UNIVERSITAT DE BARCELONA

Compositional data for analysis in economics and finance

Juan David Vega Baquero

PhD in Economics

Compositional data for analysis in economics and finance

Juan David Vega Baquero

UNIVERSITAT DE BARCELONA

PhD in Economics

Thesis tittle:

Compositional data for analysis in economics and finance

PhD candidate:
Juan David Vega Baquero

Advisor:
Miguel Santolino

Date:
September 2025

UNIVERSITAT DE BARCELONA

Acknowledgements

If I could tell the 17-year-old version of myself – the one just starting an undergraduate degree in Economics – where I am now, he would not believe a word of it. What a ride it has been! And none of this would have been possible without the support of so many people around me.

To begin, I would like to express my gratitude to my advisor, Dr. Miguel Santolino. Thank you for your patience (which I may have tested beyond its limits at times) and your unwavering support. You were always there, offering guidance whenever I felt lost, and you always believed in my abilities, even when I did not.

I would also like to thank my family. Without you even realizing it, your guidance and support since my childhood have paved the way for this moment. Even from afar, knowing I have a safe space to return to has given me the confidence to go farther than I ever imagined.

And what would this journey be without friends? I am too shy to name each of you individually (and I would surely forget someone), but you all know who you are. Each of you has been there whenever I needed, and I hope you feel the same way.

To my friends from Colombia: Life has taken us down many different paths and to many different places. Sometimes we reunite at home, sometimes on the other side of the world. Yet, it always feels like we have just met a week ago. You know you will always have a home, a safe space, wherever I am.

To my friends from university: Thank you for being yourselves and for allowing me to be myself. Thank you for sharing sports, beers, meals, concerts, seminars, and even for sharing your roots with me. Bonding over the same experiences – happiness, sadness, frustration, and hope (sometimes all in the same week) – has forged a connection I will carry with me forever.

To all the other amazing people I have met in Barcelona: Whether by chance or by fate, our paths crossed in ways that have enriched each other's lives. Thank you for all the shared moments and for helping me call this city "home".

To the academic community, thank you for creating such a stimulating environment for the development of my ideas. To the UB School of Economics, and especially to the Riskcenter, I am grateful for your ongoing support and for the knowledge we have built together over these years. Special thanks to the UBSE

programme management team, who have been our constant point of reference at every step of the way. This journey would not have been possible without you.

To the CoDA association: You are a small, strong community that has welcomed my work and made me feel valued over the years. Thank you for letting me be part of the group.

Lastly, but certainly not least, I would like to thank the external evaluators and the thesis committee. I deeply appreciate your time, effort, and thoroughness in reviewing and assessing my thesis. I hope the final outcome meets your expectations, and I am grateful for the feedback provided.

Abstract

For many multivariate problems in finance and economics relative values of the variables are more relevant than their absolute values, which is the basis of compositional data analysis. This thesis aims at contributing to the integration of the compositional framework into a relevant domain of financial analysis, namely, financial stability. The thesis is organized into three distinct parts, each focusing on a specific financial analysis.

The first one relates to the Feldstein-Horioka (F-H) puzzle, which states that liberalization of capital markets does not necessarily lead to a movement of capital looking for a better allocation of resources, as classical theory suggests. In recent years, Chile, Colombia, Mexico and Peru joined the Latin American Integrated Market through an agreement that allows investors in any of the participating markets to invest in the others. Both cross sectional and time series compositional methods were used to assess whether the creation of the joint market led to a flow of capital between markets. As a result, it was found that it is not possible to reject the F-H hypothesis, supporting the idea that the liberalization of capital markets is not enough to generate capital flows between markets.

Secondly, a concentration index for financial/banking systems via compositional methods is constructed to establish the potential existence of "too big to fail" financial entities and provide regulators with an early warning tool for this type of institution. The index was applied to the Colombian banking system and monitored over time to assess whether the financial system was becoming more concentrated. Results found that the concentration index was decreasing, and this trend would continue. From the methodological point of view, compositional models showed to be more stable and to lead to better prediction compared to classical methodologies.

Finally, the relationships between assets in a portfolio are evaluated from a compositional perspective. For many years, the issue of spurious correlations among variables expressed in relative terms, such as composition data, has received a lot of interest. As an alternative to the correlation, this thesis proposes a proportionality index for parts of a composition based on the log-ratio variance, a measure widely used when analyzing proportionality. The index was applied to a hypothetical portfolio composed of stocks from the Spanish stock market to assess the connections between the allocations generated by the Mean-Variance portfolio method. The index shed light on how the allocation process assigns the optimal allocations.

JEL classification: C01, E22, F32, G11, G17, G21, G28.
Keywords: Aitchison geometry, Feldstein-Horioka puzzle, banking concentration, systemically important banks,Mean-Variance portfolio, proportionality.

Contents

List of Figures

List of Tables

1 Introduction

Multivariate analysis focuses on datasets containing multiple measurements of each experimental unit (Olkin and Sampson, 2001). The methods used in multivariate analysis are numerous, going from covariance analysis and regression to copulas, principal component analysis or machine learning. Depending on whether the observations are cross sectional or time series, the tools employed will change. In the first case the measurements are taken over different agents, usually at one point in time, while the second are observations over the same agent over time.

Compositional data are a type of multivariate data in which the absolute values are irrelevant whilst the participation of each variable in the total is the key factor, said, the relative importance of one variable with respect to the others (Pawlowsky-Glahn et al., 2011). A composition is a vector of non-negative elements with the restriction of adding up to a constant, usually the unit. Expressing data in terms of compositions allows for the use of a different framework, which is consistent with the nature of such series. Compositional techniques have extensively been applied in fields like geology (e.g. minerals that form the soil) or sociology (e.g. socio-demographic groups within a population, depending on characteristics such as religion, race, age, etc.). Important advances in the compositional techniques have been done by researchers in catalan universities like Juan José Egozcue Rubí (Research group in Compositional and Spatial Data Analysis at UPC) or Dr. Vera Pawlowsky-Glahn and Dr. Josep Antoni Martín Fernández (Research Group Statistics and Compositional Data Analysis at UdG), for example.

In economics and finance the use of compositional techniques is still incipient. Nevertheless, there is a wide spectrum of economic variables which can be analyzed by means of compositional data. For instance, gross domestic product can be expressed in terms of the participations of each sector to the total growth or in terms of the participation of each region in the overall production. Expanding this analysis to other economic variables may lead to models with a better understanding of the dynamics of such figures. Yet, the application of compositional techniques has gained some importance. For example, Glassman and Riddick (1996) assess the performance of international portfolios using logarithmic transformations. Belles-Sampera et al. (2016) were the first to show the connection between capital allocation and compositional data, through a descriptive analysis of various solutions to

capital allocation problems. More recently, Verbelen et al. (2018) used compositional models to establish car insurance tariffs and Boonen et al. (2019) proposed time series models to study the evolution over time of capital allocations in a set of stock indices, while Fiori and Porro (2023) and Fiori and Rosazza Gianin (2025) deepen into the compositional nature of capital allocations. However, a field that has seen an important increase in the application of compositional techniques is the analysis of financial statements, including the contributions of Linares-Mustarós et al. (2018), Carreras-Simó and Coenders (2021), Arimany Serrat et al. (2022), Linares-Mustarós et al. (2022), Arimany-Serrat et al. (2023), Molas-Colomer et al. (2024), Hernandez-Romero and Coenders (2025), Coenders and Arimany Serrat (2025), and Magrini (2025).

This doctoral thesis aims at contributing to the analysis of multivariate problems in the fields of financial markets and financial stability by introducing a novel compositional approach, which emphasizes the relative nature of information (compositions) across multiple financial contexts. Indeed, there is no doubt that a proper understanding of economic crises is essential, not only to find ways to avoid them in the future, but also to be better prepared when the next one comes, given that economic cycles are inherent to the system and so are the bearish periods. The aim is to show how the compositional data analysis provides a valuable set of tools that have been under exploited for this analysis, focusing on three financial applications. First, since Feldstein and Horioka (1980) found evidence on what later was called the "home bias" of investment (a preference of investors to invest locally, even with perfect mobility of capital across countries), much effort has been put on finding the reasons behind this behavior. One of the main explanations is that risks (either exchange rate risk, volatility of capital flows, or regulatory risk, among others) play a key role on these investment decisions (Horioka et al., 2016). In this case, the compositional approach is used to provide evidence on capital flows in integrated markets (specifically the MILA market). Second, the recent global financial crisis highlighted the importance of understanding the role of large players in the financial system, meaning those with high participation in the market. This already implies the suitability of compositional methods in such analysis and here a tool to grasp concentration in the market is developed from this perspective. Finally, the need for regulation that helps reduce the impact of crises on economic agents is the base of capital and asset allocation analysis, and recent literature has shown that this is a compositional problem by construction. Here, this base is used to analyze the relationships between allocations in a portfolio.

To develop this analysis, the thesis contains a methodological chapter providing an insight of the building blocks that will support the study of the three cases mentioned and that are investigated in detail in one chapter each. Finally, the main

findings are summarized in the concluding remarks.

Chapter 2 is dedicated to explaining the methodological framework supporting the upcoming research. This chapter starts with a formal definition of compositional data and the Aitchison geometry, which is the departing point of compositional analysis. This is followed by the introduction of concepts such as subcompositional coherence and logarithmic transformations, which are key for understanding the approach used afterwards. Finally, traditional time series and its compositional approach are explained. With this toolkit, the three multivariate problems in the fields of financial markets and financial stability are analyzed.

