- **Balanced Flexible Allocation Unit-Linked**: Similar to the above funds but with a balanced approach. This indicates that the investment strategy aims to strike a balance between risk and return by investing in a mix of equities, bonds, and other assets. $RI = 0.48$.
- **Cautious Allocation Segregated Account**: A specialized investment vehicle typically offered by insurance companies or investment firms. It aims to provide investors with a conservative approach to wealth accumulation by allocating funds to a diversified portfolio of low-risk assets. $RI = 0.27$.
- **Total Return Aggressive Allocation Segregated Account**: An investment

product that aims to provide investors with aggressive growth opportunities while managing risk through diversified asset allocation. $RI = 0.88$.

2.6. Recommendation algorithm

To associate each client with the most fitting product, an index called the Risk Propensity Index is calculated for each client as follows:

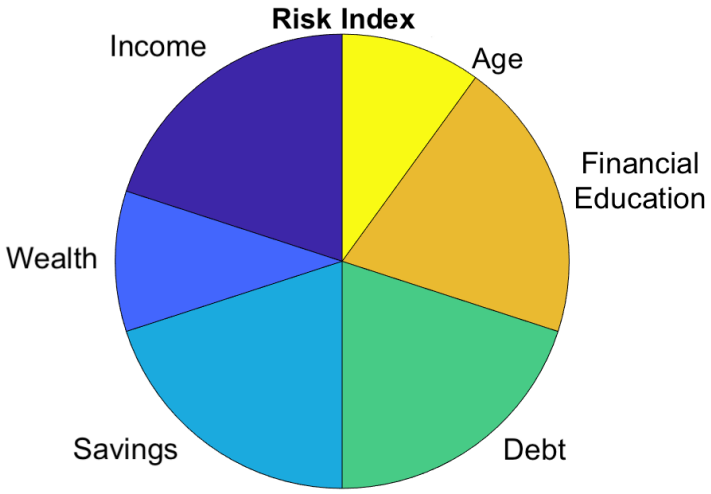

Figure 2.3: *Pie chart with the contribution of each variable to the risk index computation*

After constructing the index, the recommendation algorithm for each cluster is defined as follows:

Algorithm 2.1 Recommendation Algorithm

```
for each cluster do
    Determine whether its members are interested in capital investments or income investments
    for each client do
        Calculate their risk index
        Check the products of that type that have a risk index lower than the client's
        if the product is the one with the potential for the biggest return then
            Associate it with the client
        end if
    end for
    Select the most recurring products of that type
end for
```

2.7. Results of the Recommendation System

In this section, the results obtained for each cluster are presented. These results will form the core of the business strategy our financial institution should implement to maximize success in selling products.
The strategy emphasizes the consolidation of the 6 clusters rather than focusing on individual clients, driven by the following key reasons:

- **Time and resource savings.** This approach streamlines the analysis process, reducing the burden on human resources that would otherwise be required to individually assess and follow up with each of the 5000 clients. It also facilitates communication strategies.
- **Support for pre-positioning analysis.** Banks often engage in a practice known as "pre-positioning" of financial products in anticipation of client demand. This strategy involves banks purchasing securities, bonds, or other financial products before securing a buyer.
- **Automation.** Adopting this method fully exploits the potential of automation, establishing efficient systems that not only boost operational effectiveness but also ensure smooth handling of client engagements, while reducing manual tasks.

Cluster 1

The risk index of this cluster is low compared to the average, which suggests that clients may be inclined to invest in low-risk funds.
The most suitable product for these clients is definitely an Income Conservative Unit-Linked.
Another viable option for clients within this cluster includes a Balanced High Dividend Mutual Fund and a Fixed Income Segregated Account.

Cluster 2

The risk index of this cluster is higher than the previous one, so the most fitting products for these clients are riskier compared to those suitable for the first cluster.
Clients in this cluster prefer a Balanced High Dividend Mutual Fund, which carries more risk than the other two products.
Additionally, clients in this cluster may be willing to invest in an Income Conservative Unit-Linked. Only a small portion will opt for a Fixed Income Segregated Account.

Cluster 3

The risk index of this cluster is slightly below average.
The most fitting product for these clients is a Defensive Flexible Allocation Unit-Linked, followed by a Balanced Flexible Allocation Unit-Linked.
The least appealing option for them will be a Cautious Allocation Segregated Account.

Cluster 4

In the fourth cluster, which contains 1,386 people, all members are interested in capital accumulation investments. Therefore, for this cluster, we will consider only capital accumulation products.
The risk index of this cluster is average and more or less aligned with the previous accumulation cluster.
The most fitting product for these clients is still a Defensive Flexible Allocation Unit-Linked, but in this case, more people are willing to pursue either a Balanced Flexible Allocation Unit-Linked or a Balanced Mutual Fund.
The least appealing option remains a Cautious Allocation Segregated Account.

Cluster 5 and 6

The fifth and sixth clusters contain 770 and 544 people, respectively, all of whom are not interested in making any investments. Therefore, we will not associate them with any of the investment products.
Both clusters have low levels of risk propensity, which is consistent with their choice not to make any investments.

2.8. Conclusions

Through the examination and application of advanced cluster analysis techniques, six distinct clusters have been successfully identified, each encapsulating the diverse characteristics and needs of our clients.
Each cluster defines what we consider a "persona" (a prototype of humans with specific characteristics) that has been heavily influenced by categorical variables.
It is important to note that the dataset contained a significant number of elderly individuals with limited financial knowledge, reflecting a true spectrum of society.

Moreover, our rigorous approach has enabled us to allocate the most impactful financial products to each cluster, ensuring maximum relevance and effectiveness in meeting their financial objectives.

Lucchetti's thesis provides a solid and reliable foundation for the forecasting we have ideated and will develop in future chapters, as well as for the effective innovation that our thesis aims to bring.
Considering these clusters as the personas that compose society (or at least the subset of people who can be clients of our financial institution), we will build real-case scenarios of feature variation for individuals from our dataset and evaluate the probabilities, over a certain number of years, that people will change clusters. In our model, this implies migration from one persona to another, consequently reflecting different interests in the range of products offered by the institution.
The final goal will be to estimate, using different probabilistic models, how our current client base will evolve and which products will need to be preallocated to precisely meet their demand.
This will allow us to gain a crucial competitive advantage in a highly competitive market such as the financial sector.

3 | Models Presentation

This chapter, together with the next one, represents the core of our work, as it presents the effective proposal for addressing the agent-based recommendation problem in the near future.
Here, we will describe several possible models for client evolution over the years, along with the rationale behind each, while in Chapter 4 the results of these model simulations will be used to estimate the transition probabilities between clusters over different years, observing how many clients move or stay.

The decision to define three distinct models is crucial to convey a key message: there are innumerable ways to model client evolution.
This represents both the main challenge and the potential of this forecasting technique, as it allows us not only to tackle each problem in various ways, but also to address multiple agent-based problems across different fields of study.

In each subsection dedicated to a specific model, we will first describe the rationale behind it and then the tools used to implement that evolution.
Our choice is to prioritize the use of basic tools over overly sophisticated and complex ones, as we have found that effective means can achieve satisfactory results, as demonstrated in the previous chapters.

A common feature across all models is the stochastic simulation of client's death, which depends on the client's current age and is explained in detail in Subsection 3.1.2. The introduction of death necessitates adding a "death cluster" alongside the six "financial clusters", This death cluster is, in Markov Chain terms, an absorbing state, as once clients enter it, they will never migrate to another cluster.
Another artificial cluster acts as a storage pool, from which 200 clients are added each year to maintain a nearly constant initial client base of 5,000 (as discussed in Subsection 3.1.3).
A key hypothesis common to all models is that categorical variables (Gender, Job, Area, and City Size) remain constant over the years, along with Family Size (which, according to Lucchetti's thesis, is not a significant factor in cluster assignment). This choice stems

from the belief that these features rarely change, and implementing such rare changes would be both challenging and impractical.
In contrast, numerical variables vary differently in each model. Age increases deterministically by one unit each year, while other variables change stochastically depending on the model in question (the first model has no variation except for Age, while the last model sees all numerical features changing).
Finally, a common element across all models is the "Evanescent Investment Coefficient", described in Subsection 3.1.4. This coefficient is based on the distance metric created by Veronica in her clustering process and is used here to compute the nearest center, thereby determining the cluster to which a client belongs.

3.1. Common Traits across All Models

All models follow the same pattern: clients evolve year by year, and their nearest center is calculated to assign them to a cluster. Then, some clients die, and others are added from storage to join the client base for the current year.
The specific details are explained in the following subsections.

3.1.1. The Algorithm

In general terms, all three models follow Algorithm 3.1 steps:

Algorithm 3.1 Identification of clients' investor type each year

```
for years from 1 to 10 do
    Increment clients' Age by 1
    for each client do
        if Client died the previous year then
            Assign the client to the death cluster for the current year
        else
            Modify the client's features (according to the model used)
            Simulate the client's death through a random variable depending on gender
            and age (older clients have a higher probability of death)
        end if
    end for
    Decrease the Investor Type weight
    Compute the distance matrix of each alive client to each center of the original
    clusters
    Assign clients to the cluster with the nearest center
    Add 200 clients from a "storage" to the current client base
end for
```

3.1.2. Death Simulation

To simulate the possibility of a **client's death** in year y, we sample from a Bernoulli random variable, as shown in Equation (3.1).
If the outcome is positive, client c enters the absorbing state of dead people in year y and remains there for the rest of the simulation.

$$Death_c(y) \sim \text{Bernoulli}(p) \ , \ p = p(Age_c(y), Gender_c) \tag{3.1}$$

The probability of each client dying in a given year depends on their current age and gender. This probabilities are defined as the "probability of dying within the following year at a given age" (making them an appropriate parameter p for the Bernoulli random variable used), and are sourced from *mortality.org*[10], where such data is stored.
Our choice is to rely on data for Italy (as our clients are Italian) from 2019, the most recent year not affected by COVID-19.
The probabilities for males and females at different ages are shown in Figure 3.1, revealing significant differences between genders, particularly at advanced ages.

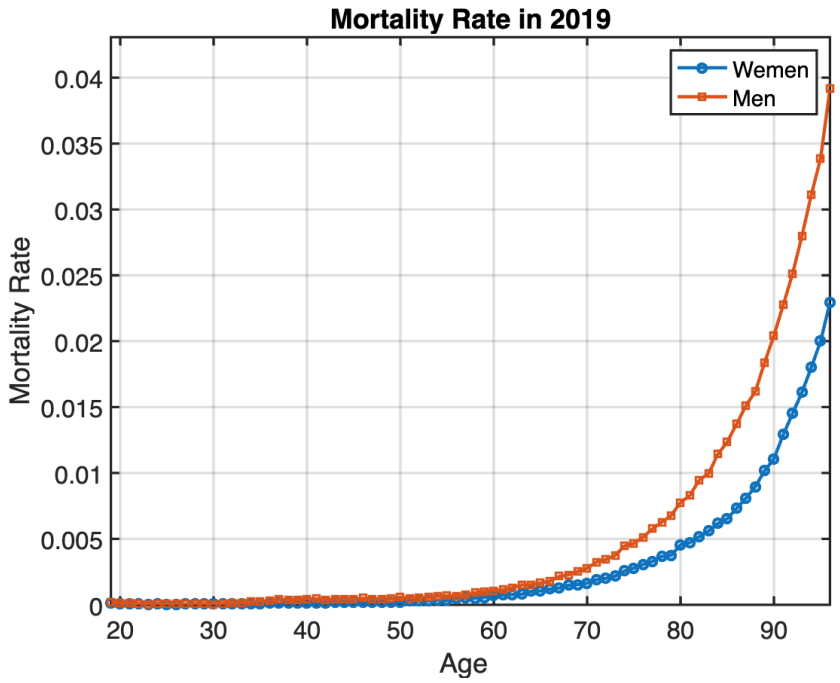

Figure 3.1: *Probability of dying within a year based on age and gender in Italy for 2019*

3.1.3. New Client Injection

The final step each year is to select 200 clients from a **synthetic data pool**, generated according to the multivariate distribution of our costumers, and add them to the current clientbase.
These clients are chosen randomly, with a probability inversely proportional to their age (younger individuals are more likely to be selected) to simulate new client subscriptions, and will be part of the client base from the following year onwards.
This approach helps maintain a nearly constant number of living clients. The synthetic data is generated using *gretel.ai*[7].

3.1.4. Vanishing Investor Coefficient

The final concept consistently implemented across all three models is the so-called **"Vanishing Investor Coefficient"**, which refers to the diminishing importance over time of the Investor Type at year 0.
In order to understand this concept, we need to recall how the distance of a client from a cluster center is computed; as defined in Section 2.2, each client's *clt* distance from a cluster center *cnt* is computed using a Mixed Distance, which is:

$$d(clt, cnt) = \sqrt{\sum_{i=1}^{n}(x_{clt} - x_{cnt})^2} + \sum_{i=1}^{c}\delta(y_{clt}, y_{cnt}) + \textbf{Weight} * \delta(Inv_{clt}, Inv_{cnt}) \quad (3.2)$$

Where $\delta(x_i, y_i)$ is the Kronecker delta function, x_i is one of the n numerical features of the client/center i, y_i is one of the c categorical features, Inv_i is the Investor Type (Lump Sum, Capital, or non-investor), and **Weight** is the Investor Coefficient.
In the distance computed on the original data (i.e. at year 0), **Weight** assumes a high fixed value to reflect the importance of the dissimilarity of the Investor Types.

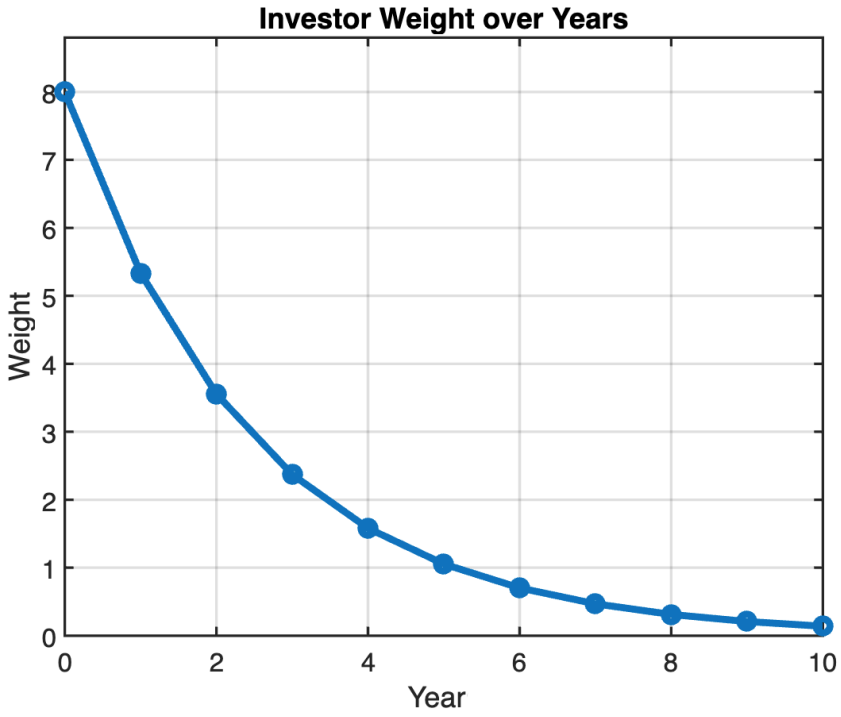

Figure 3.2: *Evanescence of "Investor Type at year 0" weight*

However, with the change of perspective in this thesis, in which years pass and clients evolve, the information about what type of investor a client was at year 0 loses importance as time goes by.
To model this dynamic, we define a distance that is the same as 3.1.4, but with the difference that the investor coefficient $Weight = Weight(y)$ depends on the year y, and in particular diminishes as the distance is computed in more advanced years (Figure 3.2).

The rationale behind is that the investor type at year 0 gradually loses relevance over time, and reducing its impact each year reflects this idea: initially, its influence is significant, but by the end, it becomes almost negligible.

3.2. First Model: Deterministic

In the first (and very simplistic) model, the client is considered frozen in time for all features except Age, which increases by one unit each year. For each client c, at year y:

$$Age_c(y) = Age_c(y-1) + 1 \tag{3.3}$$

This model is clearly an oversimplification of reality, but it serves as a useful baseline. We will examine whether significant differences exist in client migrations between clusters in this model compared to subsequent ones and determine whether introducing more feature variations is warranted.

3.3. Second Model: Age and Income Variation

In the second model, in addition to the deterministic increase in Age, we introduce variation in Income, one of the most significant features. To achieve this, we add stochastic variation that is linked to the data; specifically, the mean of the Gaussian variation is derived from the slope of the second-order least squares polynomial fitted to the Income data distribution.

The rationale behind this is that the observed trend (where income increases until around age 50 and then decreases) is significant and can be exploited.

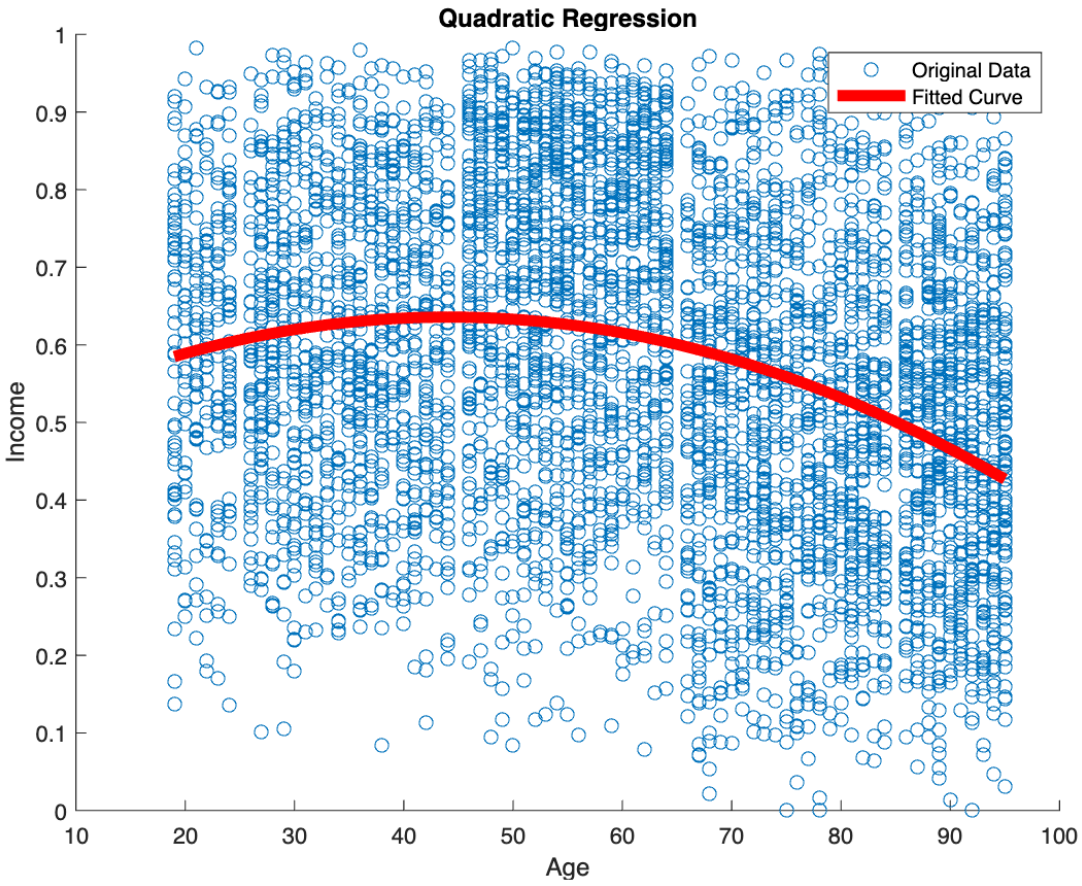

Figure 3.3: *Second-order regression polynomial of Income distribution*

Formally, the variables that vary for client c at year y are:

$$\begin{cases} \text{Age}_c(y) = \text{Age}_c(y-1) + 1 \\ \text{Income}_c(y) = \text{Income}_c(y-1) + \epsilon_c(y) \end{cases}, \ \epsilon_c(y) \sim \mathcal{N}\left(\text{slope}_{income}(\text{Age}_c(y)), 0.01\right) \quad (3.4)$$

3.4. Third Model: Variation in Multiple Features

In the third model, we introduce Gaussian noise to the multivariate distribution of all numerical variables, while keeping categorical variables fixed. This involves adding a Gaussian perturbation to the vector of numerical variables $x_c(y) \in [0,1]^{10}$ for client c at year y.

The perturbation is defined as follows: the mean $\mu_c(y)$ is 0 for all components except for Financial Education, Income, and Wealth. Regarding Financial Education, a positive value is assigned because it is difficult to lose once acquired. Income and Wealth follow

the same approach as in the second model, with the slope of the second-order polynomial used for adjustment.

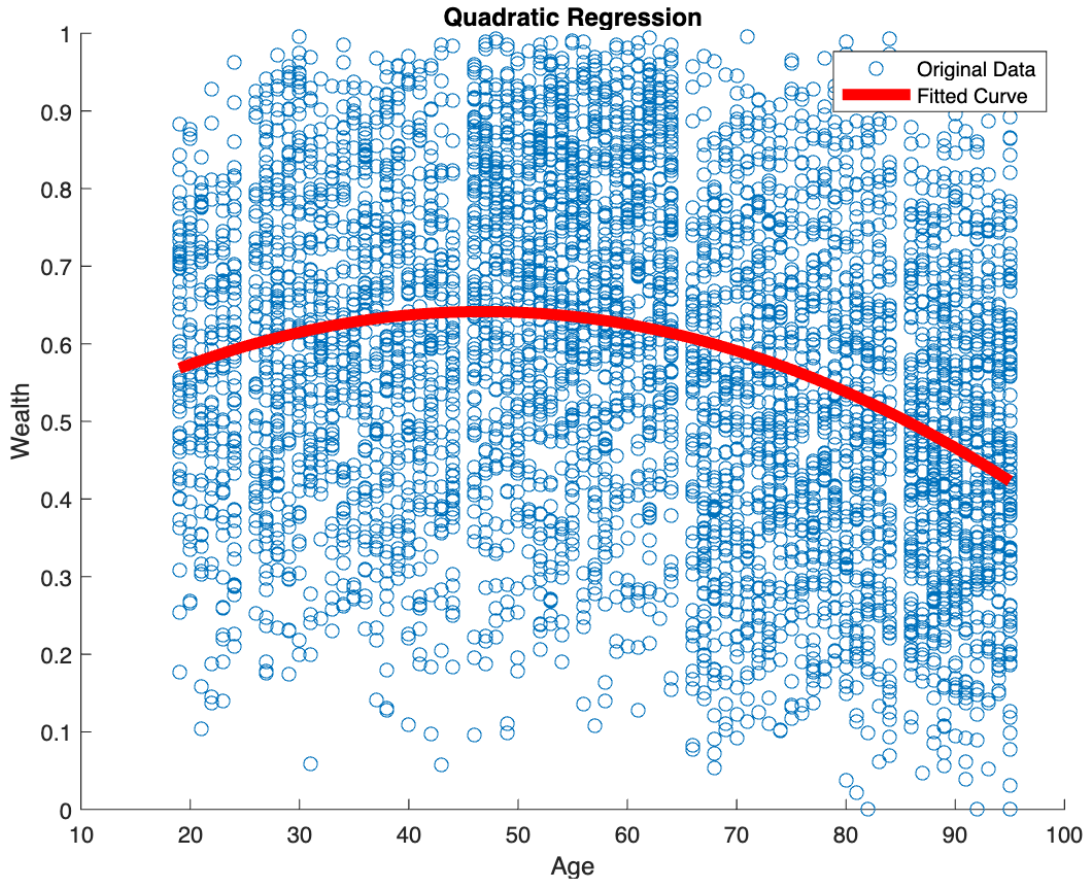

Figure 3.4: *Second-order regression polynomial of Wealth distribution*

The covariance matrix Σ for the Gaussian perturbation is derived from the original data (Figure 1.5, multiplied by an adjustment factor), preserving the relationships and correlations between variables.

Formally, the variation of numerical features of client c at year y is defined as:

$$\begin{cases} \text{Age}_c(y) = \text{Age}_c(y-1) + 1 \\ \begin{bmatrix} \text{Income}_c(y) \\ \text{Wealth}_c(y) \\ \text{Debt}_c(y) \\ \text{FinEdu}_c(y) \\ \text{ESG}_c(y) \\ \text{Digital}_c(y) \\ \text{BankFriend}_c(y) \\ \text{LifeStyle}_c(y) \\ \text{Luxury}_c(y) \\ \text{Saving}_c(y) \\ \text{Debt}_c(y) \end{bmatrix} = \begin{bmatrix} \text{Income}_c(y-1) \\ \text{Wealth}_c(y-1) \\ \text{Debt}_c(y-1) \\ \text{FinEdu}_c(y-1) \\ \text{ESG}_c(y-1) \\ \text{Digital}_c(y-1) \\ \text{BankFriend}_c(y-1) \\ \text{LifeStyle}_c(y-1) \\ \text{Luxury}_c(y-1) \\ \text{Saving}_c(y-1) \\ \text{Debt}_c(y-1) \end{bmatrix} + \epsilon_c(y) \end{cases} \tag{3.5}$$

$$\epsilon_c(y) \sim \mathcal{N}\left(\begin{bmatrix} \text{slope}_{inc}(\text{Age}_c(y)) \\ \text{slope}_{wlt}(\text{Age}_c(y)) \\ 0 \\ 0.05 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma/100 \right) \tag{3.6}$$

4 | Models Results

This chapter presents the results obtained from the Monte Carlo simulations for each of the three models described in the previous one.
Section 4.1 and Section 4.2 provide a brief theoretical recap of the methods and tools used, as well as their key properties.
For each simulation result, a comparison is made to assess how differences between the models affect the predicted client base, especially over time as uncertainty increases.

The final goal of these results is to produce a realistic configuration of the composition of the clusters in the future, particularly to understand specific trends of increase or decrease. These results have countless uses, and the next chapter will describe a possible one.

From a technical point of view, starting now state -1 will represent the storage pool of new customers entering the client base each year, state 0 will be the absorbing state for deceased clients, and states 1 through 6 will represent the personas (and associated investor types) described in Section 2.4.

4.1. MonteCarlo Method Implementation

The First Model is deterministic, so multiple simulations are unnecessary.
By contrast, the other models are stochastic, making the **Monte Carlo method**[14] the most appropriate tool for simulating various outcomes. Averaging the transition matrices from different trials allows us to obtain accurate estimates of the transition probabilities between states over time.
Firstly, the number of clients moving (or remaining) in clusters are normalized by rows to generate proper transition matrices in the Markov Chain sense.
The transition matrix $P_y^{(m)}$ of Model m between year $y-1$ and y is computed as:

$$P_y^{(m)} \approx \frac{1}{n} \sum_{i=1}^{n} (P_y^{(m)})_i \tag{4.1}$$

where $(P_y^{(m)})_i$ represents the transition matrix for i-th simulation of Model m at year y. Our estimates are based on $n = 100$ trials. While increasing this number would improve accuracy, it would come at a significant computational cost. For the scope of this work, we believe 100 trials offer a good balance between precision and computational efficiency.

4.2. Time-Variant Markov Chain

The result of the Monte Carlo simulation is a **Time-Variant (non-homogeneous) Markov Chain**[13].
Unlike a (canonical) time-invariant Markov Chain, where the probability of transitioning from one state to another depends only on the current state, in a time-variant Markov Chain, these probabilities also depend on the year.
The representation of a time-variant Markov Chain is a transition matrix P_y, where the probability of moving from state i to state j between years $y-1$ and y is denoted $P_{i,j|y}$. The transition matrix $P_{0\to n}$ for an n-year jump from year 0 is computed as the product of the transition matrices over the interval $[0, ..., n]$:

$$P_{1\to n} = P_1 \times P_2 \times \cdots \times P_n \tag{4.2}$$

The distribution of clients π_n at year n given the distribution π_0 at year 0 is:

$$\pi_n = \pi_0 \times P_1 \times P_2 \times \cdots \times P_n \tag{4.3}$$

In our case study, the initial distribution π_0 is given by:

$$\pi_0 = \begin{bmatrix} 2000 & 0 & 737 & 759 & 804 & 1386 & 770 & 544 \end{bmatrix} \tag{4.4}$$

This shows that the storage contains 200 clients to be released each year for 10 years (first position, cluster -1), and at year 0, no clients have died (second position, cluster 0).