Chapter 3 analyzes the Feldstein-Horioka puzzle from a compositional perspective. Feldstein and Horioka (1980) found through an analysis of the investment-saving relationship that the liberalization of capital markets is not enough to generate net transfers of financial capital between countries and that the integration of goods markets is also necessary to compensate the transfers of capital (Ford and Horioka, 2017). During the recent years, many authors have tried to find evidence and explanations in favor and against this theory. For instance, Narayan (2005) finds that the Feldstein-Horioka puzzle is verified in China between the 1950s and the 1990s, while Fouquau et al. (2008) conclude through a transition regression model that openness, country size and current account size have an influence in the relationship between investment and saving. On the other hand, authors like Coakley et al. (1996) and Drakos et al. (2017) try to find different explanations for the correlation between investment and savings that debate Feldstein and Horioka's argument. For the case of Latin America, Bellod-Redondo (1996) analyzes the capital flows in Mexico, before and after the integration to the North Atlantic Free Trade Agreement (NAFTA), concluding that saving rates limit the investment. On the other hand, Ibarra-Yunez (2008) tried to verify the Feldstein–Horioka puzzle in the region finding that, despite the opening in the economies during the recent years, there is no evidence of capital movements. Under this scenario, compositional data is used to analyze the Latin American Integrated Market (MILA) which is an agreement between the stock markets of Chile, Colombia, Mexico and Peru, that created a common stock market in which all participants should be able to invest in stocks from the four markets as if they were investing in their local markets. The analysis will intend to determine whether the composition of the investment across the four markets is different after the entry into force of the agreement, implying that there was a movement of capitals from one country to another. As can be seen, the analysis focuses on a specific type of capital flow, contrary to the approach of Feldstein and Horioka, which uses the aggregated data of the capital account of the balance of payments. Similar analyses of capital movements have been conducted by Adedeji and Thornton (2007) and Vieira (2003), not concentrating on the Feldstein–Horioka

puzzle itself, but rather on the flows of capital.

In this case, compositional methods are used to analyze the behavior of the participation of each market before and after the entry into force of the interconnected market. The idea is to determine whether there has been a change in the composition of the integrated market after investors were given the opportunity to trade in all the markets. The time frame selected (2010 to 2020) contains observations within the three periods: before June 2011 (when Chile, Colombia and Peru founded the joint market), from this moment until December 2014 (when Mexico joined) and from December 2014 onwards, with the current configuration of the interconnected market in operation. Both cross sectional and time series compositional methods are applied to find if there was a change in the composition of the investment in the four markets produced by the creation of the joint market and, thus, if there was evidence to reject (or not) the Felstein-Horioka hypothesis, supporting the idea that liberalization is not enough to generate capital flows between markets.

In Chapter 4 compositional data is applied to create an indicator of concentration for the Colombian banking system. This kind of analysis has been of relevance during the last decades, considering the potential exposures of highly concentrated systems to (negative) shocks in one of the main participants. Indeed, these participants have received the name of "too big to fail" (TBTF) institutions and the discussion about them dates back to the 1980s. However, there was still no evidence, nor consensus, about the actual impact of these institutions to the overall risk of the market. For instance, Mishkin et al. (2006) summarize the debate in literature regarding this topic. It was during the financial crisis in 2008 that the attention was turned towards these institutions, because of the large amounts of public funds invested by governments to prevent them from failing. The main critique to this response from governments and regulators comes from the moral hazard implied: as these banks know about their systemic importance, they have incentives to act in a risk lover fashion, expecting governments to come in their rescue in case of materialization of risk. The analysis during the recent years has been focused on whether this kind of institutions should reduce their size and, therefore, their systemic importance (together with the risk exposure of the system to their behavior), or how to regulate them to reduce such risk.

Undeniably, literature on TBTF entities started growing in the following years, including analyses like Sorkin (2010), Shull (2010) and Zhou (2010). Later on, the definition of TBTF institutions expanded after the Basel Committee on Banking Supervision (2013) defined systemically important banks (SIBs) also in terms of interconnectedness, substitutability, cross-national activity and complexity. On the other hand, a revision of literature by Moch (2018) concluded that the size of banks has a nonlinear relationship with systemic risk. This means that systemic

risk increases more than proportionally when the size of banks increases. Additionally, it was found that interconnectedness and concentration sharpen this nonlinear effect. More recent approaches to the matter include the risk indicator proposed by Bezrodna et al. (2019), which aimed to connect the size of a bank in terms of assets to the riskiness of the institution, and the leave-one-out method used by Li et al. (2020), who intended to determine the effect of each single institution in the systemic risk. The amount of literature on TBTF institutions shows the relevance of the topic and the continuous search for new methodologies that can help keep track of them, as well as all potential tools that can be used to minimize the risk associated with them.

Thus, considering each financial institution as a part of the system, compositional methods are applied to elaborate a concentration index for the financial system. This methodology is applied to the Colombian banking system, assessing its trend through the recent years. Furthermore, the model is used for forecasting the behavior of the index in the following years. Summarizing, the concentration index proposed is expected to be considered as an early warning, to alert the policy makers and regulators about the existence of potential TBTF entities in the system.

Chapter 5 assesses the relationships between allocations within a portfolio. Indeed, asset and capital allocations are compositional by nature. One of the most well-known asset allocation methods used in portfolio theory is the Mean-Variance portfolio (MVP) theory from Markowitz (1952), and provides the optimal participation of each asset within a portfolio based on their expected returns and volatility. This optimization process does not only consider the individual volatilities (measured through the variance), but also the covariances (thus correlations) between them. However, the existence of spurious correlations firstly noticed by Pearson (1897) and later analyzed by Chayes (1960) gave rise to the use of log-ratios to overcome the limitations of the correlation as a measure of relationship between variables. For instance, Egozcue and Pawlowsky-Glahn (2023) show through a simple example how correlations of parts of compositions represented in closed form (further explained in section 2.1) are inconsistent depending on which and how many parts conform the composition. Hence Lovell et al. (2015) and later Egozcue and Pawlowsky-Glahn (2023) used the log-ratio variance defined by Aitchison (1986) as the building block of proportionality analysis and propose an index that suits better the properties of compositions than correlations.

Further developing on proportionalities, this chapter elaborates on the statistical properties of the log-ratio variance to propose a new proportionality index that allows to assess the relationships between parts of a composition. This methodology is used to assess the relationships between allocations in a hypothetical portfolio formed with the top five stocks in the IBEX35 (the reference index of the Spamish

stock market). The allocations are obtained by using the minimum variance method (a particular case of the MVP mentioned earlier), hence the need for using compositional methods. The methodology sheds light on how the proportionalities between the allocations can be used to understand the way allocations are assigned by the Mean-Variance method.

In Chapter 6 concluding remarks are summarized. Overall, the use of compositional methods in the three problems analyzed show the suitability of such methods in financial stability analysis. Additionally, this chapter offers an introduction to new paths of research, by employing the tools from compositional data analysis to answer relevant questions in economics and finance.

2 Compositional methods

This chapter is dedicated to explaining the methodological background that is common to the three upcoming chapters. As mentioned before, the backbone of the thesis is compositional data analysis, a full set of methodologies to study vectors which components are part of a whole.

Hence, Section 2.1 introduces the Aitchison geometry, which is the basis of the compositional data analysis and Section 2.2 elaborates on the concept of subcompositional coherence. Finally, Section 2.3 explains time series analysis focusing on its implementation for compositional data.

2.1 Aitchison geometry

Formally, a composition is a vector $X = [x_1, \ldots, x_n]$ of size $n > 1$, whose elements carry only relative information, such that $\sum_{j=1}^{n} x_j = \kappa$ (Pawlowsky-Glahn et al., 2011). Therefore, it is defined in the simplex:

$$\mathcal{S}^n = \{X \in \mathbb{R}_+^n | x_i \geq 0, i = 1, \ldots, n, \sum_{i=1}^{n} x_i = \kappa\} \tag{2.1}$$

For simplicity, the value of κ is set to 1, without loss of generality. The characteristics of $\mathcal{S}^n$ impose several restrictions at the moment of modeling the series. Despite being an issue for several years, it was Aitchison (1982, 1983, 1984, 1986) who defined desirable features of compositional methods and what is referred to as the Aitchison geometry. This includes, among other methods, transformations to express compositional series (defined in the sample space $\mathcal{S}^n$) in $\mathbb{R}^n$ and then being able to apply conventional statistical techniques to the transformed series.

As mentioned, the value of κ can be set to a desired value to improve interpretability of the compositions: 1, 100, 1000, 1 million, etc. For the sake of this thesis, the value $\kappa = 1$ is chosen. This is done by using the closure operation:

$$\begin{aligned} \mathcal{C}(X) &= \left[\frac{\kappa x_1}{\sum_{i=1}^{n} x_i}, \frac{\kappa x_2}{\sum_{i=1}^{n} x_i}, \ldots, \frac{\kappa x_n}{\sum_{i=1}^{n} x_i}\right] \in \mathcal{S}^n \\ \mathcal{C}(X) &= \left[\frac{x_1}{\sum_{i=1}^{n} x_i}, \frac{x_2}{\sum_{i=1}^{n} x_i}, \ldots, \frac{x_n}{\sum_{i=1}^{n} x_i}\right] \in \mathcal{S}^n \end{aligned} \tag{2.2}$$

Furthermore, Aitchison (1986) defined other elements of the algebraic-geometric structure of the simplex. For instance, the perturbation operation $\oplus$ within the simplex is defined as $X \oplus Y = \left(\frac{x_1 \cdot y_1}{\sum_{i=1}^{n} x_i \cdot y_i}, \ldots, \frac{x_n \cdot y_n}{\sum_{i=1}^{n} x_i \cdot y_i}\right)$ for $X, Y \in \mathcal{S}^n$. Moreover, the powering operation $\odot$ is defined as $\lambda \odot X = \left(\frac{x_1^{\lambda}}{\sum_{i=1}^{n} x_i^{\lambda}}, \ldots, \frac{x_n^{\lambda}}{\sum_{i=1}^{n} x_i^{\lambda}}\right)$ for $\lambda \in \mathbb{R}$.

With this in mind, Aitchison (1986) also expressed the compositional mean for M compositions as $AM_{\Delta}(X_1, \ldots, X_M) = \frac{1}{M} \odot \bigoplus_{m=1}^{M} X_m$. Additionally, according to Aitchison geometry the distance between two compositions X and Y is defined by Equation 2.3:

$$AD_{\Delta}(X,Y) = \| X \ominus Y \|_{\Delta} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\right)^2} \tag{2.3}$$

where $\| \cdot \|_{\Delta}$ is the norm and $X \ominus Y = X \oplus [(-1) \odot Y]$. This definition brings to light an important drawback of the Aitchison geometry: logarithmic transformations do not allow for zeros in the compositions. The usual approach when dealing with zeros in compositions is to replace them with sufficiently small values (Martín-Fernández and Thió-Henestrosa, 2013). Nevertheless, the issue of zeros remains a point of debate, following the contributions of Simone (2014), Templ et al. (2011) and Bouzid and Kervrann (2019).