4.3. Effect of the Evanescent Investor Coefficient

The first result we want to highlight is the impact of the Evanescent Investor Coefficient in cluster computation.

We recall that the Investor Coefficient is the weight given in 3.1.4 to the similarity or dissimilarity of a client's Investor Type with the cluster center (and that each client is

associated with the cluster that has the nearest center), and that this weight decreases as years pass.

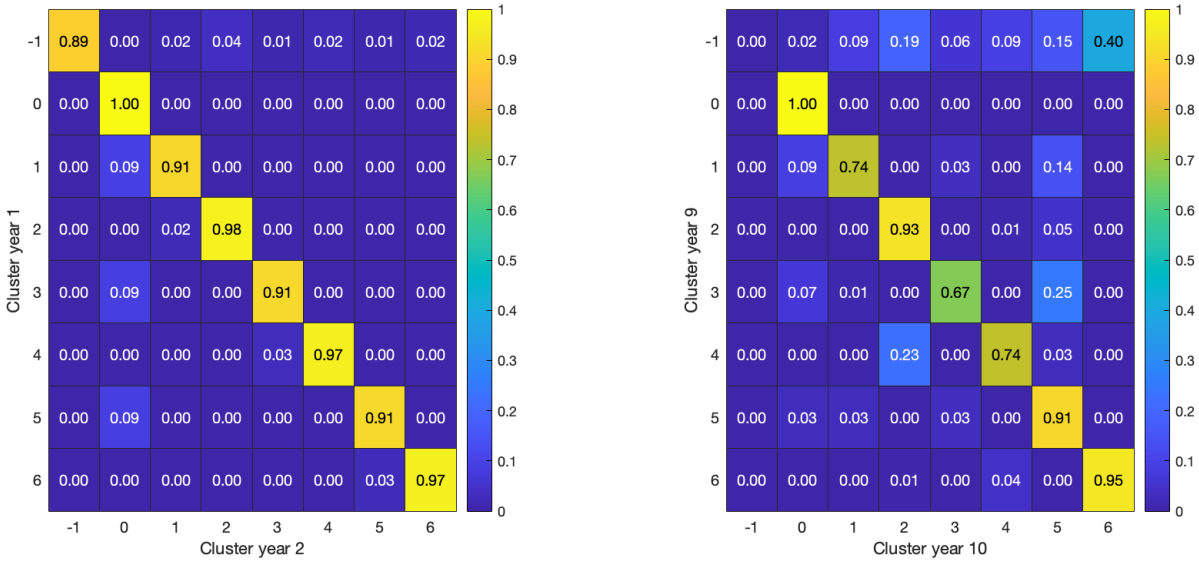

Figure 4.1: *Difference between transition matrices in early and advanced years*

Figure 4.1 shows the transition matrices $P_2^{(1)}$ (transition matrix for the first model between years 1 and 2) and $P_{10}^{(1)}$ (transition matrix for the same model between years 9 and 10).

In the early years there is no possibility for a client to transition to a persona characterized by a different investor type; for example, examining the submatrix $P_2^{(1)}(1:6, 1:6)$ the only nonzero values outside the main diagonal are $P_{2,1|2}^{(1)}$, $P_{4,3|2}^{(1)}$ and $P_{6,5|2}^{(1)}$. However, the investor type does not vary between these clusters, meaning there is no chance to change investor type in the early years.

This behavior arises because the difference between investor types carries significant weight in the computation of distances between points and cluster centers, making transitions practically impossible. Over time, however, the investor coefficient gradually reduces his weight, as described in Section 3.1.4. As a result, transitions between different investor types become more likely, as shown by the increased transition probabilities between different investor clusters in $P_{10}^{(1)}$ (Figure 4.1). For example, a non-zero probability is observed in $P_{4,2|2}^{(1)} = 0.25$.

This modeling reflects the intuition that, in the early years, the investor type at year 0 holds significant relevance and certainty, but over time, its importance diminishes, becoming almost irrelevant after many years (in the context of a human lifespan).

In the following sections, we will present the simulation results for each model, particularly focusing on the client base evolution at years 5 and 10.
Conclusions about product sales will be based on these two points, and we highlight the key results for each model, including $P_{0\to5}$, $P_{0\to10}$, π_5 and π_{10}.

4.4. First Model Results

The first model varies only the Age of current client base, as described in Section 3.2.
The first output of the algorithm is the transition matrix between each step; Figure 4.2 shows the matrices with transition probabilities for a client moving between clusters at year 5 and year 10 starting from a given cluster at year 0.

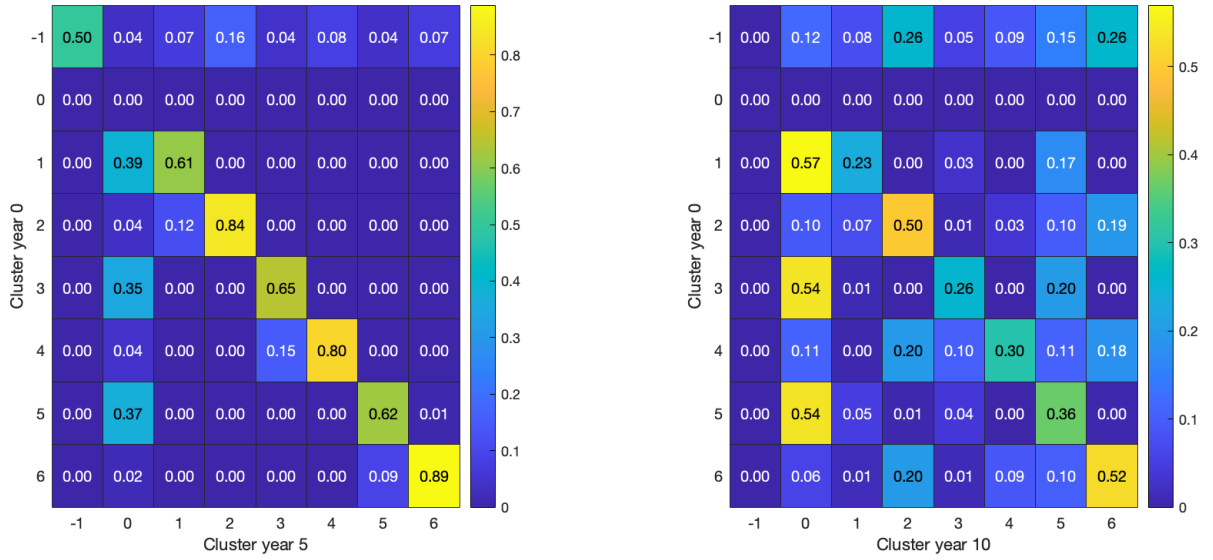

Figure 4.2: *Transition matrices* $P^{(1)}_{0\to5}$ *and* $P^{(1)}_{0\to10}$

The population of clusters at each year is shown in Figure 4.3.
Highlighting the results at year 5 and 10, we obtain:

$$\begin{cases} \pi_0^{(1)} = \begin{bmatrix} 2000 & 0 & 737 & 759 & 804 & 1386 & 770 & 544 \end{bmatrix} \\ \pi_5^{(1)} = \begin{bmatrix} 1000 & 1022 & 679 & 964 & 818 & 1277 & 608 & 632 \end{bmatrix} \\ \pi_{10}^{(1)} = \begin{bmatrix} 0 & 1751 & 438 & 1274 & 514 & 676 & 1144 & 1203 \end{bmatrix} \end{cases} \tag{4.5}$$

The following conclusions can be drawn:

- After five years, significant probabilities exist for moving only between clusters of the same Investor Type or dying.

- More than half the clients in the elder clusters (1,3 and 5) are likely to die within ten years, with over 35% dying within five years.
- It is more likely for new clients to enter younger clusters (as their probability of entering the client base is inversely proportional to age).
- After ten years, some significant migration occurs between different types of investors, particularly from 1 to 5, 2 to 6, 3 to 5, 4 to 2 or 6, and 6 to 2.
- The most populated cluster at year 0, cluster 4, remains constant until year 6, after which it experiences a drastic decline.
- Non-investor clusters (5 and 6) are sparsely populated in the first years, but then sharply increase and become two of the three most populated clusters by year 10.
- Cluster 2 experiences a sudden increase in members and becomes the most populated cluster after ten years.

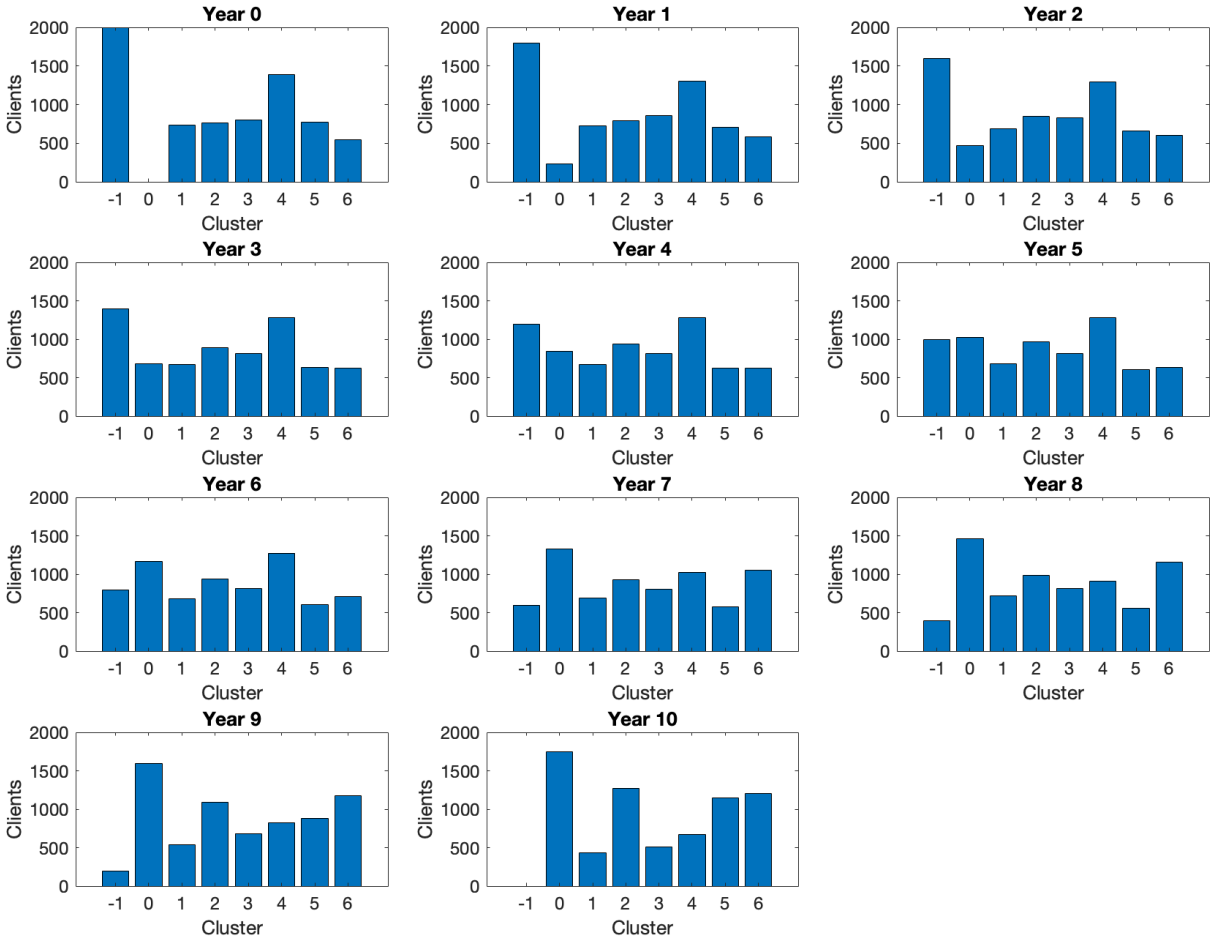

Figure 4.3: *Clients distribution at each year (First model)*

4.5. Second Model results

The second model generated by our algorithm is very similar to the first one, with the only change being the income of each client, as described in Subsection 3.3.
The transition matrices for a client moving between clusters at year 5 and year 10, starting from a given cluster at year 0, are shown in Figure 4.4.

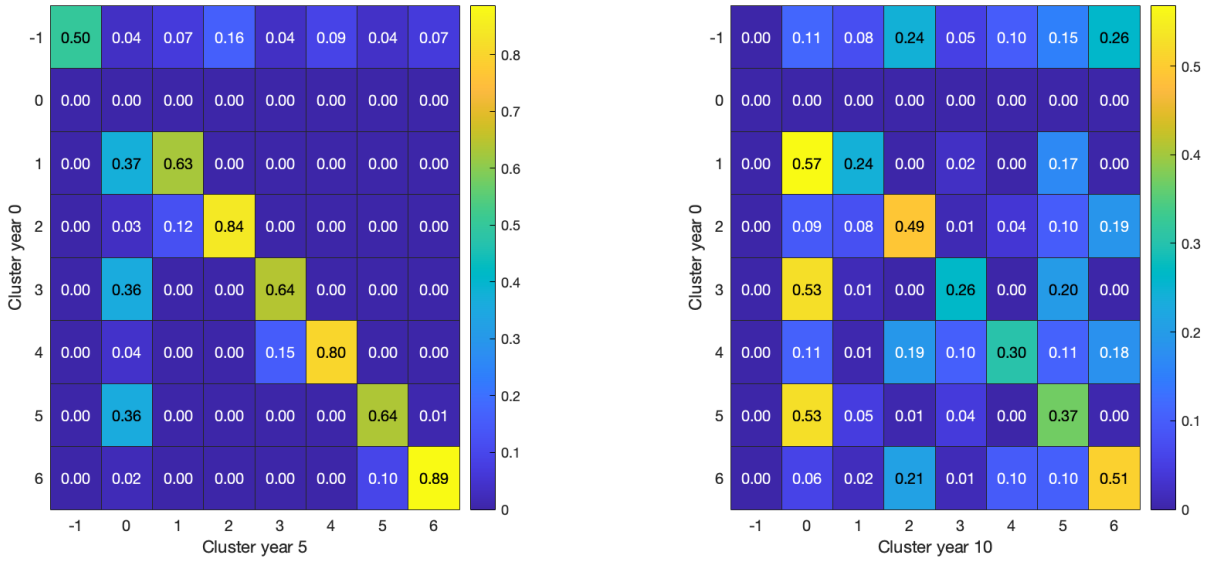

Figure 4.4: *Transition matrices* $P^{(2)}_{0\to5}$ *and* $P^{(2)}_{0\to10}$

The client distribution across clusters at each year is shown in Figure 4.5.
Highlighting the results at years 5 and 10, we obtain:

$$\begin{cases} \pi^{(2)}_0 = \begin{bmatrix} 2000 & 0 & 737 & 759 & 804 & 1386 & 770 & 544 \end{bmatrix} \\ \pi^{(2)}_5 = \begin{bmatrix} 1000 & 1006 & 689 & 962 & 805 & 1296 & 619 & 623 \end{bmatrix} \\ \pi^{(2)}_{10} = \begin{bmatrix} 0 & 1727 & 457 & 1243 & 524 & 699 & 1154 & 1196 \end{bmatrix} \end{cases} \tag{4.6}$$

As expected, these results are very similar to the first model, leading to the same conclusions:

- After five years, significant probabilities exist for moving only between clusters of the same Investor Type or dying.
- More than half the clients in the elder clusters (1,3 and 5) are likely to die within ten years, with over 35% dying within five years.
- It is more likely for new clients to enter younger clusters (as their probability of

entering the client base is inversely proportional to age).

- After ten years, some significant migration occurs between different types of investors, particularly from 1 to 5, 2 to 6, 3 to 5, 4 to 2 or 6, and 6 to 2.
- The most populated cluster at year 0, cluster 4, remains constant until year 6, after which it experiences a drastic decline.
- Non-investor clusters (5 and 6) are sparsely populated in the first years, but then sharply increase and become two of the three most populated clusters by year 10.
- Cluster 2 experiences a sudden increase in members and becomes the most populated cluster after ten years.

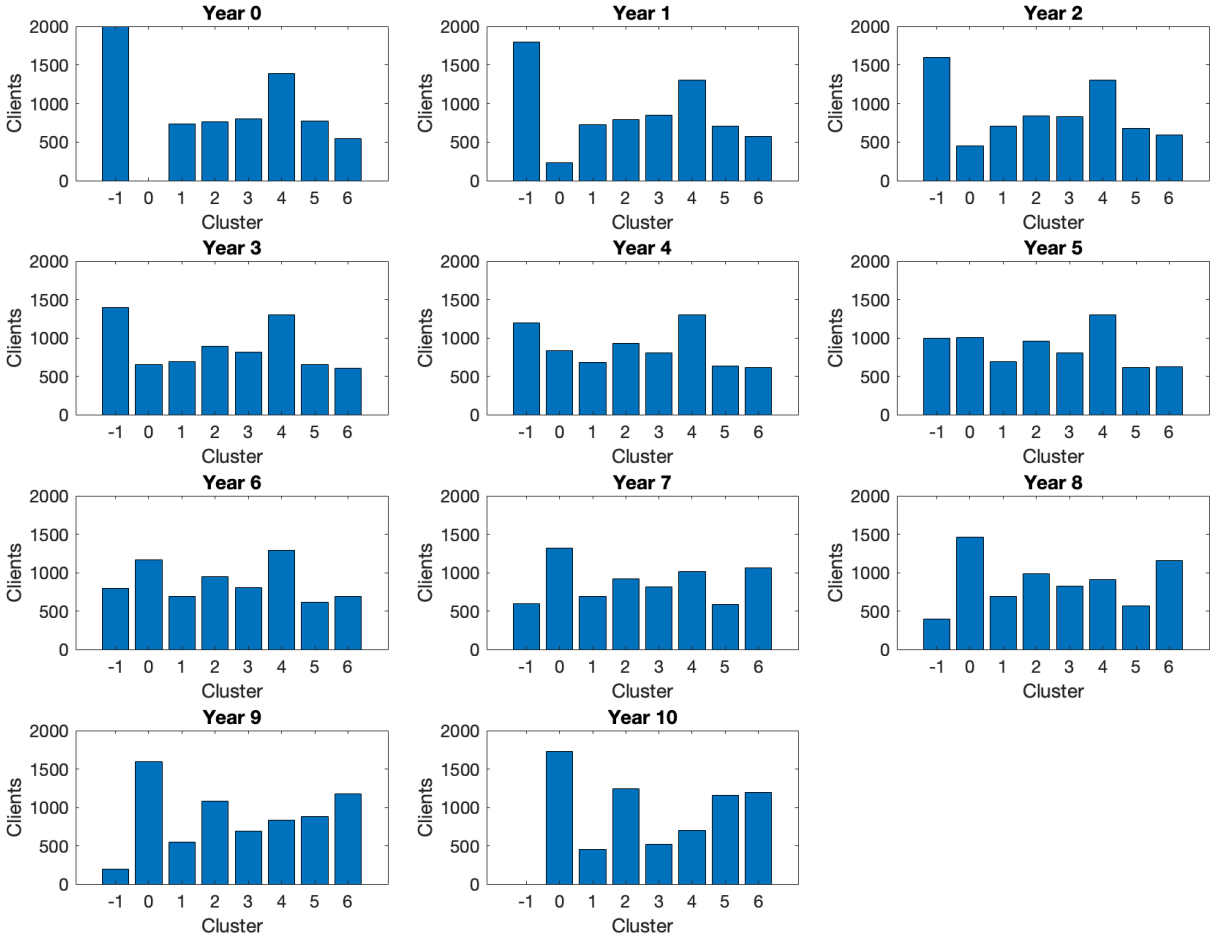

Figure 4.5: *Clients distribution at each year (Second model)*

4.6. Third Model results

The third model is the most detailed, as described in Section 3.4.
The first outcome of the algorithm is the transition matrix between each step.
Figure 4.6 shows the transition probabilities for a client to be in a cluster at year 5 and year 10, starting from a given cluster at year 0.

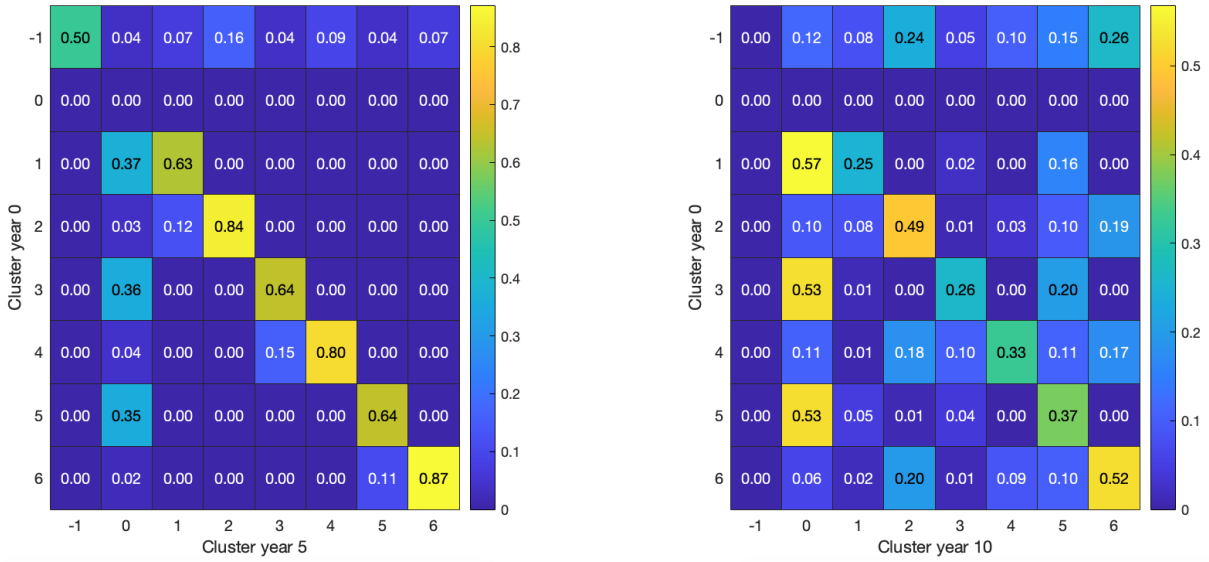

Figure 4.6: *Transition matrices* $P^{(3)}_{0\to5}$ *and* $P^{(3)}_{0\to10}$

The results with client distribution in each cluster over the years are shown in Figure 4.7.
Highlighting results at years 5 and 10, we have:

$$\begin{cases} \pi^{(3)}_0 = \begin{bmatrix} 2000 & 0 & 737 & 759 & 804 & 1386 & 770 & 544 \end{bmatrix} \\ \pi^{(3)}_5 = \begin{bmatrix} 1000 & 1005 & 692 & 960 & 807 & 1295 & 634 & 608 \end{bmatrix} \\ \pi^{(3)}_{10} = \begin{bmatrix} 0 & 1730 & 466 & 1210 & 522 & 727 & 1148 & 1196 \end{bmatrix} \end{cases} \tag{4.7}$$

Surprisingly, the conclusions are the same as the other models (this similarity will be analyzed in Chapter 6):

- After five years, significant probabilities are only for moving between clusters of the same Investor Type or dying.
- More than half of the clients in the elder clusters (1, 3, and 5) are likely to die within ten years, with over 35% dying within five years.
- It is more likely for new clients to enter younger clusters (their probability to join

the client base is inversely proportional to age).

- After ten years, significant migration occurs between different types of investors, particularly from 1 to 5, 2 to 6, 3 to 5, 4 to 2 or 6, and 6 to 2.
- The most populated cluster at year 0, cluster 4, remains constant until year 6, then experiences a drastic population decrease.
- Non-investor clusters (5 and 6) are sparsely populated in the first years, but sharply increase in the later years and become two of the three most populated clusters.
- Cluster 2 sees a sudden increase in members and becomes the most populated cluster after ten years.

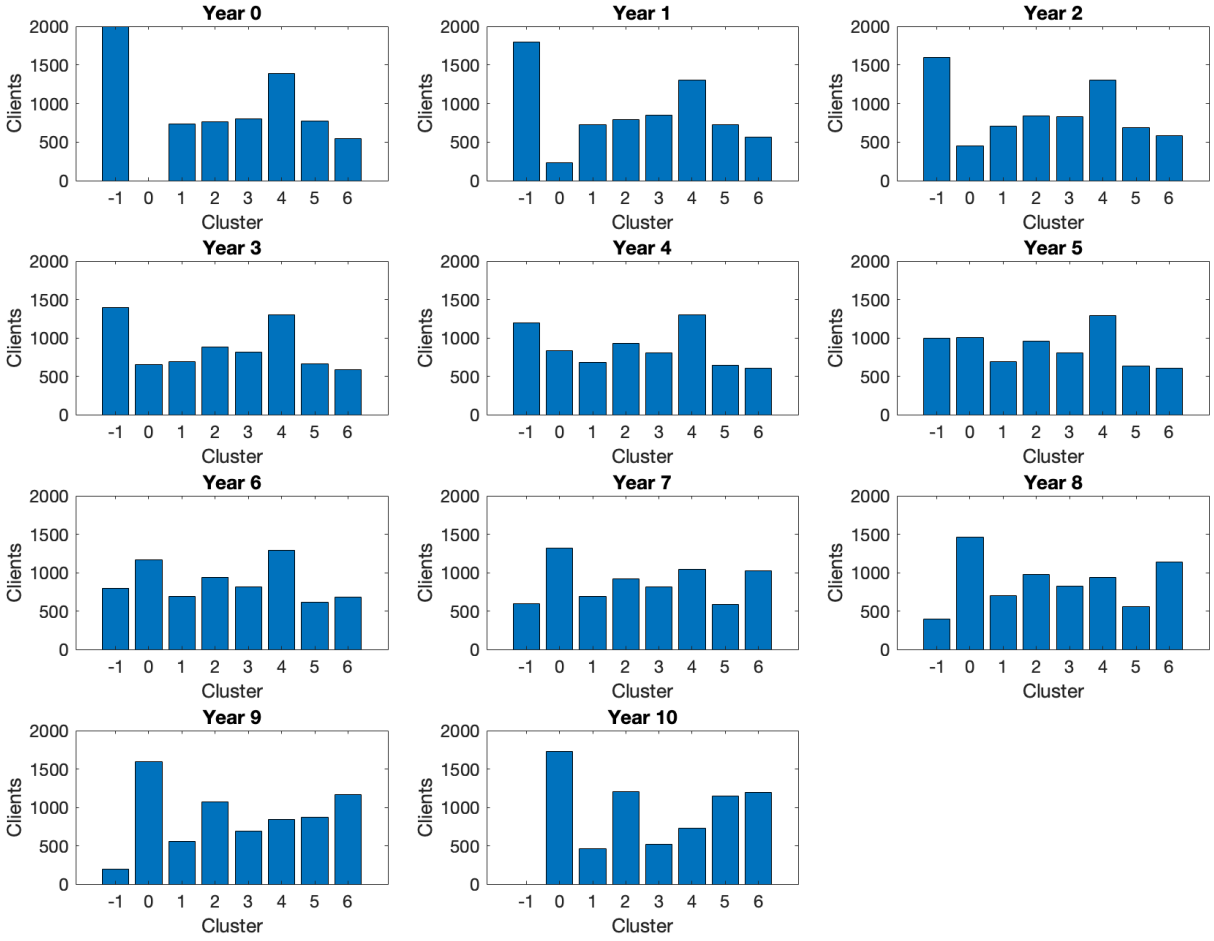

Figure 4.7: *Clients distribution at each year (Third model)*

4.7. Results comparisons

At first glance, the results from the three models appear closely aligned, with similar estimations. To assess their true similarity beyond this initial impression, we compare transition matrices and client distributions across different models for corresponding years, using both metrics and visualizations.

The cornerstone for each model is the transition matrices of the three models, $P^{(m)}_{0\to5}$ and $P^{(m)}_{0\to10}$, and the distributions at year 5 and 10, $\pi^{(m)}_5$ and $\pi^{(m)}_{10}$, where $m \in \{1, 2, 3\}$ represents the model.
Figure 4.8 shows the transition matrices from all three models are quite similar, exhibiting comparable migration patterns.