Despite the advantages of using the compositional data framework to analyze multivariate series, the sum of the vector X being equal to a constant is a constraint. This limitation causes issues when the usual multivariate econometric models are applied. To overcome this issue, the elements of $X \in \mathcal{S}^n$ need to be translated into elements of $\mathbb{R}$. Aitchison (1986) suggested a first attempt by defining the centered log-ratio (clr) transformation:

$$clr(X) = \left[\ln \frac{x_1}{g(X)}, \ldots, \ln \frac{x_i}{g(X)}\right] \text{ with } g(X) = \left(\prod_{i=1}^{n} x_i\right)^{1/n} \tag{2.4}$$

where $g(X)$ is the component-wise geometric mean of the composition. As can be seen, because the clr is a component-wise transformation, it follows that $clr(X) \in \mathbb{R}^n$. However, by definition, the elements of $clr(X)$ add up to zero, which

imposes a new restriction.[1]

To apply the usual econometric models, X must be translated into the Euclidean space without restrictions. This can be achieved through the isometric log-ratio (ilr) transformation (Pawlowsky-Glahn et al., 2011). Starting from an orthonormal basis $\mathbf{e} = \{e_1, ..., e_{n-1}\}$, it is possible to create the matrix V, whose rows are defined as $clr(e_j)$. Therefore, V is $n-1 \times n$. Given its construction, V satisfies: $V \cdot V' = \mathbf{I}_{n-1}$ and $V' \cdot V = \mathbf{I}_n - (1/n)\mathbf{1}'_n\mathbf{1}_n$ where $\mathbf{I}_n$ is the identity matrix of size n and $\mathbf{1}_n$ is the n-row vector of ones. With V, the ilr transformation can be defined as:

$$ilr(X) = clr(X) \cdot V \tag{2.5}$$

Another advantage of the ilr transformation is that $\Delta(X,Y) = d(ilr(X), ilr(Y))$, where $d(\cdot,\cdot)$ is the Euclidean distance. More details can be found in Aitchison (1986) and Egozcue et al. (2003).

Now, the issue relies on defining the orthonormal basis $\mathbf{e}$. The most common method used in compositional data literature is what Egozcue et al. (2003) proposed through binary partitions. This method has the advantage of interpretability because the partitions can be such that the groups can be translated into a principal component analysis. In this process, the parts of the composition are divided into two groups. The elements of one of the groups will be assigned $+1$ while those of the other will be given -1. This way, the elements of one group balance with those of the other. In the subsequent steps, one of the groups is taken and divided again into two subgroups, coded $+1$ and -1, while the elements of the other group(s) are assigned 0. Therefore, the balances between two groups are created at each step. This process has to be completed $n-1$ times, until each subgroup contains only one element. The result is a matrix $n-1 \times n$ filled with $+1$, -1 and 0, which can be used as the orthonormal basis for the V matrix and the ilr transformation. After the transformation, the components of X are now represented by coordinates in $\mathbb{R}^{n-1}$, which allows using conventional statistical methods for multivariate analysis.

2.2 Subcompositional coherence

In the same line, a subcomposition is defined as a subset of components from a composition Aitchison (1982). Thus, for a subset of components $m \subseteq [1, \ldots, n]$ the subcomposition is[2]:

[1] This implies that the covariance matrix of $clr(X)$ is singular, that is, the determinant is zero (Pawlowsky-Glahn et al., 2011).

[2] Note that they do not necessarily correspond to the first m parts of the composition.

$$X_m = [x_j]_{j \in m}, \quad \text{for } m \subseteq [1, \ldots, n] \tag{2.6}$$

This implies that the size of X_m must be larger than one (by the definition of the composition X) and smaller or equal to n. Then, by applying the closure operation:

$$\mathcal{C}(X_m) = \left[\frac{x_i}{\sum_{j \in m} x_j} \right]_{i \in m} \tag{2.7}$$

This means that the subcomposition can be seen as a composition itself and can be analyzed independently.

Considering the nature of compositional data, Aitchison (1983, 1984, 1986) defined desirable features of compositional methods. One of these characteristics is subcompositional coherence, which implies that the conclusions obtained for elements in a subcomposition should be equal to those obtained in the full composition.

More formally (Egozcue and Pawlowsky-Glahn, 2023), a function $f_n : \mathcal{S}^n \to \mathbb{R}$ is subcompositionally coherent if it is:

1. Scale invariant: For a positive real constant α, if it holds that $f_n(\alpha \cdot X) = f_n(X)$, then it is scale invariant. This, translated into compositional terms means that for $f_m : \mathcal{S}^m \to \mathbb{R}$ containing only the parameters of the elements in the subset m, it should hold that $f_n(X) = f_m(X_m)$ for any subcomposition X_m. This is referred to as an invariant function under subcomposition.

2. Subcompositionally dominant: If it holds that either $f_n(X) \geq f_m(X_m)$ (non-increasing dominance) or $f_n(X) \leq f_m(X_m)$ (non-decreasing dominance), then f_n is subcompositionally dominant with respect to f_m under the subcomposition X_m.[3]

To acknowledge this, Aitchison (1986) proposed the logratio variance to measure the association between components. Thus, for two components x_i and x_j, the logratio variance is defined as:

$$\tau_{ij} = Var\left(log(x_i/x_j)\right) \tag{2.8}$$

This measure has desirable characteristics: in particular, it fulfills subcompositional coherence, since the ratios x_i/x_j are independent of the other components.

[3] A full, detailed definition of subcompositional coherence can be found in Section 3 of Egozcue and Pawlowsky-Glahn (2023)

2.3 Compositional time series models

2.3.1 Time series analysis

Following Fuller (1996), a time series is a real function $z(t,\omega)$ for $t \in T$ (time) and $\omega \in \Omega$ (all possible realizations of the variable). Therefore, z is defined on $T \times \Omega$. For a fixed t, x is a random variable on a probability space. On the other hand, for a fixed ω, z is a function of time called a realization or sample function, and it is what can be actually observed in practice. Therefore, each one of the observations in the time series is a random variable itself and has its own distribution. When looking at the distribution of the observations over time (the sample function), it was found to be a joint distribution function of all individual random variables.

With this in mind, a time series is told to be strictly stationary if the joint distribution is the same, independent of the time t. Nevertheless, as mentioned before, in reality, it is only possible to observe one realization of the time series, which makes it impossible to obtain the joint distribution function. Therefore, it is common to consider the weak stationarity of time series by examining only the first two moments of the distribution. Indeed, a time series is said to be weakly stationary if the expected value of z is constant for all t and the covariance matrix is only a function of the distance between the realizations and does not depend on t itself.

In practice, the stationarity of a time series is tested using unit roots tests. The augmented Dickey-Fuller test (Said and Dickey, 1984) assesses the null hypothesis of unit roots in the characteristic equation of the time series, against the alternative of a stationary series. The idea behind this is that all roots of the characteristic equation of a stationary series should lie inside the unit circle.

For stationary time series, it is possible to find a stochastic difference equation of the form:

$$\sum_{i=0}^{p} \alpha_i z_{t-i} = \xi_t \text{ or the equivalent } z_t = \sum_{i=1}^{p} \beta_i z_{t-i} + \xi'_t \tag{2.9}$$

where α_i is a scalar coefficient with $\alpha_0 \neq 0$ and $\alpha_p \neq 0$, p is the order of the autoregressive time series, ξ follows a normal distribution with mean zero and variance σ_ξ, $N(0,\sigma_\xi)$, β_i is a coefficient in terms of α and ξ' follows $N(0,\sigma_{\xi'})$. The latter equation is known as the generic form of an autoregressive process.

In practice, to define the number of lags p to be used in the model, the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are used. The AIC is a likelihood criterion for measuring the fitness of a statistical model to a sample. It is defined as $AIC = -2\ell + 2K$ where ℓ is the log-likelihood of the estimation and K corresponds to the number of parameters in the model. Similarly,

the BIC is defined as $BIC = -2\ln(\ell) + K\ln(n)$, where n is the sample size. As can be seen, in both cases the higher the goodness-of-fit with respect to the number of parameters used, the lower the values of the information criteria. Therefore, a lower value of the criteria will indicate a better model.

The vector autoregressive (VAR) model can be seen as an extension of the autoregressive model defined previously for which the variable z is replaced for a vector $Z = [z_1, \ldots, z_n]$, usually referred to as a multivariate time series. With this in mind, the general form of the VAR model is:

$$Z_t = \sum_{i=1}^{p} B_i Z_{t-i} + \epsilon_t \tag{2.10}$$

where B_i is a matrix of parameters in $\mathbb{R}^{n \times n}$ and ϵ_t is the error term with multivariate normal distribution with zero means and a covariance matrix Σ_ϵ, $N_n(0, \Sigma_\epsilon)$. The model in (2.10) can be extended by adding a vector $A_t = [a_1, \ldots, a_l]$ of control variables:

$$Z_t = \sum_{i=1}^{p} B_i Z_{t-i} + \gamma A_t + \epsilon_t \tag{2.11}$$

where γ will be the $l - 1 \times 1$ vector of coefficients associated with the control variable.

However, in real life, many observed series are not stationary. Even though the VAR model can be estimated in the absence of stationarity because the estimators exist and are consistent (Sims et al., 1990), this is one of the main assumptions to determine the estimators' distribution and be able to make an inference.

If there is no stationarity, classical theory usually proposes two ways of modeling the time series: a VAR model in differences and a vector error correction (VEC) model (Lütkepohl, 2007). In the first case, the non-stationary series are differenced to obtain stationarity. Then, a VAR model is estimated with the new variables. This approach has a limitation regarding interpretability because the new coefficients will not refer to the effect of one variable on the other, but to the effect of changes in one variable on changes in the other one. The VAR in differences model is defined as:

$$\Delta Z_t = \sum_{i=1}^{p} B_i \Delta Z_{t-i} + \epsilon_t \tag{2.12}$$

where Δ is the difference operator defined as $\Delta Z_t = Z_t - Z_{t-1}$.