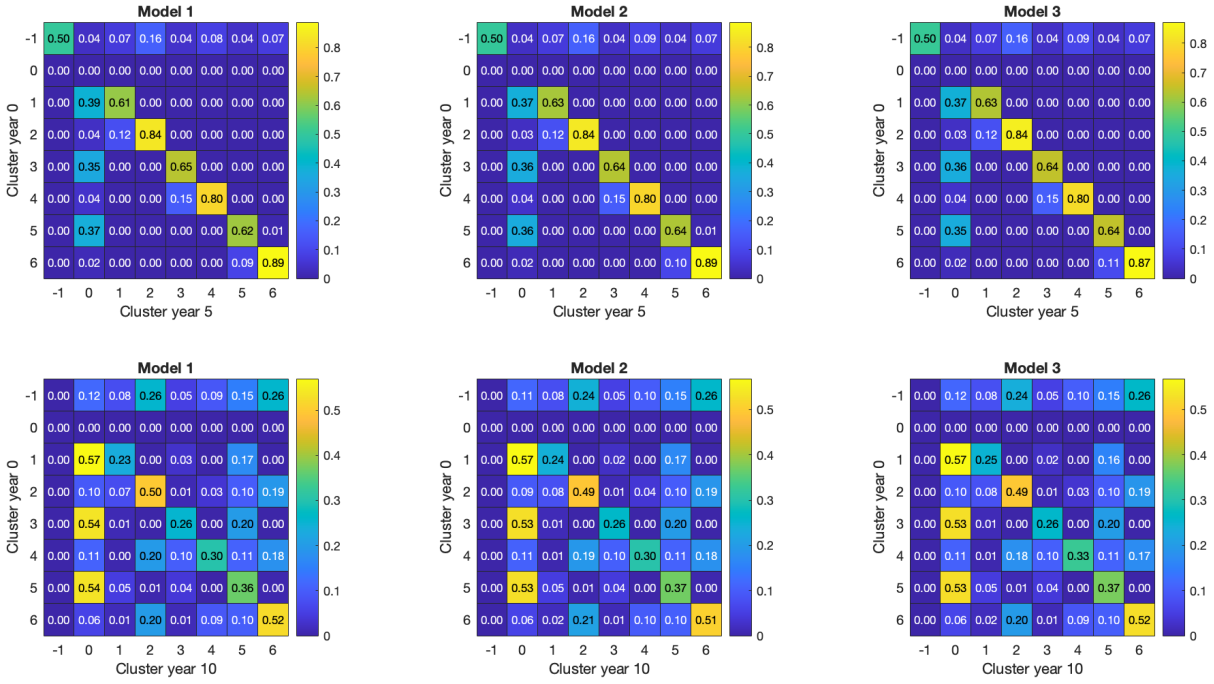

Figure 4.8: *Comparison of transition matrices $P_{0\to5}$ and $P_{0\to10}$*

To quantify the similarity between matrices, we use the **Frobenius Norm**.
The Frobenius Norm[9] of a matrix $A \in \mathbb{R}^{n\times m}$ is defined as $\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|a_{ij}|^2}$.
To evaulate how much two matrices A and B differ, an useful metric is the normalized norm of the difference:

$$d_F(A,B) = \frac{\|A-B\|_F}{\sqrt{\|A\|_F^2 + \|B\|_F^2}} \tag{4.8}$$

The results are as follows:

- $d_F(P^{(1)}_{0\to5}, P^{(2)}_{0\to5}) = 0.0105$
- $d_F(P^{(1)}_{0\to5}, P^{(3)}_{0\to5}) = 0.0146$
- $d_F(P^{(2)}_{0\to5}, P^{(3)}_{0\to5}) = 0.0079$
- $d_F(P^{(1)}_{0\to10}, P^{(2)}_{0\to10}) = 0.0156$
- $d_F(P^{(1)}_{0\to10}, P^{(3)}_{0\to10}) = 0.0232$
- $d_F(P^{(2)}_{0\to10}, P^{(3)}_{0\to10}) = 0.0188$

There is a slight increase in divergence between the five-year and ten-year estimates, as the five-year estimate is included in the ten-year one. However, the key takeaway is that all distances are below the 0.05 significance threshold, indicating that the matrices are comparable.

Another meaningful indicator of matrix similarity is the **Comparison of Eigenvectors**. While the eigenvectors are eight-dimensional, they can be visualized by reducing the dimensionality and considering the three-dimensional subspace generated by the first three principal components, which is the most representative subspace.

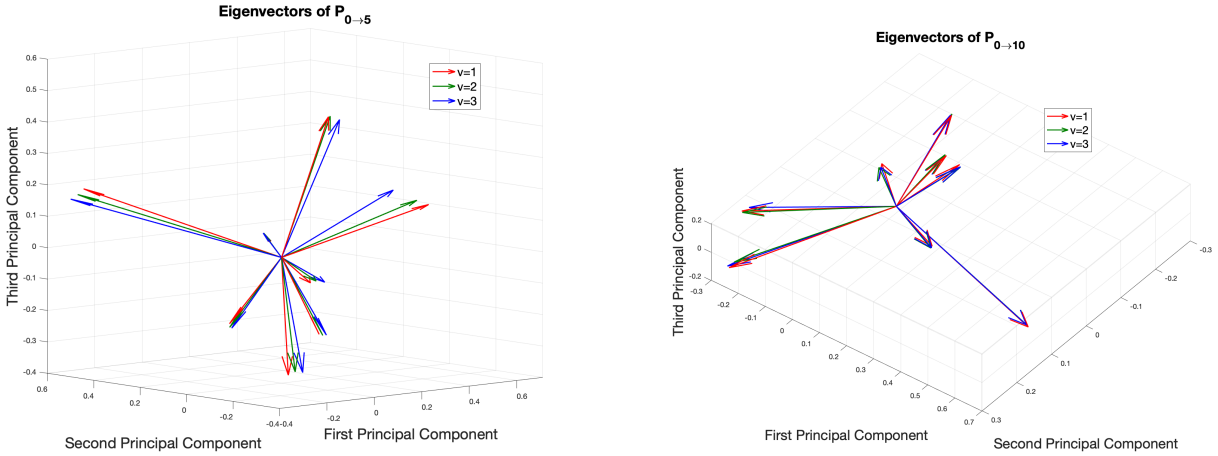

Figure 4.9: *Eigenvectors of the transition matrices in the subspace generated by the first three principal components*

Figure 4.9 illustrates that the eigenvectors align well, providing further visual confirmation of the matrices' similarity. Alongside the normalized Frobenius norm difference, this visual alignment reinforces that the matrices convey the same information.

Lastly, we compare **client distributions** at years 5 and 10 for each model. Figure 4.10 shows the population comparisons, which reveal a high degree of alignment across models.

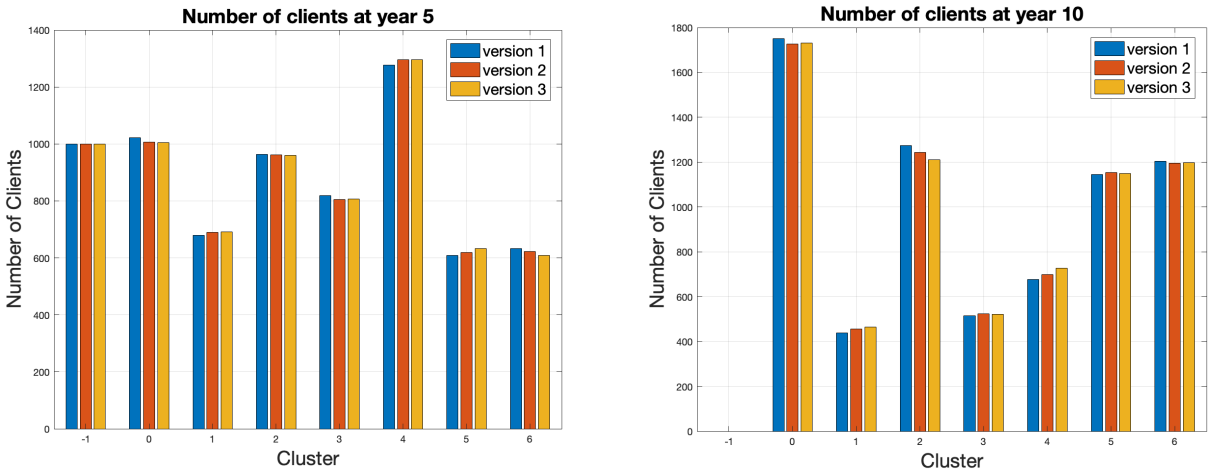

Figure 4.10: *Population distribution across clusters at year 5 and year 10*

The conclusion, supported by both graphical representations and numerical calculations, is that there is no need to prefer one model over another; the choice of model does not significantly influence the final results, particularly the client distribution after five and ten years. This finding, somewhat surprising, will be analyzed in Chapter 6.

Given this similarity, from now on, the number of clients each year will be estimated as the sample mean of the three models' results (which is very close to each individual value) and will be used to evaluate variations in the client base and guide decisions about the institution's future strategies.
Each estimate, rounded to the nearest integer, has a standard deviation of less than 2% of the mean, which we consider satisfactory for future developments.
The estimates, in numbers, are as follows:

$$\begin{cases} \pi_0 = \begin{bmatrix} 2000 & 0 & 737 & 759 & 804 & 1386 & 770 & 544 \end{bmatrix} \\ \pi_5 = \begin{bmatrix} 1000 & 1011 & 687 & 962 & 810 & 1289 & 620 & 621 \end{bmatrix} \\ \pi_{10} = \begin{bmatrix} 0 & 1736 & 453 & 1242 & 520 & 701 & 1149 & 1199 \end{bmatrix} \end{cases} \tag{4.9}$$

5 | Strategic Adaptation to Client Distribution Changes

Through the models presented and implemented in the previous chapters, we have obtained a realistic estimate of the populations of various personas over the years. To illustrate the usefulness and power of this methodological approach, we will present a concrete example that directly impacts Italian savers and the intermediaries who serve them. This example involves modeling changes in client nature and behavior and supporting a financial institution in managing these changes effectively, particularly in terms of **offering products and services**.
Clearly, this is just one of the many possible applications of our model, but we believe it is crucial to show one of them to highlight the model's potential.

Considering again the case analyzed in the previous chapters, one of the main results identified is a decline in investor clients and a rise in non-investors. Therefore, the strategy for the coming years must address these changes, including offering products tailored to non-investors. These clients tend to have a higher level of rejection towards traditional investment products, so we must focus on offering low-risk options closely tied to critical needs that they perceive as urgent and essential.

A fundamental suggestion for developing our operational strategy comes from the "Italy's Health Profile 2023" report[6] by the Organisation for Economic Co-operation and Development (OECD) and the European Observatory on Health Systems and Policies.
It highlights that health insurance, in various forms adapted to different client profiles, is quickly becoming essential to ensure timely access to primary healthcare services. As public healthcare systems face increasing demand, private insurance can help individuals avoid long waiting times and receive prompt medical attention, particularly in preparing for unforeseen events that may lead to minor or major health issues.

From a practical standpoint, we will now exclude the deceased clients (cluster 0) and storage clients (cluster -1) from our analysis. These clusters represent clients we cannot target, except through marketing strategies aimed at non-clients who may join our client

base. However, this expansion strategy falls outside the scope of our work, as we are focusing on the current client base and those expected to remain in the coming years.
Clearly, advertising campaigns are a critical element for industry growth and will be essential for increasing the number of clients. Our primary focus, however, will be on retaining and meeting the needs of existing clients.

5.1. Current Operational Strategy

Our strategy builds on the institution's current strategy, as outlined by Veronica Lucchetti in her thesis[12].
This plan primarily involves segmenting the client base (as of year 0) and using a recommendation system that, given a set of products, suggests the most suitable ones for each persona, typically recommending 3-4 products per cluster. The products, detailed in Subsection 2.5 are divided into two main categories:

Capital Accumulation Products

- Balanced Mutual Fund (ID 1)
- Defensive Flexible Allocation Unit-Linked (ID 6)
- Balanced Flexible Allocation Unit-Linked (ID 8)
- Cautious Allocation Segregated Account (ID 9)

These products are suitable for investors from clusters 3 and 4.

Income Accumulation Products

- Income Conservative Unit-Linked (ID 2)
- Balanced High Dividend Mutual Fund (ID 4)
- Fixed Income Segregated Account (ID 10)

These products are intended for clients in clusters 1 and 2.

Each product is identified by a product ID, which will be useful for the graphical analysis in subsequent sections (ID 0 indicates clients not interested in any product).
The preferences and quantities for each cluster are shown in Figure 5.1.

For future projections, we assume that the proportions of products requested by clients in investor clusters (1, 2, 3, and 4) will remain unchanged. This means that the distribution identified by Lucchetti is considered representative of the clients within each cluster. For instance, in cluster 1, 229 out of 737 clients in year 0 chose product 4, representing 31%.

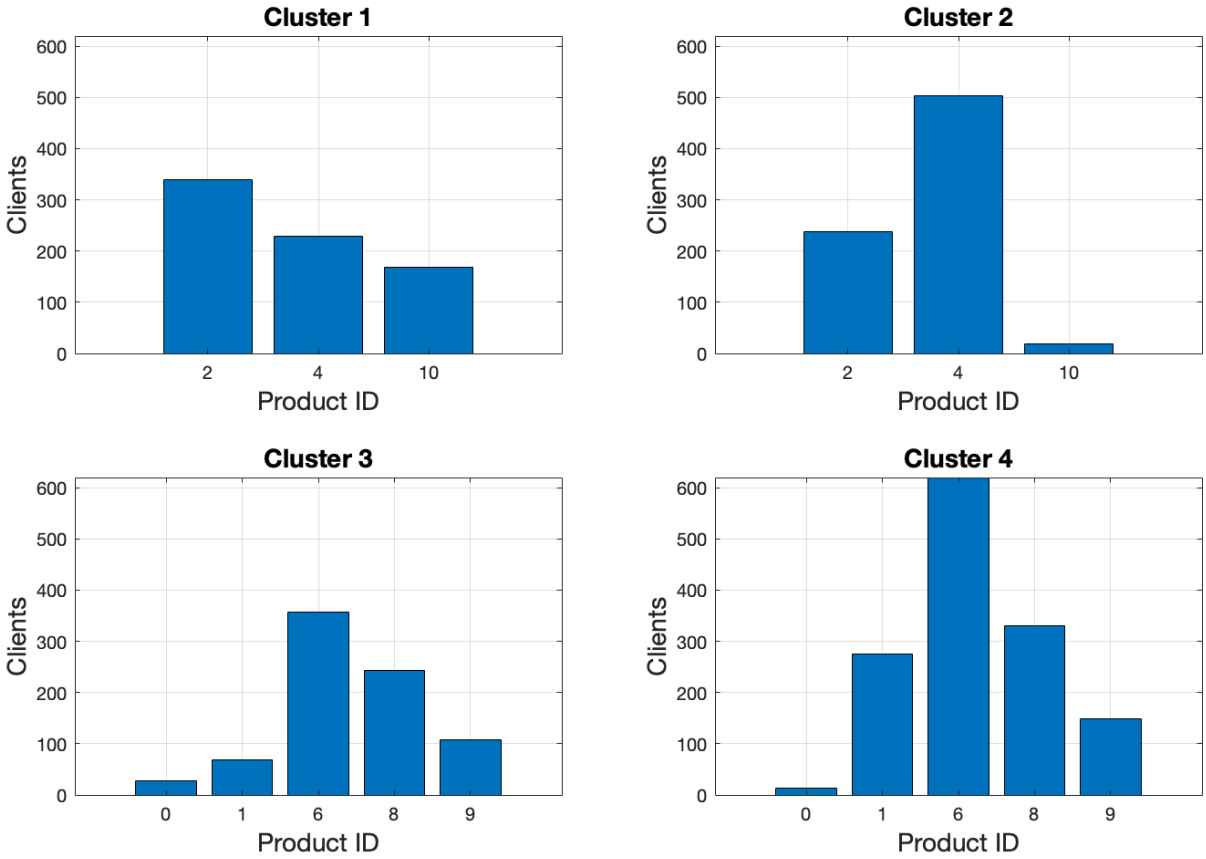

Figure 5.1: *List of products and corresponding units purchased at year 0*

We will maintain this proportion and apply it to the client population for clusters 1 at years 5 and 10.
Additionally, we assume that clients purchase a product when they enter an investor cluster (1, 2, 3, and 4) and discontinue it when they exit. This allows us to estimate the number of products purchased each year based on the number of clients in each cluster.