A second approach relies on the fact that even if the series are not stationary, they can maintain a long-term relationship that may be stationary. Two non-stationary time series are cointegrated if it is possible to find a linear combination of them that

forms a stationary series. Therefore, relying on this assumption, the VEC estimation intends to model this long-term relationship and add it to the VAR model. Thus, the VEC model is defined as:

$$\Delta Z_t = \eta + H \cdot Z_{t-1} + \sum_{i=1}^{p-1} B_i^* \Delta Z_{t-i} + \epsilon_t \tag{2.13}$$

where η is a vector of parameters in $\mathbb{R}^n$, $H = \sum_{s=1}^{p} B_s - I$ and $B_r^* = -\sum_{s=r+1}^{p} B_s$ (B_s corresponds to the matrix of the VAR model's coefficients in Equation 2.10, meaning both H and B_r^* are of size $n \times n$). Furthermore, $H = \alpha \cdot \beta'$ can be defined, where α will denote the size of the cointegration effects and β the cointegration matrix leading to the stationarity of the series. The dimensions of both α and β will be determined by the rank of β, which is explained below. Consequently, the VEC model can be interpreted as composed via the long-term dynamics (contained in H) and the short-term effects (modeled through B_r^*).

To estimate a VEC model, the variables' order of integration must be known, meaning how many times it is necessary to difference each variable to obtain a stationary series, and whether the variables are cointegrated, that is if a long-term relationship exists between the variables and how many variables are needed to obtain it, which would correspond to the rank of the cointegration matrix β.

2.3.2 Compositional time series

The time series approach can be extended to compositional data. In this case, after applying the ilr transformation, it is possible to use conventional statistical models and estimate a VAR model. Therefore, the generic VAR model for a composition X (after applying the ilr transformation) is defined as:

$$ilr(X_t) = \sum_{i=1}^{p} \varrho_i ilr(X_{t-i}) + \varepsilon_t \tag{2.14}$$

And the extended form would be:

$$ilr(X_t) = \sum_{i=1}^{p} \varrho_i ilr(X_{t-i}) + \gamma A_t + \varepsilon_t \tag{2.15}$$

Similar to the previous case, if the transformed series is not stationary, then it is possible to differentiate the series by applying the Δ operator, and, if the series in differences is stationary, then the model to estimate would be:

$$\Delta ilr(X_t) = \sum_{i=1}^{p} \varrho_i \Delta ilr(X_{t-i}) + \varepsilon_t \tag{2.16}$$

Finally, for the VEC compositional model it would be:

$$\Delta ilr(X_t) = \mu + M \cdot ilr(X_{t-1}) + \sum_{i=1}^{p-1} \varrho_i^* \Delta ilr(X_{t-i}) + \epsilon_t \tag{2.17}$$

Now that the general methodology for the upcoming chapters has been explained, it is time to apply these concepts in different issues in economics and finance. Nevertheless, each chapter contains an explanation on which parts and how these tools are used to contribute to each particular matter and generate results and conclusions.

3 Capital flows in integrated capital markets: MILA case[4]

The aim of this chapter is to analyze if there are movements of capital across markets when they become fully integrated. The Feldstein-Horioka (F-H) puzzle states that the liberalization of capital markets is not enough to generate net transfers of capital between countries (Feldstein and Horioka, 1980). This study analyzes investment flows in the Latin American Integrated Market (MILA), which is an agreement between four markets (Chile, Colombia, Mexico and Peru), that allows perfect mobility between them. The participation of each stock exchange in the joint market is examined over time to assess whether the entrance in force of the agreement generated a recomposition of the MILA, debating the F-H hypothesis. Thus, this approach does not enquire into the overall investment of a country, but focuses on portfolio investment, which is expected to be less subject to restrictions and less dependent on other types of international transfers, such as those in goods and services markets.

Since the assessment focuses on the participation, compositional methods are a natural approach to the problem. These methods have been widely developed in the recent decades, but applications in economics remain rare[5]. Therefore, this is also an opportunity to employ novel methods and contribute to the debate around the F-H puzzle.

In this case, there are two options: first, data can be treated as cross-sectional, applying compositional models and measures of distance to assess the hypothesis; second, treat data as time series, using these models to search for a change in the dynamics of the market structure. Therefore, both perspectives are used to verify the F-H puzzle, and the results are evaluated to determine which provides better insight into the problem.

The paper is structured as follows. Section 3.1 explains the F-H puzzle and the

[4]This chapter is based on an article published on Quantitative Finance and Economics (Vega Baquero and Santolino, 2022a).

[5]See, for example, Belles-Sampera et al. (2016), who studied capital allocation problems using the compositional framework, or Boonen et al. (2019), who forecasted risk allocations through compositional data.

analysis that gives motivation for this study. Then, section 3.2 provides insight into the MILA market and its participating countries. Section 3.3 describes the methodology used to assess the F-H hypothesis, and Section 3.4 presents the dataset. Finally, Section 3.5 shows the main results of the evaluation of the F-H puzzle and Section 3.6 concludes the study.

3.1 The Feldstein-Horioka puzzle

The findings of Feldstein and Horioka (1980) marked the beginning of a discussion in economics: Does the liberalization of capital markets generate net transfers of financial capital between countries? Classical theory would say yes, but the authors concluded that the data do not agree. In particular, they used data on 21 Organisation for Economic Co-operation and Development (OECD) countries for the period 1960–1974 to assess the relationship between domestic saving and domestic investment. The idea behind is that with perfect mobility of capital, there should be no correlation between these two variables since the investment decisions would respond to the opportunities in the global market. Indeed, they found that *"international differences in domestic savings rates among major industrial countries have corresponded to almost equal differences in domestic investment rates"* (Feldstein and Horioka, 1980, p. 328). As explained later by Ford and Horioka (2017), the liberalization of capital markets is not enough to generate net transfers of financial capital between countries, so the integration of goods markets is also necessary to compensate the transfers of capital.

Following these findings, the literature on the matter has been prolific, trying to argue both,in favor and against the conclusions drawn. For instance, Narayan (2005) used a cointegration approach to verify whether saving and investment were correlated in China between the 1950s and the 1990s, concluding that, due to the fixed exchange rate regime operating during the largest part of the period, the Feldstein-Horioka (F-H) puzzle holds and the correlation between savings and investment is high. On the other hand, Fouquau et al. (2008) made use of panel threshold regression models to assess the impact of other variables on the relationship between savings and investment, finding that openness, country size and current account size have an influence.

Conversely, Coakley et al. (1996) tried to explain the correlation between savings and investment through the use of a theoretical model in which the correlation is due to the cointegration of the variables, as explained in the stationarity of the current account balance (since the current account is equal to savings minus investment). Furthermore, Drakos et al. (2017) analyzed the relationship between savings and

investment from long-term and short-term perspectives, seeking for an explanation of the high correlation between the variables; they found that this is consistent with the existence of solvency constraints in the long run.

More specifically, in Latin America, Bellod-Redondo (1996) analyzed the capital flows in Mexico before and after the integration into the North Atlantic Free Trade Agreement (NAFTA), concluding that saving rates limit the investment. Another enquiry on the topic for Latin American countries was conducted by Sinha and Sinha (1998), who used a cointegration test to determine whether this can be the actual explanation for the high correlation between savings and investment and found that, for some of the countries in the sample, there was a long-term relationship between the two variables, while, for others, there was not; this coincided with high macroeconomic instability. More recently, Ibarra-Yunez (2008) tried to verify the F–H puzzle in the region, finding that, despite the opening of the economies during the recent years, there is no evidence of capital movements.

Additionally, there is extensive research on spillovers and connections in financial markets. Some of the most recent approaches include those by Chen and Dong (202), Dong (2020) and Jia (2021).

3.2 MILA market

The MILA is an agreement signed by the stock exchanges from Chile, Colombia and Peru, which started operating as an interconnected market in 2011. By the end of 2014, Mexico joined the three founding members to form the current MILA. The four countries from the Pacific Alliance (a Latin American trading bloc) signed the agreement which allows investors from any of the participating markets to invest in stocks from any of the exchanges. By the end of 2020, the market operated with more than 700 listed companies and had a capitalization above USD 770 billion. Operationally, the instruments are kept in the four separated exchanges, but they are interconnected for investors to be able to trade in any of the markets. A short anecdotal review of the history of each one of the markets is presented below in order to understand the different backgrounds that led to the foundation of the current MILA.

3.2.1 Chile

The Bolsa de Comercio de Santiago was founded in 1893[6], working during several years in parallel with the Bolsa de Corredores de Valparaíso. During part of the 20th

[6]https://www.bolsadesantiago.com

century, the latter was the main stock market in Chile because of the important economic activity taking place in Valparaíso, the country's main port. Nevertheless, as the years passed, and due to several regulatory changes, the Bolsa de Comercio de Santiago became the most important stock market in Chile. Since 1908, investors in both exchanges had the possibility to trade stocks in both markets through Inter Exchange Operations. In 2005, the two exchanges, together with the Bolsa Electrónica (founded in 1989), signed agreements to allow investors to trade stocks in all markets. In 2018, the Bolsa de Corredores de Valparaíso ceased operations.

As part of the participation in international entities and the alliances with other stock markets around the world, the Bolsa de Comercio de Santiago joined, as a founding member, the Ibero-American Federation of Stock Exchanges and Securities Markets in 1973 and, in 1991, the World Federation of Exchanges. In 2000, the Mercado de Valores Extranjeros (Foreign Stock Exchange) was created as a platform for local investors to trade foreign stocks. Finally, in 2011, the Bolsa de Comercio de Santiago joined the MILA, creating the biggest stock market in the region.

3.2.2 Colombia

The Bolsa de Valores de Colombia[7] was born in 2001 after the fusion of the Bolsa de Bogota (created in 1928), the Bolsa de Medellin (created in 1961) and the Bolsa de Occidente (created in 1983). Since the creation of the Bolsa de Bogota, the trading of stocks in the country has been closely linked to the performance of the business activity. Indeed, following the economic instability after the Great Depression, the first years of the trading activity were slow and the strongly restrictive legal frame during the following decades did not allow for important growth of the trading activity. It was not until the creation of the Bolsa de Medellin that the stock markets started gaining some dynamism, paired with the important growth in the economic activity during the 1960s. This led to the creation of the Bolsa de Occidente a few years later.

The technological improvements achieved in the 1990s helped the three markets experience an increase in their activity. Starting in the 2000s, the aim of the three markets to attract more international investment led to their fusion and the creation of the current Bolsa de Valores de Colombia.

[7] https://www.bvc.com.co

3.2.3 Mexico

After several years of informal trading, in 1894, the Bolsa Nacional was born in Mexico City[8]. A few months later, another stock exchange, the Bolsa de Mexico, started operating in the same city. Later, both exchanges found that it was more beneficial to work together and decided to merge under the name Bolsa de Mexico. In 1933, with the new legislation about the stock exchanges, the Bolsa de Valores de México was created as the sole exchange in the country. In the 1950s, the Bolsa de Monterrey and the Bolsa de Occidente (in Guadalajara) were created to stimulate the economic activity in those regions. In 1975, the new legislation on capital markets in Mexico led to the merge of all stock exchanges into the new Bolsa Mexicana de Valores.

By the end of the 1980s, and through the 1990s, the Bolsa Mexicana de Valores gained considerable dynamism due to the electronic negotiation, the incursion of international stockbrokers and an important increase in the number of listed companies. Additionally, in 2003, it started the investment of local agents in international stocks, continuing the internationalization of the Mexican stock market. In 2014, the Bolsa Mexicana de Valores joined the Latin American Integrated Market (MILA), after several years of technical adjustments.

3.2.4 Peru

The Bolsa de Comercio de Lima started operating in 1861[9]. For several years, there were no shares from private companies traded in the stock exchange, but it was creating valuations for them. After the high inflation period in the 1870s, the stock market regained some dynamism. The name was changed to Bolsa Comercial de Lima in 1898, with an increased offer for investors and changes in the benefits for traders and other agents in the market. The beginning of the 20th century brought important growth in the trading activity of the Peruvian stock market, which was stopped by the uncertainty generated by the Great Depression and World War II. This led to large reforms and the creation of the new Bolsa de Comercio de Lima in 1951.

At the beginning of the 1960s, the hundredth anniversary of the Peruvian stock market was a transition period in which the main objective was to strengthen the market and increase capitalization by incorporating new firms. This growth period led to further reforms and the creation of the current Bolsa de Valores de Lima (BVL) in 1971. During the following years, the BVL experienced important dy-

[8] https://www.bmv.com.mx
[9] https://www.bvl.com.pe

namism, marked by the incorporation of new technologies, it then became one of the most profitable stock markets in the world by the beginning of the 1990s. In 1995, the BVL started operating electronically, and from 2002, the market information was available for all users online, increasing the transparency and efficiency of the market. In 2011, the BVL joined the MILA with Chile and Colombia, with the aim of increasing the investment possibilities for all agents. In recent years, more innovations have come to the market, such as the creation, in 2013, of the Mercado Alternativo de Valores (Alternative stocks market), which is for the negotiation of shares from smaller companies, the incorporation of new indices and a remarkable internationalization strategy.

3.2.5 Joint market

The four stock markets that form the MILA have important participation in the company shares, although private and public bonds, derivatives and currencies are also traded. Regarding the main sectors trading in the exchanges, it is worth noting that all of them have important participation in the financial services, energy and mining and food and beverage companies within the most traded stocks.

Additionally, the Bolsa de Comercio de Santiago has important companies in sectors such as retail, real state and transportation services, while the Bolsa de Valores de Colombia has important public services firms participating in the market. On the other hand, the Bolsa Mexicana de Valores combines public services and transportation enterprises in the main traded shares, and the Bolsa de Valores de Lima has retail and industrial companies among the most relevant in the exchange market.

Although each stock market continues to have its own indices, financial services firms like the S&P Dow Jones Indices[10] carry indices of the joint market, such as the S&P MILA Pacific Alliance Select, which includes the largest, most liquid companies of the MILA; or the S&P MILA Andean 40, which includes the 40 largest and most liquid stocks in Chile, Colombia and Peru.

3.3 Methodology

This section is dedicated to introducing the methodology used through the analysis. The compositional data framework explained in Chapter 2 is used as the building block to construct several approaches to assess the F-H puzzle in the proposed setting. First, the cross-sectional methods with compositional data are shown. Af-

[10] https://www.spglobal.com/spdji

ter, considering that the data corresponds to a multivariate time series, this framework is explained comparing the traditional and the compositional data approaches. Throughout the analysis, the focus is on assessing whether there was a change in the composition of the MILA market in June 2011 (when the markets of Chile, Colombia and Peru created the joint market) or in December 2014 (when Mexico joined the three founding members). The dataset (explained in more detail in Section 3.4) includes monthly observations from January 2005 to December 2020, meaning that there are observations before and after these breaking points. Thus, the composition treated in this chapter is $X = [x_1, x_2, x_3, x_4]$ where each x_i is the participation of one market in the total MILA market.

3.3.1 Cross-sectional approach using compositional data

The first approach is to analyze the compositional series as cross-sectional data. Therefore, it is possible to estimate a mixture model adjusted for compositional data as explained by Comas-Cufí et al. (2016). In this case, it is assumed that the dataset follows a linear combination of a finite set of K distributions and has a probability density function (pdf) of the form $\pi_1 f_1(\cdot;\theta_1) + \cdots + \pi_K f_K(\cdot;\theta_K)$, where $\theta_1, \cdots, \theta_K$ are the parameters of the pdfs $f_1, \cdots, f_K$, respectively, and $\pi_1, \cdots, \pi_K$ are positive numbers with $\sum_{k=1}^{K} \pi_k = 1$ and the weights of each individual pdf (McLachlan and Peel, 2004). Considering that compositional data are defined in the simplex $\mathcal{S}^n$, the Dirichlet distribution is commonly used for modeling. However, Comas-Cufí et al. (2016) explained that an orthonormal transformation can be used to transform the data so that they can be modeled using distributions defined in the real space. In this case, the ilr transformation defined in Equation 2.5 is applied, so a Gaussian distribution can be used and a Gaussian mixture model is estimated. Therefore, in this estimation, each one of the f_k distributions corresponds to a Gaussian distribution with the parameters $\theta_k = [\mu_k, \sigma_k]$ for $k = 1, \ldots, K$.

The idea behind this approach is that, assuming the data comes from a mixture of $K = 3$ Gaussian distributions (one for each period analyzed), the model should assign the observations to these distributions following the three periods under analysis. If the model assigns the observations to the three distributions in a different fashion, then there is no evidence of a change in the composition of the MILA market due to the entrance in force of the agreement and the F-H puzzle could not be rejected.

On the other hand, it is possible to use different measures of distance between compositions to check whether the entrance in force of the agreement produced a change in the composition of the joint market, meaning that there was a potential movement of capital that would contradict the F-H puzzle. For each one of the

three measures of distance, it is calculated for every period X_t with respect to the previous one X_{t-1} to see if there was an immediate shift in the composition of the MILA market that would indicate any potential movement of capital at the moment of implementation of the agreement. This was done by comparing the results for the breaking points of the MILA market (detailed in Section 3.4) with those of the rest of the observations to obtain an idea of a "normal" value of each distance measure. If the variation at the period of interest appears as "atypical", it could indicate a potential flow of capital induced by the entrance in force of the agreement.

The first of these measures is the Aitchison distance defined in Equation 2.3 for two compositions X_t and $X_{t-1} \in \mathcal{S}^n$:

$$AD_\Delta(X_t, X_{t-1}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\ln\frac{x_{i,t}}{x_{j,t}} - \ln\frac{x_{i,t-1}}{x_{j,t-1}}\right)^2} \tag{3.1}$$

Also, following Thomas and Lovell (2014), it is possible to use measures like the compositional Kullback-Leibler divergence, which measures the difference of two compositions and is defined by Martín-Fernández et al. (ND) as:

$$KL_\Delta(X_t, X_{t-1}) = \sum_{i=1}^{n} \ln\frac{x_{i,t}}{x_{i,t-1}} \tag{3.2}$$

Thomas and Lovell (2014) also proposed the cosine similarity, which ranges between –1 (completely opposite vectors) and 1 (completely proportional vectors), with 0 meaning completely orthogonal vectors. The cosine similarity is defined as

$$CS_\Delta(X_t, X_{t-1}) = \frac{\sum_{i=1}^{n}\frac{x_{i,t}}{x_{i,t-1}}}{\sqrt{\sum_{i=1}^{n} x_{i,t}^2}\sqrt{\sum_{i=1}^{n} x_{i,t-1}^2}} \tag{3.3}$$

3.3.2 Time series approach

Another way to assess the F-H puzzle is by estimating a VAR model as defined in Equation 2.10 to find relationships between the market capitalization of the participants of the MILA and see whether the market capitalization of one market or its past observations are able to explain the market capitalization in the others. Thus, in this case the variables $Z = [z_1, z_2, z_3, z_4]$ are the market capitalization of each exchange. However, if the augmented Dickey-Fuller test shows that the series are not stationary, then the model to estimate will be the one in Equation 2.12.

Also, the compositional models from Equations 2.14 and 2.16 are estimated for the composition X defined at the beginning of this chapter.

For the estimation, it is necessary to decide which models fit better to the data,

including the proper number of lags p for each model by using the AIC and BIC criteria. Then it is time to examine whether the F-H puzzle holds for this specific setting. In order to do this, the model will be first estimated for the whole sample (in this case, from January 2005 to December 2020, as explained in more detail in Section 3.4). Later, the model will be estimated for three different subsamples: before June 2011 (when the MILA market started operating with Chile, Colombia and Peru), from July 2011 to November 2014 and from December 2014 onward (when Mexico joined the MILA). Finally, the coefficients obtained in the four models will be compared in order to see if the entrance in force of the agreement produced a change in the participation of each market in total, meaning that there was a recomposition of the investment induced by the agreement. Thus, if the coefficients of the general model are different from those for the models in the subsamples, this would mean that the dynamics of the series changed and there was a recomposition of the market.

3.4 Dataset

The variable to be used is the monthly market capitalization of each stock market. The data are publicly available from the World Federation of Exchanges[11] and are expressed in US dollars as common currency. The dataset is composed by utilizing the market capitalization of the four markets from January 2005 to December 2020. Figure 3.1 shows the dataset in levels, while Figure 3.2 shows the composition formed by the four markets. In both figures, there are two white lines dividing the graph: the first one corresponds to June 2011, when Chile, Colombia and Peru joined the MILA, and the second one to December 2014, when Mexico joined the three founding members and formed the current MILA. Therefore, the analysis will be concentrated on finding differences in the market capitalization of the four markets before and after the entrance in force of the agreement.