5.2. Analysis of Variations

Figure 5.2 shows the projected variation in client numbers across different clusters for years 0, 5, and 10, allowing for easy comparison.

The percentage change in client numbers from year 0 are as follows:

- **Cluster 1**: −6.8% by year 5, −38.5% by year 10
- **Cluster 2**: +26.7% by year 5, +63.6% by year 10
- **Cluster 3**: +0.8% by year 5, −35.3% by year 10

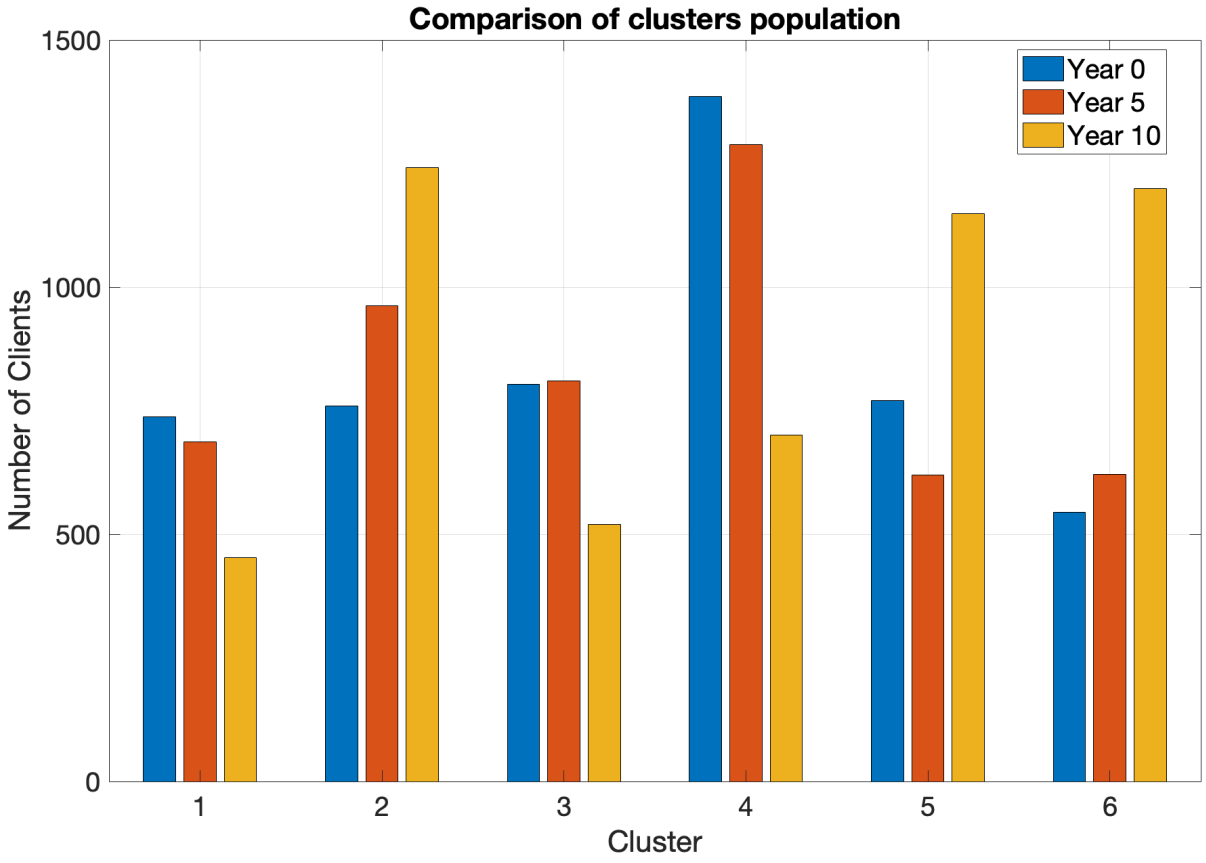

Figure 5.2: *Number of clients in each cluster at year 0, 5 and 10*

- **Cluster 4**: −7% by year 5, −49.4% by year 10
- **Cluster 5**: −19.5% by year 5, +49.2% by year 10
- **Cluster 6**: +14.2% by year 5, +120.4% by year 10

By combining these changes with the products associated with each cluster shown in Figure 5.1, we can predict the overall trend for product usage (Figure 5.3).

A significant insight is the marked increase in non-investor clients (clusters 5 and 6). These clients have not been prioritized in our current product offerings, as they were previously considered uninterested in financial products. However, it is now evident that we must address their needs by identifying products that can enhance their security and quality of life.

5.3. Future Strategies Recommendations

We strongly believe there is a product for everyone, as all individuals have needs related to their future. These needs may range from ambitions to safety concerns for themselves

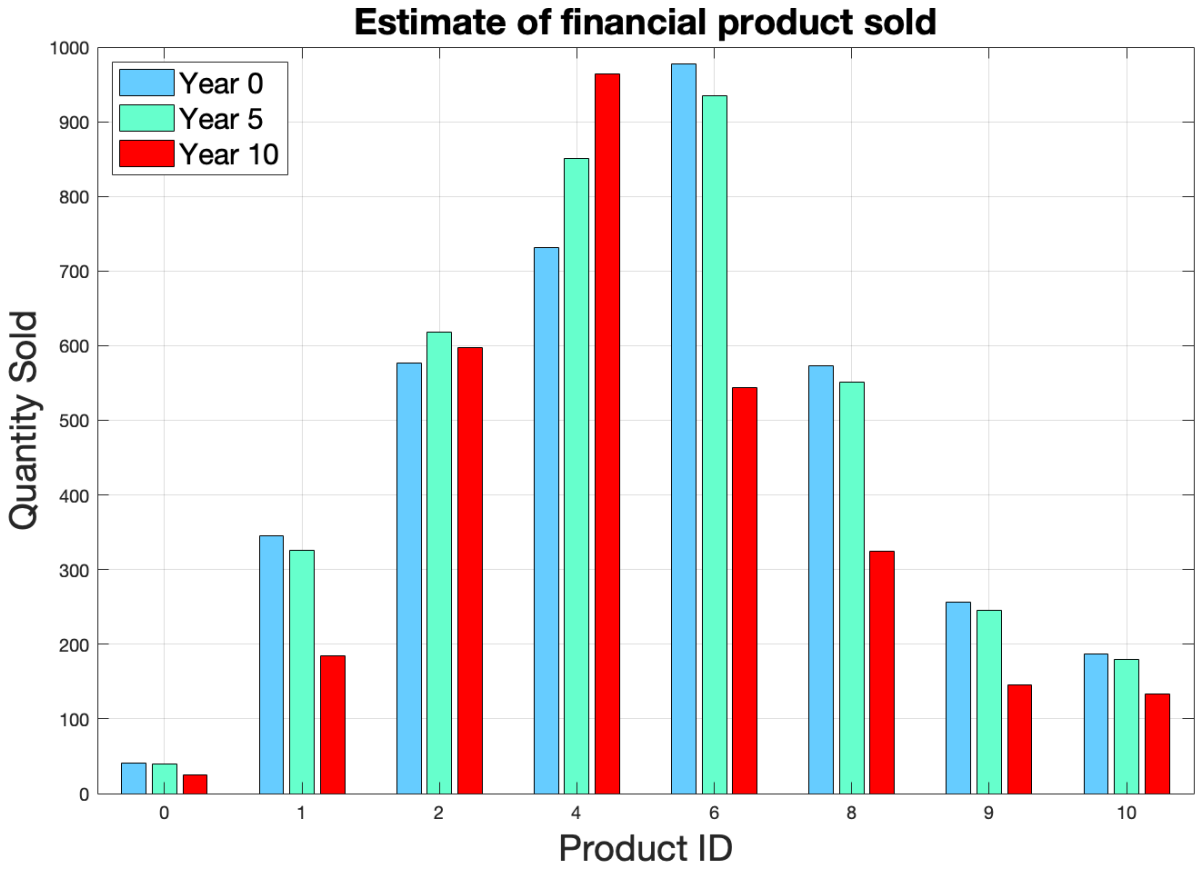

Figure 5.3: *Trends product purchases based on predicted client distribution*

and their loved ones.

Financial products, such as those in our current portfolio, are primarily tailored to individuals with specific financial plans who are capable of handling a certain degree of financial risk. This is the case for clients from clusters 1 and 2 (income investors) and clusters 3 and 4 (capital accumulator investors).

To address the needs of clients in non-investor clusters, we need to broaden our approach and consider insurance products that meet the specific needs of these segments.

Some of these products are already present in our current institutional portfolio, but two main issues prevent non-investors from being interested in them: firstly, the link between payments and financial indices, which makes them riskier than they could be (as these clients are not interested in handling market fluctuations); secondly, they are general insurance policies. The market is clearly shifting toward more specific policies that cater to the unique needs of each client based on their life stage.

In the following subsections, we will analyze each of the non-investor clusters in more detail to better understand their structure and characteristics. Using categorical variables, we will build a decision-tree-like structure to detect subclusters and better segment clients,

particularly to identify needs that can be matched by new insurance products. One crucial difference between cluster 5 and 6 is age, which clearly separates them, simplifying the potential segmentation.

5.3.1. Cluster 5: Scission and Insurance Needs

Cluster 5 consists of clients who are 58 years or older (considering age increments in 5 and 10-year steps). The age-based division creates just two branches. After testing different decision trees to identify specific client needs, we arrive at the definitive structure shown in Figure 5.4, where each node indicates the estimated number of clients at 5 and 10 years, and each leaf suggests suitable products for the subcluster defined by that leaf. Along with personalized recommendations, health insurance is becoming crucial in Italy and is a suitable suggestion for everyone.

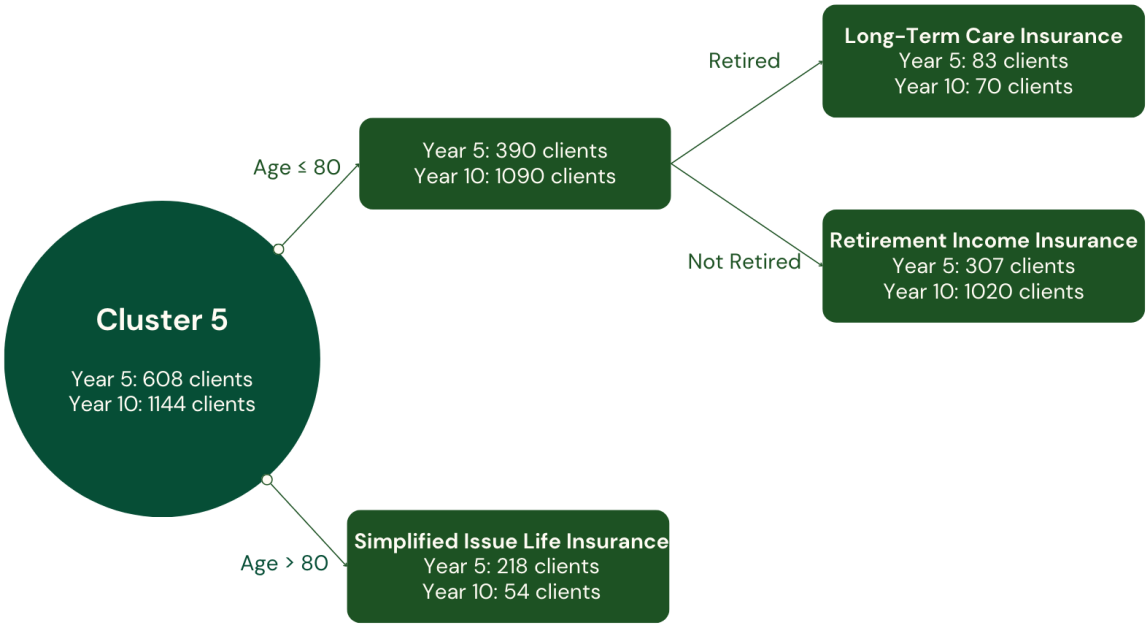

Figure 5.4: *Subclusters of Cluster 5 with client numbers and suggested products*

The product choices are explained as follows:

- **Clients under 80 years old and already retired** (83 after 5 years and 70 after 10 years): a Long-Term Care Insurance, combined with a Whole Life Insurance, would address the unpredictability of later life. Also, an Indexed Universal Life Insurance, linked to a financial index like the S&P 500, could be suitable.
- **Clients under 80 years old and not yet retired** (307 after 5 years and 1020 after 10 years): a Retirement Income Insurance, combined with a Health Insurance, would help with medical expenses and provide income post-retirement.

- **Clients over 80 years old** (218 after 5 years and 54 after 10 years): a Simplified Issue Life Insurance offers lower premiums without requiring medical exams, making it a favorable option for this group.