The use of this dataset has an important implication: by using a common currency (to be able to compare the data from the four markets), the fluctuations in the exchange rates in each country are disregarded. Therefore, an assumption is required to proceed: the four countries are affected in the same sense by shocks in the currency exchange markets. This means that, if one country suffers devaluation, the others will also go through the same process. To assess this assumption, the exchange rate fluctuations of the four currencies were verified for the period of interest. When checking for their pairwise correlations, all of them were above 0.59, meaning that there is a strong correlation between the exchange rates of all coun-

[11] https://www.world-exchanges.org/.

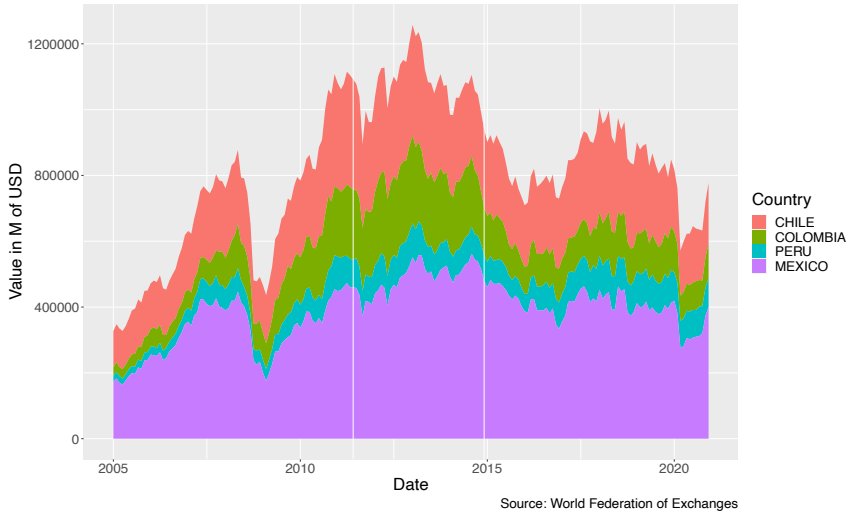


Figure 3.1: Market capitalization of MILA exchanges

tries against the US dollar, validating the assumption made. However, there is still room for a caveat regarding the potential effects of currency exchange fluctuations in the results.

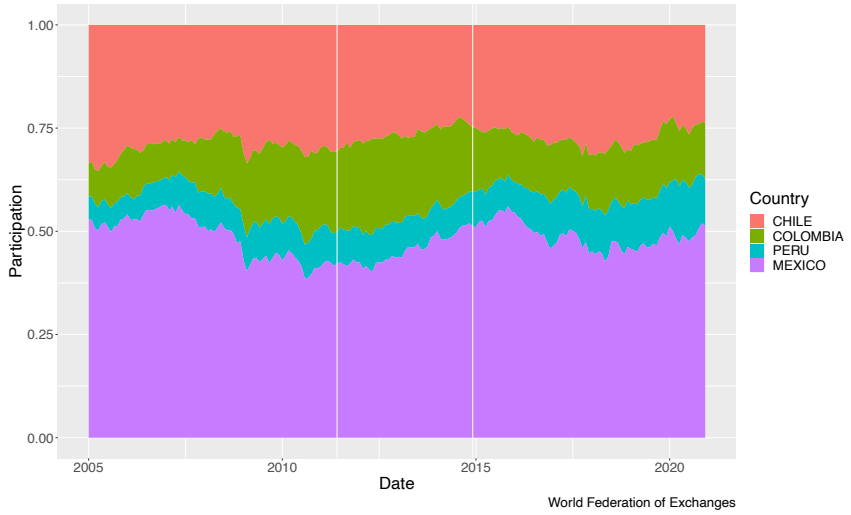


Figure 3.2: Composition of the market capitalization of MILA exchanges

3.5 Evaluation of F-H puzzle in MILA market

3.5.1 Cross-sectional analysis results

Figure 3.3 shows the results of the estimation of the Gaussian mixed model with three distributions. There seem to be two breaking points in the series, but they do not correspond to the entrance in force of the MILA market, which means that it is not possible to find a recomposition of the capital market caused by the agreement. Therefore, it is not possible to reject the F-H puzzle.

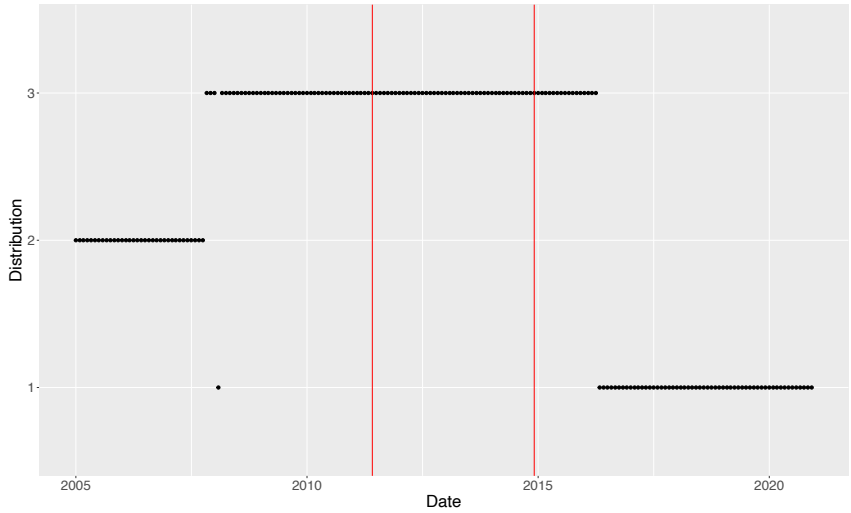


Figure 3.3: Gaussian mixed model results.

Regarding the results of the distance measures, Figure 3.4 shows that the distance between the composition of the MILA market each period, with respect to the period before, has been close to zero along the whole period under analysis, without any important variation around the moments of implementation of the MILA market.

Likewise, the cosine similarity estimation results are presented in Figure 3.5. It is possible to see that the values were close to one along the whole period, particularly in the moments of the integration of the markets, which means that there is no evidence of recomposition of the MILA market with the entrance in force of the agreement.

Finally, Figure 3.6 shows similar results for the compositional Kullback-Leibler divergence. Thus, across the period under analysis, the divergence between the composition of the MILA market for a month and the month before remained close to zero, particularly in the periods of interest when the joint market started operating.

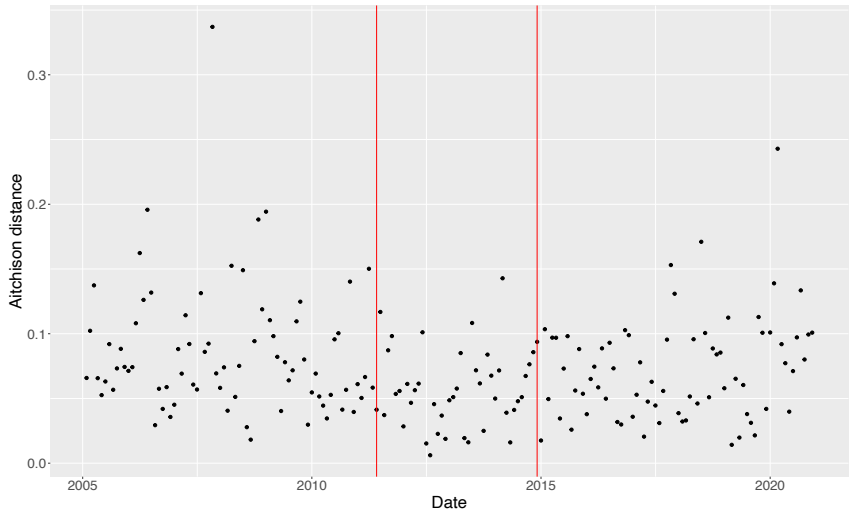


Figure 3.4: Aitchison distance results.

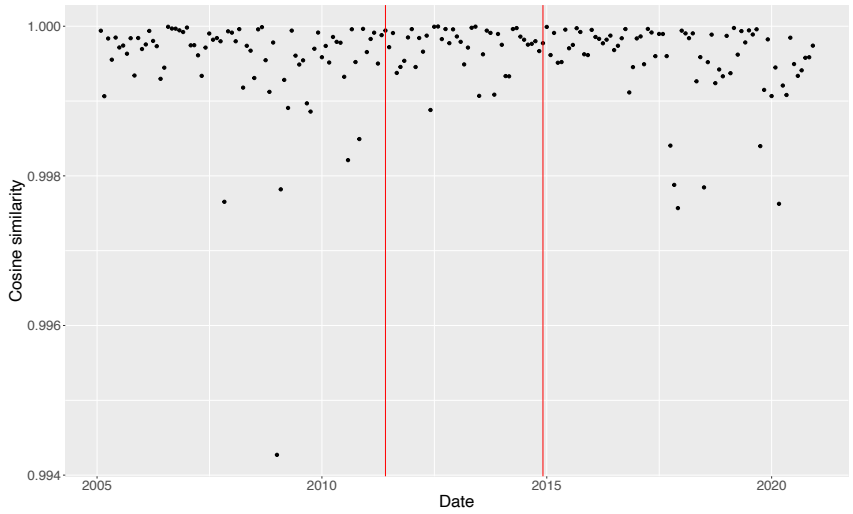


Figure 3.5: Cosine similarity results.

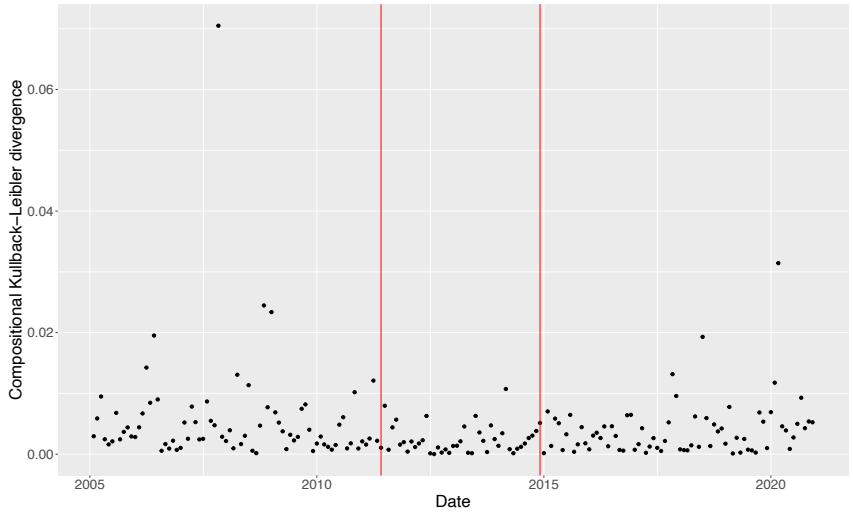


Figure 3.6: Compositional Kullback-Leibler divergence results.

As can be seen, the cross-sectional methods used agree with the F-H puzzle, as none of them have revealed any change in the composition of the market capitalization in the MILA with the entrance in force of the agreement.

3.5.2 Compositional time series results

Time series model selection

To define the (compositional) time series models to estimate, it is necessary to perform some diagnostic tests in the series. First, the stationarity of the series was tested since it is the first assumption when using time-series techniques. The augmented Dickey-Fuller test assesses the null hypothesis of a unit root in the characteristic equation of the series, which implies that the series is non-stationary. As shown in Table 3.1, none of the four series of market capitalization showed stationary in levels. After differentiating the series one time, the results of the test show that the series are stationary (Table 3.1), meaning that the series are integrated of order one and the model to be estimated is the one in Equation 2.12. In the case of the compositions, the situation is similar. Table 3.2 reports the results for the ilr transformed series. Recall that the size of $ilr(X)$ is 3 and there is no direct correspondence with the four countries that form the original composition because of the construction of the transformation. Therefore, the variables are renamed V.1, V.2 and V.3. The test results show that the levels are not stationary, unlike the first differences; so, the model in Equation 2.16 is the one to estimate.

	Levels		First differences	
	Statistic	P-value	Statistic	P-value
Chile	-1.90	0.62	-4.69	0.01
Colombia	-1.50	0.79	-5.43	0.01
Peru	-2.69	0.29	-5.16	0.01
Mexico	-2.63	0.31	-5.31	0.01

Table 3.1: Augmented Dickey-Fuller test for market capitalization series

	Levels		First differences	
	Statistic	P-value	Statistic	P-value
V.1	-1.81	0.66	-5.72	0.01
V.2	-1.70	0.70	-5.68	0.01
V.3	-2.24	0.48	-5.60	0.01

Table 3.2: Augmented Dickey-Fuller test for ilr transformed series

At this point, it is necessary to determine how many lags are needed in each model by using the AIC and BIC criteria. As shown in Table 3.3, both the AIC and the BIC conclude that, for the model with the differences of the market capitalization, the more lags added the best. The number of lags was tested to up to 36, and the AIC continued to decrease at each time. Considering that the model should be parsimonious, the chosen number of lags for the estimated model was 12, consistent with the periodicity of the series. Therefore, the model to be estimated is expressed in Equation 3.4:

$$\Delta Z_t = \sum_{i=1}^{12} B_i \Delta Z_{t-i} + \epsilon_t \tag{3.4}$$

For the case of the compositional model, the results vary substantially. Thus, the information criteria show that the models with only one lag had the best adjustment, meaning that the model to be estimated is the one in Equation 3.5:

$$\Delta ilr(X_t) = \varrho \Delta ilr(X_{t-1}) + \varepsilon_t \tag{3.5}$$

No. of lags	Market capitalization differences		*Ilr* transformed series differences	
	AIC	BIC	AIC	BIC
1	16085.49	16150.43	-1843.73	-1804.76
2	16010.39	16127.09	-1823.78	-1755.70
3	15932.34	16100.63	-1803.59	-1706.50
6	15728.65	16050.69	-1749.24	-1565.68
12	15317.52	15942.24	-1620.94	-1267.14
24	14379.77	15589.55	-1481.95	-799.11
36	12544.07	14309.25	-1510.65	-515.45

Table 3.3: AIC and BIC results for the proposed models

Time series model estimation results

The results for the model in Equation 3.4 show that almost all coefficients are non-significant except, for a few coefficients in random lags which do not have further interpretation. For instance, in the equation for Chile, only the second lag of Mexico, the third of Peru, the eleventh of Chile and the twelfth of Colombia are significant; in the equation for Colombia, only the eleventh lag of Peru is significant. In the equation for Peru, the significant coefficients are its first and eighth lags; lags 2, 3, 8 and 11 of Chile; Mexico's eighth lag and Colombia's eleventh lag. Finally, in the equation for Mexico, its second lag, Peru's second and ninth lags and Chile's eleventh lag are significant. These results mean that it was not possible to find a dynamic process generating the data of the market capitalization of the MILA market.

This was confirmed by the Granger causality test (Granger, 1969), which shows that none of the variables contain information that can be used to explain the others. Table 3.4 reports the statistic and p-value results for the test with the null hypothesis that each dependent variable is explained by the others. Furthermore, when estimating the three separated models for the periods under study, the results were found to be very similar to those of the first model, i.e., most of the coefficients were not significantly different from zero. This implies that the model also has little predictive power in the subsamples, and that there is no difference between the three periods studied. The latter can be seen as a confirmation of the F-H puzzle, as there is no change in the coefficients of the model in the periods under analysis.

For the compositional model (Equation 3.5) the results are similar. Figure 3.7 reports the coefficient results for the compositional VAR model. In the left panel, V.1 corresponds to the proportion of Chile, Colombia and Peru with respect to Mexico, as explained by the first lags of itself and the other variables. V.2, in the middle,

Dependent variable	Statistic	P-value
Chile	1.33	0.10
Colombia	1.05	0.40
Peru	1.15	0.26
Mexico	0.83	0.75

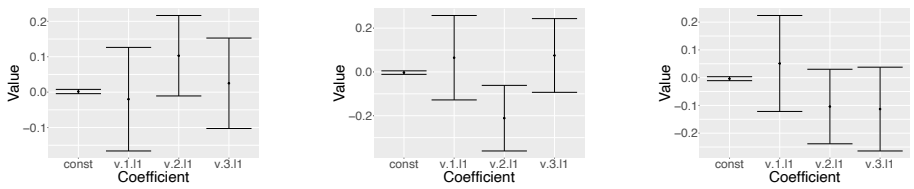
Table 3.4: Granger causality test for market capitalization model

corresponds to the proportion of Chile and Colombia with respect to Peru, as explained by the same variables as before; the right panel shows the estimation results for V.3, which is the proportion of Chile with respect to Colombia.

As can be seen in Figure 3.7, almost all coefficients were non-significant, except for the equation for V.2, in which its own lag is significant. Also, the separated models for the three periods exhibited a similar result. Furthermore, the Granger test concludes that there is no causality between the variables and the previous observations of the others (Table 3.5). Furthermore, when comparing the coefficients of the general model for the whole sample and the models for the three subperiods, the results were statistically the same for most of them, which means that it is not possible to find a recomposition of the integrated capital market after the entrance in force of the agreement. This would mean that the F-H puzzle still holds in this specific setting.

Dependent variable	Statistic	P-value
V.1	1.67	0.01
V.2	1.06	0.39
V.3	1.06	0.38

Table 3.5: Granger causality test for compositional model



(a) Estimated coefficients for V.1 (b) Estimated coefficients for V.2 (c) Estimated coefficients for V.3

Figure 3.7: Estimated coefficients compositional model in differences.

3.6 Conclusions

Different methods were applied to assess the F-H hypothesis, according to which the liberalization of capital markets does not necessarily lead to a movement of capital between countries. The particular case of the MILA was used by considering the specificity of the setting, in which only stock markets were liberalized to allow investors to trade shares in other markets as they would do locally. The hypothesis testing strategy was based on the market capitalization of the MILA market and the relative importance of each of the participants within the total, looking for any change in the composition of the market that could indicate a flow of capital between the stock exchanges. As a result, it is not possible to find a change in the participation of the four markets in the MILA caused by the entrance in force of the agreement. Therefore, there was no indication of capital flows within the markets after the implementation of the agreement, meaning that the F-H puzzle holds also in this specific setting.

From the methodological point of view, it was possible to use compositional methods to assess the hypothesis by analyzing the series as cross-sectional data, achieving consistent results. On the other hand, compositional and traditional time series models did not provide useful information to assess the hypothesis, as the estimated models did not deliver significant results. This sheds light on the potential of compositional methods for analyzing problems for which other approaches fail to come up with meaningful results.

This study contributes to the debate on the effects of capital markets' liberalization, and its findings can provide policymakers with arguments in favor or against the applications of certain policies, particularly in the field of international investment. Further research on the topic could aim to implement similar methodologies in different settings, such as the European Union, which not only allows for the mobility of capital, but also incorporates other factors, such as labor.

4 Too big to fail? An analysis of the Colombian banking system through compositional data[12]

This chapter constructs a concentration index for financial/banking systems via compositional analysis to establish the potential existence of "too big to fail" financial entities. The intention is to provide an early warning tool for regulators about this type of institution, so they can define thresholds and measures depending on their risk appetite and the systems' specificities. The index is applied to the Colombian banking system and assessed over time with a forecast to determine whether the system is becoming more concentrated. Both traditional and compositional time series methods are used for the forecast, comparing different alternative models to define which one is the most suitable for the problem in hand.

This chapter is divided into five sections. Section 4.1 introduces and motivates the analysis of this chapter, while Section 4.2 describes the methodology, emphasizing the compositional data framework and its application. Section 4.3 explores the Colombian banking system and the data set used in the analysis, which is included in section 4.4, together with the comparison and assessment of the proposed model. Finally, section 4.5 summarizes the findings and concludes.

4.1 Introduction

The term "too big to fail" (TBTF) has been in use since the 1980s for institutions that can pose "significant risks to other financial institutions, to the financial system as a whole, and possibly to the economic and social order" (Stern and Feldman, 2004, p. 1). Nevertheless, some authors did not consider that the TBTF concept should be central to banking regulation (Mishkin et al., 2006) until the financial crisis in 2008, when evidence emerged of how the effect of shocks in these institutions can expand without control. An anecdotal explanation on the development of the crisis

[12]This chapter is based on an article published on Latin American Journal of Central Banking (Vega Baquero and Santolino, 2022b).

can be found in Sorkin (2010). Some of the initial literature on TBTF institutions after the crisis include Shull (2010) and Zhou (2010). More recently, authors such as Barth and Wihlborg (2017), Ioannou et al. (2019), Omarova (2019) and Cetorelli and Traina (2021) have continued the discussion on TBTF institutions, focusing on their development after the crisis.