5.3.2. Cluster 6: Separation and Needs

Cluster 6 includes clients no older than 58 (again considering 5 and 10-year increments). The age-based division allows for a simple two-branch separation. After testing different decision trees to identify the best segmentation, we present the final structure in Figure 5.5. Each node reports the estimated number of clients at 5 and 10 years, and each leaf presents the product suitable for the subcluster defined by that leaf.
Here too, health insurance is becoming increasingly important and is recommended for everyone.

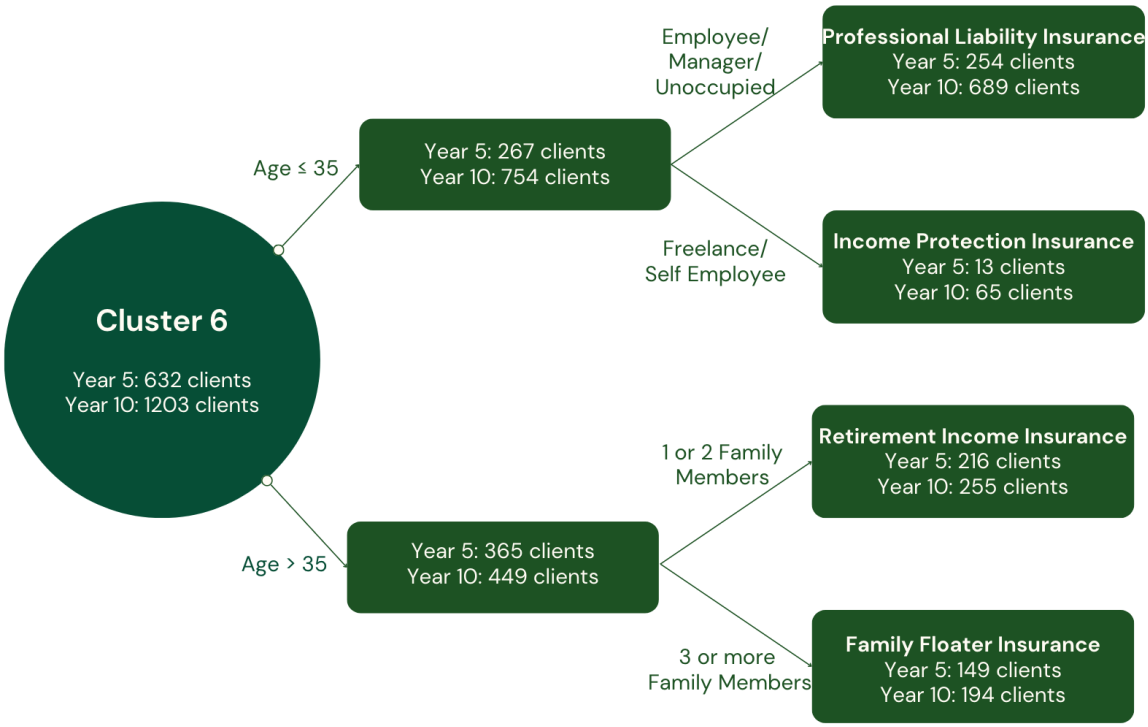

Figure 5.5: *Subclusters of Cluster 6 with client numbers and suggested products*

The product choices are explained as follows:

- **Young Freelancers or Self-Employed** (under 36 years old, 13 after 5 years and 65 after 10 years): a Professional Liability Insurance protects against negligence claims, or a Business Owner's Policy covers a range of damages and losses.
- **Young Employees or Unemployed** (under 36 years old, 254 after 5 years and 689

after 10 years): Accident Insurance or Income Protection Insurance offers regular payments to replace part of their income if they're unable to work due to illness or an accident.

- **Adults with Young Families** (over 36 years old, with more than three family members, 216 after 5 years and 255 after 10 years): College Savings Plans or policies to provide income protection in case of work disability are recommended. A Family Floater Insurance is another option for covering the entire family.
- **Adults with No Family** (over 36 years old, with fewer than three family members, 149 after 5 years and 194 after 10 years): a Health Insurance combined with a Retirement Income Insurance would provide both medical coverage and a pension supplement.

5.3.3. Summary of the Results and Forecasts

This section focuses on forecasting future product sales.

Although we cannot accurately predict how many clients from non-investor clusters will purchase the suggested products, we can rely on research and assumptions. A study from Bain[1] estimates that in 2023 25% of Italians have health insurance, and this number is steadily increasing. Since health insurance is a subset of the products we consider, we estimate that 40% of clients from clusters 5 and 6 will purchase these products.

This estimate is reasonable given the growing interest in personalized insurance products, and we expect that over a 10-year period, demand will continue to grow. Figure 5.6 presents the projected increase in product volume.
We are aware of the roughness of the estimate, which assigns the same value to products that are quite different. However, we believe it can still serve as a useful basis for comparison and for assessing the growth or decline of our institution.

Data indicates that the total volume of annual purchases increases by 15.1% at year 5 compared to year 0, and by 4.6% at year 10.
While we are satisfied with these estimates, we must address the dip between years 5 and 10, which will be the focus of a future business plan, when part of the future predicted through our models will become a reality rather than a stochastic estimate.

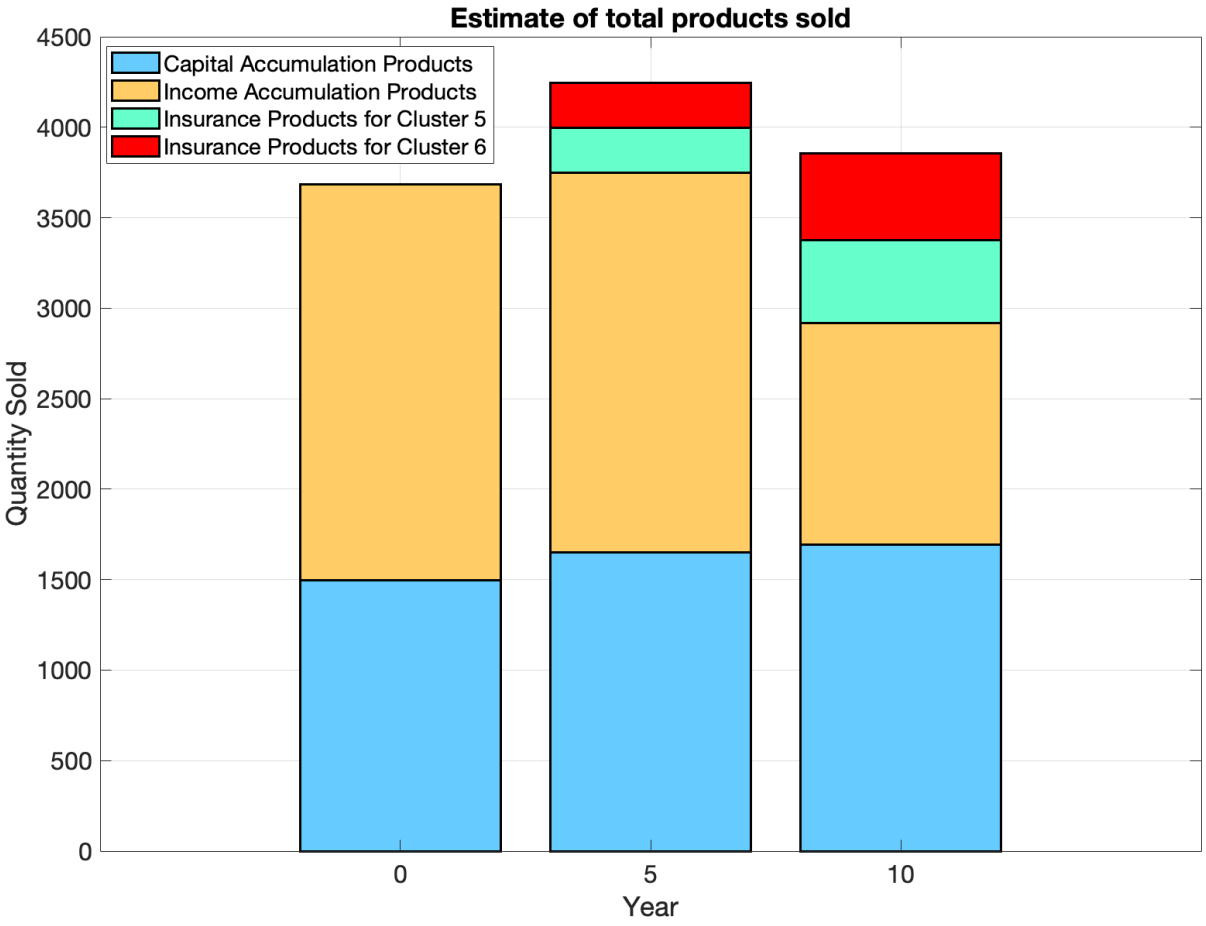

Figure 5.6: *Comparison of total purchase volumes across different years*

6 | Discussions

After conducting the quantitative analysis and gaining a comprehensive view of our work, we can discuss the results and the process implemented.
The primary takeaway is not computational but methodological: we believe that our model extends beyond the financial domain, as will be discussed in Section 6.3.
The aim of this thesis was not to claim the ability to predict the future of a financial institution (a task fraught with uncertainty, especially over a ten-year horizon) but rather to create a plausible, real-world scenario based on sound and reliable assumptions (Section 6.1), of how the client distribution may evolve over the years.

To demonstrate the utility of this model, we utilized it in one of its many potential applications: making informed decisions to address the evolving client base indicated by the model, as well as suggesting potential future directions, particularly in forecasting client needs and identifying suitable financial or insurance products for them.
In light of the outcomes, we can confidently assert that the model's primary objective has been achieved: the development of a flexible system that, in synergy with expert insights and intuition, can serve as a critical tool for industries to navigate changes and plan for the future.

6.1. Key Findings

The model we developed, with minimal recommendations and adjustments, is versatile and applicable to a wide range of scenarios, not just within the financial sector.
Its core strength lies in having made the fewest assumptions possible:

- The availability of a reliable dataset with variables of different types, where a quantity of interest (in our case, the Investor type) is defined for each client at year 0, alongside other features that describe the client base.
- Clients can be categorized into a small number of personas, meaning individuals who can be clearly identified and described through specific categories and numbers, and grouped based on similar needs.

- The possibility to model the evolution of clients over time, driven by rational assumptions.

In our particular case study, the results presented in Chapter 4 provide a plausible forecast of the client base over 5 and 10 years, with key findings such as:

- A significant increase in non-investor clients, especially after 10 years (Clusters 5 and 6, +78.6%).
- A notable decline in investor clients after 10 years (Clusters 1, 2, 3, and 4, -20.9%), leading to a corresponding reduction in financial product sales.
- A refined clustering system that enables the company to offer tailored insurance products to non-investor clients based on their subcluster, potentially expanding the company's offerings.
- Age emerges as a key distinguishing factor between clusters defined by the same Investor type, confirming findings from Lucchetti's thesis[12].

The scenario we have outlined is realistic and effectively addresses the main issues discussed in the Introduction, namely the development of a wide range of products dedicated to older people, and the significant differences between the products demanded by younger and older individuals.
In the next section, we will provide our interpretation of these results as the authors and developers of this model.

6.2. Interpretations

The first key interpretation concerns the increase in non-investor clients. This outcome is closely linked to both the evolutionary causes driving client behavior and an additional component: the way the "storage cluster" is modeled, which reflects a real-world phenomenon where younger individuals are more likely to enter the client base.
Since non-investors tend to be younger, this modeling approach significantly boosts the proportion of non-investors entering the client base.

Another striking result, and perhaps the most surprising, is the consistency of outcomes across the three different evolutionary models presented in Chapter 3.
Specifically, the transition probabilities of the $P_{0\to10}$ matrices remain similar, even between relatively distant clusters (Figure 4.8).
This subtle finding may be due to the synthetic nature of the data, a possibility we raised when analyzing the data distribution in Chapter 1 With few numerical differences between

the data points (except for Age, which, as discussed in Chapter 2, is a crucial factor in distinguishing client clusters), the driving factors seem to be categorical rather than numerical variables.
As the importance of the investor coefficient diminishes over time, clients are categorized into clusters based more on their categorical features and less on numerical ones, as highlighted by Lucchetti's cluster characterization in Section 2.4).
Despite these nuances, the overall scenario of client evolution appears realistic. Approximately 40% of clients do not change clusters over the course of 10 years, and around a quarter transition between investor clusters. The remaining clients either move to storage or exit (via death), which does not count as a change in investor type. These results give us confidence in the parameter tuning used to quantify the variance in model definitions (Chapter 3).

As for the strategy development, this section is meant to showcase just one practical application of the model, using a decision-tree-like structure to segment clusters effectively. Although our decisions are not the only possible ones, age, as the primary factor in grouping personas, naturally forms the first level of branches in Clusters 5 and 6. Subsequent divisions are made primarily using categorical variables (with Family Size, an ordinal variable, also playing a role), aiming to define personas that are understandable and actionable.

6.3. Implications

Our aim in this work was not merely to create a financial evolutionary scenario showing how investors change over time, but rather to develop a broader model that, with minimal adjustments, could serve as a cornerstone in various contexts involving the evolution of people and the categorization of personas.
This could range from patients with health issues to client profiling, student classifications, and many other scenarios. The evolutionary nature of personas makes it an adaptable system for modeling any field where individuals change and shift between different groups. Even the potential applications of the outcomes are numerous, given the variety of personas that can be studied.
While we showed the implementation of a specific strategy, this framework could also be used to manage medicine storage based on patient types and quantities, develop targeted marketing strategies for predicted clients, optimize transportation line energy consumption based on future citizen behavior, and explore many other possibilities.

Regarding the state of the art, the use of Markov Chains (and more boadly, Markov

Processes) in financial agent-based models (to which our study belongs), such as analyzing optimal markets for investment, calculating returns, or managing portfolios, has already been explored in various works, including those by Mettle and colleagues [16], [15], Chorn [4], and Vaninsky [23].
However, our work presents an innovative approach by focusing on real-world future scenarios of investors, specifically by identifying personas and the evolution of clients, rather than focusing on markets and stocks.
Another relevant body of literature related to our work involves agent-based evolutionary models that explain observed processes, while less attention has been given to forecasting future scenarios. Current methods excel at short-term market forecasting, but they often struggle to capture the long-term, individual-level evolution of investors.
This path is less explored (given its intrinsic complexity) but is crucial. Our approach addresses this gap by concentrating on persona evolution and providing insights into the changing needs and behaviors of investors.

The scalability of this model allows it to be applied to a broader range of clients, including those from different socioeconomic backgrounds, international markets, and larger datasets. Its flexible structure also makes it easily integrable with other predictive technologies, offering a comprehensive framework for understanding how clients evolve across diverse cultural and financial contexts.
Future advancements could incorporate machine learning techniques to enhance the model's predictions by introducing additional variables, such as sentiment analysis or behavioral economics. Additionally, integrating this model with big data analytics could enable more accurate projections, offering deeper insights into the dynamic changes within increasingly complex financial systems.

6.4. Limitations

The dataset used is complete and free from missing or unreliable data, which is one of its major strengths. While having more than 5,000 clients would have been ideal, the use of generated data as a storage mechanism was an effective way to address the smaller sample size.

Regarding the specific models of our case study, we assumed that each client always belongs to a single cluster at any given time. However, this could be modified into a probability distribution where clients belong to multiple clusters in varying percentages simultaneously.
In addition, our model assumes that the number of personas remains fixed over time,

whereas an alternative approach could be to create new groups when many clients deviate significantly from existing persona centers.
Another limitation stems from the fixed nature of the categorical variables in the data. This could be addressed if temporal trends were available, allowing for dynamic modeling of factors such as changes in living location or employment.

However, the main focus of our thesis is not on the specifics of our model but rather on the fact that the proposed ones are just one example among many possible implementations. For instance, an entirely different model could use partial or complete time series of the relevant features, allowing randomness to be more influenced by data trends. A financial institution could realistically gather such data, which would also facilitate the identification of distinct evolutionary paths specific for each personas.

6.5. Recommendations

The decision to use a time-variant Markov chain was driven by its adaptability, which surpasses that of homogeneous models. One of its key advantages is the ability to account for dynamics as the "Evanescent Investor Coefficient" (described in Subsection 3.1.4), resulting in transition matrices that vary significantly over time (as shown in Section 4.2). This cannot be adequately captured by time-homogeneous transition matrices.
The trade-off of using this tool, compared to time-homogeneous models, is the loss of certain asymptotic properties, which we consider of secondary importance in our work since the time horizon is well-defined and stationary distributions are not involved.

The main takeaway is that our framework provides extensive flexibility, not only in parameter settings (though even discussing about parameters is somewhat reductive in this case) but also in a broader sense, allowing for infinite variations in modeling.

To navigate this flexibility, we, as authors, have two primary recommendations:

- The entire framework must rely on a strong definition and profiling of personas, such as in Lucchetti's work, because everything is built upon this foundation. If the characterization misses part of the dataset or fails to differentiate between personas with significant differences, the process of classifying clients into these groups over time becomes meaningless.
- When evaluating the evolutionary model, it is essential to test the "parameters" of the random variables through several trials. Specifically, extreme trials can provide valuable insight into how each parameter influences the evolution of the system. This trial process serves as an important intermediate step in understanding the

dynamics of the model.

We firmly believe that our models can contribute valuable insights and practical suggestions to the complex and expansive field of agent-based recommendation systems, where forecasting is as challenging as it is fascinating.

Conclusions

The study successfully demonstrated, through a concrete case study, the potential of Time-Variant Markov Chains to forecast how the client base of a financial institution can evolve over a period of interest and how to address these changes.

Based on a reliable algorithm for implementing agent-based recommendation systems (described in Veronica Lucchetti's work[12]), three different models of client feature variations were created, each with increasing complexity (from varying only Age to including all numerical features in the evolution).
The surprising result was that all three evolutionary models led to the same outcome, which can be summarized as an increase in non-investor clients and a decrease in investors, resulting in fewer financial products being sold.
To address these findings and show their practical use, we proposed an example of operational strategy for the future of the institution, which includes adding insurance products tailored to different subclusters within the non-investor groups.

This work not only provides a practical example of a plausible future scenario (and its implications) but also offers insight into how these models (just requiring evolutionary rules and a suitable clustering technique) can be easily extended beyond the financial sector to all contexts involving agent-based recommendation systems.
The model's potential and adaptability are its main strengths, making it a potentially winning tool for designing successful strategies in any field where agent evolution is involved, of which there are countless.

Bibliography

[1] Bain & Company. Sanità italiana in crescita: le assicurazioni devono investire ed evolvere, 2023. URL https://www.bain.com/it/about-bain/media-center/press-releases/italy/2023/Sanita-italiana-in-crescita-assicurazioni-devono-investire-ed-evolvere/. Accessed: 2024-09-09.

[2] Bank of America. Bofa private bank study of wealthy americans finds generational shift in attitudes towards wealth, June 2024. URL https://newsroom.bankofamerica.com/content/newsroom/press-releases/2024/06/bofa-private-bank-study-of-wealthy-americans-finds-generational-.html. Accessed: 2024-09-09.

[3] D. D. O. Batista and G. A. Giraldi. Comparative study of cluster validation indices in the context of violent crime hot spots detection. *Expert Systems with Applications*, 65:98–107, 2016. doi: 10.1016/j.eswa.2016.08.033.

[4] L. Chorn. Portfolio management framework for multistage investments. *Proceedings - SPE Annual Technical Conference and Exhibition*, 09 2004. doi: 10.2118/90700-MS.

[5] Citi. Research report, 2024. URL https://www.citifirst.com.hk/home/upload/citi_research/rsch_pdf_30259888.pdf. Accessed: 2024-09-09.

[6] European Commission. Italy: Country health profile 2023, 2023. URL https://health.ec.europa.eu/document/download/67cd0b86-b081-4fa5-84a8-f4487e912320_en?filename=2023_chp_it_english.pdf. Accessed: 2024-09-03.

[7] Gretel.ai. Gretel.ai - synthetic data for developers, 2024. URL https://gretel.ai/. Accessed: 2024-08-13.

[8] R. W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, 1950. doi: 10.1002/j.1538-7305.1950.tb00463.x.

[9] R. Herzog, F. Köhne, L. Kreis, and A. Schiela. Frobenius-type norms and inner

products of matrices and linear maps with applications to neural network training, 2023. URL https://arxiv.org/abs/2311.15419.

[10] Human Mortality Database. Human mortality database, 2024. URL http://www.mortality.org. Accessed: 2024-07-27.

[11] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967. doi: 10.1007/BF02289588.

[12] V. Lucchetti. Data-driven customer segmentation: A needs-based cluster analysis for optimizing financial product recommendations. Unpublished master's thesis, 2023-2024.

[13] S. J. K. Marco Bazzi, Francisco Blasques and A. Lucas. Time varying transition probabilities for markov regime switching models. Discussion Paper 14-072/III, Tinbergen Institute, 2014. URL https://papers.tinbergen.nl/14072.pdf.

[14] N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[15] F. O. Mettle, E. N. B. Quaye, and R. A. Laryea. A methodology for stochastic analysis of share prices as markov chains with finite states. *SpringerPlus*, 3(1):657, 2014. doi: 10.1186/2193-1801-3-657. URL https://doi.org/10.1186/2193-1801-3-657.

[16] F. O. Mettle, E. K. Aidoo, C. O. N. Dowuona, and L. Agyekum. Analysis of investment returns as markov chain random walk. *International Journal of Mathematics and Mathematical Sciences*, 2024(1):3966566, 2024. doi: https://doi.org/10.1155/2024/3966566. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2024/3966566.

[17] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901. URL https://doi.org/10.1080/14786440109462720.

[18] P. Rousseeuw, I. Ruts, and J. Tukey. The bagplot: A bivariate boxplot. *American Statistician - AMER STATIST*, 53:382–387, 11 1999. doi: 10.1080/00031305.1999.10474494.

[19] S. Senthilnathan. Usefulness of correlation analysis. *SSRN Electronic Journal*, 07 2019. doi: 10.2139/ssrn.3416918.

[20] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[21] J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531, Vancouver, 1975.

[22] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf`.

[23] K. Vaninsky, S. Myzuchka, and A. Lukov. A multi-agent nonlinear markov model of the order book. *arXiv: Trading and Market Microstructure*, 2012. URL `https://api.semanticscholar.org/CorpusID:154154486`.

List of Figures

1.1 Investor Type Distribution 5
1.2 Age Distribution 5
1.3 Debt Distribution 5
1.4 Job Distribution 5
1.5 Pairwise correlation between quantitative variables 7
1.6 Variables' scores for the first four principal components 8
1.7 Bagplot of three pairs of variables highlighting client 2218 as an outlier in the bivariate distributions, all related to Financial Education 10
1.8 Univariate (normalized) densities divided by type of investors 11

2.1 Datapoints distribution in 2-dimesional and 3-dimensional t-SNE reduction 15
2.2 Dendrogram obtained with Average Linkage 16
2.3 Risk Propensity Index composition 22

3.1 Probability of dying within a year based on age and gender in Italy for 2019 30
3.2 Evanescence of "Investor Type at year 0" weight 31
3.3 Second-order regression polynomial of Income distribution 33
3.4 Second-order regression polynomial of Wealth distribution 34

4.1 Difference between transition matrices in early and advanced years 39
4.2 Transition matrices $P^{(1)}_{0\to5}$ and $P^{(1)}_{0\to10}$ 40
4.3 Clients distribution at each year (First model) 41
4.4 Transition matrices $P^{(2)}_{0\to5}$ and $P^{(2)}_{0\to10}$ 42
4.5 Clients distribution at each year (Second model) 43
4.6 Transition matrices $P^{(3)}_{0\to5}$ and $P^{(3)}_{0\to10}$ 44
4.7 Clients distribution at each year (Third model) 45
4.8 Comparison of transition matrices $P_{0\to5}$ and $P_{0\to10}$ 46
4.9 Eigenvectors of the transition matrices in the subspace generated by the first three principal components 47
4.10 Population distribution across clusters at year 5 and year 10 48

5.1 List of products and corresponding units purchased at year 0 51
5.2 Number of clients in each cluster at year 0, 5 and 10 52
5.3 Trends product purchases based on predicted client distribution 53
5.4 Subclusters of Cluster 5 with client numbers and suggested products 54
5.5 Subclusters of Cluster 6 with client numbers and suggested products 55
5.6 Comparison of total purchase volumes across different years 57

List of Tables

2.1 "Retirees from the North" numerical features 17
2.2 "Young Employees from the North" numerical features 17
2.3 "Affluent Elderly Employees from the North" numerical features 18
2.4 "Young Employees from the South" numerical features 19
2.5 "Retired Non-Investors" numerical features 19
2.6 "Young Wealthy Sensible Savers" numerical features 20

List of Algorithms

2.1 Recommendation Algorithm from Lucchetti's thesis 22
3.1 Identification of clients' investor type each year 29

Acknowledgements

Riconoscimenti

Questa tesi è la conclusione di un lunghissimo percorso che ho iniziato tanti anni fa; non è stato facile e men che meno perfetto: è stato la mia vita, l'ho affrontato a modo mio, ho imparato tantissime cose e ora posso dire di andarne fiero.

Per iniziare, ringrazio il professor Marazzina per essere stato il mio relatore e Veronica per tutto il lavoro fatto insieme, l'infinito supporto e la positività che mi ha sempre trasmesso. Ringrazio poi il professor Zenti, per come ha messo a mia disposizione le sue profonde conoscenze della materia e per avermi mostrato concretamente un metodo di lavoro preciso e acuto, che porterò sempre con me come esempio. Lo ringrazio anche per la fiducia che è riuscito a trasmettermi nei momenti in cui ne avevo bisogno; sono orgoglioso della tesi che abbiamo prodotto, e gran parte del merito è suo.

Ringrazio tutte le altre persone che hanno fatto parte del mio percorso universitario: in particolare voglio citare Ale, che dal primo giorno mi ha fatto sentire a casa ogni volta che eravamo insieme, Debe, per la genialità e per essere stato sempre e inguaribilmente sè stesso, e Gio, perchè insieme abbiamo vissuto di tutto e condiviso dubbi, fatiche, traguardi e successi.
Grazie anche a tutti i coinquilini, compagni e professori che hanno condiviso qualcosa con me, sono tantissimi e li ricorderò con affetto.

Nel momento poi di pensare a tutte le altre persone che fanno parte della mia vita, mi rendo conto di quante sono e della fortuna che ho.
Ringrazio innanzitutto mamma e papà, per l'infinito amore che provate per noi tre e che ci mostrate ogni giorno, siete sempre stati la mia ispirazione e lo sarete sempre; vi ringrazio anche perchè avete avuto tanto coraggio ad ammettere a voi stessi che non si può essere perfetti, nè come persone nè tantomeno come genitori, e per aver anteposto il mio bene a qualsiasi altra cosa. Non avrei potuto desiderare nulla di più.
Ringrazio Guido e Albi per le infinite cose che ci uniscono e per tutto quello che viviamo insieme; ognuno di noi ha il suo stile inconfondibile, conoscerci e capirci sarà sempre la

ricchezza più grande.
Ringrazio Anna, per aver visto ogni difficoltà e avermi spronato sempre, per l'affetto e la comprensione che sai dimostrarmi, per il tuo modo di essere che è speciale. Poter condividere la vita con te è un privilegio immenso, e non immagini la forza che mi dà averti accanto ogni giorno.
Ringrazio il nonno, la nonna Pina, la nonna Lidia, lo zio, la zia, Marco, Giulio e la Ciana per la famiglia che siamo e per quanto importanti siete per me.
Ringrazio Andrea, per la sincerità con cui ci raccontiamo la vita e perchè non ti stanchi mai di ripetermi lezione più importante: che la felicità è imparare ogni giorno ad essere sè stessi.
Grazie a Marco, Gio, Albo, Chicco, Paco, Filo e Tito per la fratellanza che ci unisce, al mio Clan e alla mia Comunità Capi per tutta la strada fatta insieme, a Ciuppi e Carlo perchè l'esempio che mi avete dato mi ispira e mi sprona a cercare di essere per gli altri ciò che voi siete stati per me.

Infine, un ringraziamnto ad una persona che difficilmente leggerà queste parole ma a cui tengo particolarmente, che è Neil Gaiman, per aver scritto un capolavoro come The Sandman.
Lo ho scoperto per caso, ma la provvidenza ha voluto che lo leggessi nel periodo in cui lavoravo alla tesi; ho alternato spesso la fantasia dell'uno alla la razionalità dell'altra, e l'ispirazione che mi ha dato questa combinazione è stata immensa.
Sono profondamente convinto che opere come questa arricchiscano il mondo, ed anche di questo sono grato.