As a response to the crisis from the regulatory side, the Basel Committee on Banking Supervision (BCBS) started working on a new framework to reduce the risk of such a disruption in the future. This led to new definitions, especially those of global systemically important banks (G-SIBs) and domestic systemically important banks (D-SIBs)[13]. In general, Systemically Important Banks (SIBs) are large, interconnected financial institutions whose failure could generate widespread disruptions in the financial system, both domestically and internationally (Alzoubi et al., 2022). In the case of the D-SIBs, the definition includes not only the size, but also other characteristics such as interconnectedness, substitutability and complexity of the financial institutions, while for G-SIBs, the cross-national activity should also be considered. Both definitions are complementary, depending on the nature of each entity. The BCBS also mentions the importance of considering the potential impact of a D-SIB at an international level, either regional or bilateral, even if the effect at a global level might be insignificant. Moch (2018) carried out an interesting review of recent literature about SIBs. Among the most important conclusions, the author noted that there was consensus that the relationship between the size of banks and systemic risk was nonlinear. In other words, the risk increases more than proportionally to increases in the size of banks. Furthermore, the author found that concentration and interconnectedness in systems could lead to an amplified effect of size on systemic risk. Li et al. (2020) developed a recent attempt to measure the effect of SIBs on systemic risk. They tried to determine the importance of an institution in the system via changes in the systemic risk measured with and without it (a leave-one-out method), using z-scores as the risk measure. Another approach can be found in Bezrodna et al. (2019), who proposed a risk indicator that considered the importance of banks in terms of the size of their assets with respect to the total assets the banking system held and connected this with the riskiness of the entity to determine whether an SIB needed greater supervision than others did. This approach relies on the assumption that highly concentrated, interconnected systems are more vulnerable to distress in one SIB.

More specific to the United States context, the concentration of the financial system and the risk it imposes on competition and stability has been largely studied. For instance, before the crisis Cetorelli et al. (2007) studied the effects of concen-

[13] Basel Committee on Banking Supervision (2022).

trated markets to stability. More recently, Vives (2016) focused on competition and its effect on stability in light of the events of the financial crisis and Bikker and Spierdijk (2019) published a handbook discussing the methodologies available for policymakers to assess competition in the financial system. Indeed, regulators in the US have a wide range of indicators that allow them to estimate the status of competition in the financial sector such as the Lorenz Index and the Herfindahl-Hirschman Index. Notably, the latter is used in assessing of mergers and acquisitions.[14]

As can be seen, the size of banks and financial institutions continues to play a key role in determining whether an entity is considered TBTF. This study proposes a novel approach to evaluating concentration in financial systems through constructing a concentration index based on the relative size of each entity in terms of assets. Then, the main assumption of this analysis is that the importance of a financial entity on a system depends on its relative size, that is, its size in relation to the size of the other entities within the system. The composition of the relative participation of each financial entity on the whole system will determine the system's degree of concentration. The methodology presented uses the compositional data framework to estimate an indicator of concentration that can be tracked over time, to gain insight into a defined system's current and expected concentration. In the compositional data framework, relative information is more relevant than absolute values are. Individuals (in this case, financial institutions and more specifically banks) are not considered independently but as part of a whole (here the financial/banking system), measured in relative terms instead of absolute terms (van den Boogaart and Tolosana-Delgado, 2013).

For this study, the variable of interest is a bank's relative weight with respect to others in terms of assets. Therefore, the value of assets each bank holds is expressed in compositional terms to compare the actual composition of the banking system with the benchmark, defined as an ideal composition in which all entities have the same participation. This comparison is made through the Aitchinson distance (Aitchison, 1986), which measures the distance between two compositions, to create an indicator for the concentration of the system. However, this indicator per se does not explain the evolution of the system's concentration level. Hence, the indicator is assessed over time to define the system's concentration trend. Furthermore, a time series model could forecast the future behavior of the concentration of the system, which can be used as an early warning for regulators. For instance, regulators could set thresholds (following their risk appetite) for the deviations from the forecasted behavior or rules to react to facing changes in the trend toward a more concentrated system.

[14]More information on the application of this index in https://www.federalreserve.gov/bankinforeg/competitive-effects-mergers-acquisitions-faqs.htm.

This methodology is applied to the Colombian banking system, to estimate the concentration trend over the last decade and try to forecast its behavior in the following years. Furthermore, the estimated model is compared with other potential estimation methodologies and assessed for its estimation hypotheses.

Compositional methods have scarcely been used in finance and economics. Previous applications of this methodology in economics include Belles-Sampera et al. (2016), who made an initial approach to understanding capital allocation problems as compositional problems, and Boonen et al. (2019), who went beyond this to forecast risk allocations using the compositional data framework. Furthermore, approaches to forecasting compositional data have been taken by Mills (2010), Kynčlová et al. (2015) and Zheng and Chen (2017). All of them showed the benefits of using compositional data framework in different fields.

4.2 Methodology

This section explains the methodology used for the analysis, based on the definition of the compositional data framework and the centered log-ratio (clr) and isometric log-ratio (ilr) transformations explained in Chapter 2. Additionally, the criterion for the model selection and assessment of results are explained.

4.2.1 Compositional VAR model

Considering that the main objective of the models to be estimated is forecasting the composition of the Colombian banking system, several specifications are assessed to choose the one that suits the most the problem in hand and then use this model to see the expected behavior of the concentration of the market in the upcoming years. This modeling strategy follows the approach in Boonen et al. (2019), who analyzed the market risk in a stock portfolio.

Thus, for the models in Equations 2.10, 2.11, 2.12, and 2.13 the variables $Z = [z_1, \ldots, z_n]$ will be the value of the assets held by each institution. The final figure of institutions included will be defined in Section 4.3.

On the other hand, the compositional approach uses the participation of each institution to construct the composition $X = [x_1, \ldots, x_n]$. This is done my applying the closure operation from Equation 2.2 on the vector Z mentioned previously. Thus, this is the variable to use in the estimation of Equations 2.14, 2.15, 2.16, and 2.17.

Furthermore, for the estimation of Equations 2.11 and 2.15 the control variable A_t will be the total value of assets in the market at moment t, defined as $A_t = \sum_{i=1}^{n} z_{i,t}$.

The variable will be used in logarithm to eliminate the scale effect and lagged one period, meaning that for the estimation of Z_t the control variable will be $\log(A_{t-1})$.

To define the number of lags to be used in every model, the AIC from Chapter 2 is used. Additionally, the dataset will be diagnosed to test for the stationarity and cointegration assumptions using the augmented Dickey-Fuller test. With the information obtained, the models that suit the data's characteristics will be estimated.

4.2.2 Model comparison

When the estimated models have been obtained, the next step is to compare them. In this case, the AIC cannot be used because the models to be compared now have different dependent variables. Therefore, this comparison is achieved through the accuracy of the forecasts the models make (considering that the intended use for the estimated models is forecasting).

In this case, for a sample with T periods, an $h > 1$ number of periods can be selected. Then, for each period $k \in \{T-h, \ldots, T-1\}$, the model can be estimated with the information up to k. Next, with the estimated model, forecast periods $\{k+1, \ldots, T\}$ and the forecasted values, the deviation from the observed values can be calculated through the mean Aitchison distance of prediction errors (MADPE):

$$MADPE(k,m) = \frac{1}{T-k} \sum_{t=k+1}^{T} AD_{\Delta}(X_t, \hat{X}_t^{k,m}) \tag{4.1}$$

where $AD_{\Delta}(\cdot,\cdot)$ is the Aitchison distance defined in Equation 2.3 and $\hat{X}_t^{k,m}$ is the forecast for the period t with the model m estimated with information up to k. Again, as the MADPE measures the forecasts' accuracy through the distance between the observed and predicted values, the model with the lowest MADPE will be best for forecasting.

4.3 Colombian financial system data

The Financial Superintendence of Colombia[15] (financial regulator) defines credit establishments as all the entities that channel resources from the public (capturing them as deposits) to individuals and companies in need of liquidity (by providing credit). These are divided into four groups, depending on the means used to channel the resources: banking establishments that obtain resources from the market via current accounts or term deposits to provide credit; financial corporations,

[15] https://www.superfinanciera.gov.co

which channel resources to companies to promote their growth; commercial financing companies, which raise funds through fixed-term deposits to provide finance for the commercialization of goods and services, and leasing operations; and financial cooperatives that are authorized to provide credit to non-associated parties. Additionally, the Colombian financial system has special official institutions, which are government-financed entities providing development financing for specific purposes or to specific clients that the legal act creating each entity defines.

The study aims to analyze the financial system's risk of concentration with a low number of institutions managing most of the assets. Notably, the gorvernment funds special official institutions (exclusively or in a high proportion). In most cases, they are created with the purpose stabilizing the system through liquidity in situations of disruption or to provide liquidity to a specific public. Therefore, considering that their activity responds to rules different from those of the free competitive activity that govern the banking establishments, these entities will be excluded from the analysis.

Following financial regulation, all credit establishments ought to provide key financial indicators monthly. Considering that the aim is to analyze the financial system through the relative importance of each entity within the system, assets would be a good size indicator. The total assets in credit establishments have been growing in recent decades with banking establishments' considerable participation, which on average was 95.6% during the last decade. The next type of entity is financial corporations with 2.6%, followed by commercial financing companies with 1.4%, and finally financial cooperatives with 0.4%. Therefore, the composition of credit establishments shows very high relevance of banking establishments, while the rest are irrelevant regarding total assets. Furthermore, financial corporations, commercial financial companies and financial cooperatives use specific means to fund their credit operations. Therefore, the risk they are exposed to differs in some sense from that of banking establishments. Consequently, the data set to be used for the analysis will only contain banking establishments.

4.3.1 Colombian banking system

The data set considered is from January 2010 to April 2020, which provides a base availability of 124 observations for each entity. The series consists of 26 banking establishments. However, the financial system is dynamic. Therefore, four establishments did not have observations for the full period, either because they started operating after January 2010 or they ceased operations before April 2020. These institutions participated marginally in the system. Therefore, it is assumed that their effect on the outcome is negligible, and they have been excluded to avoid issues

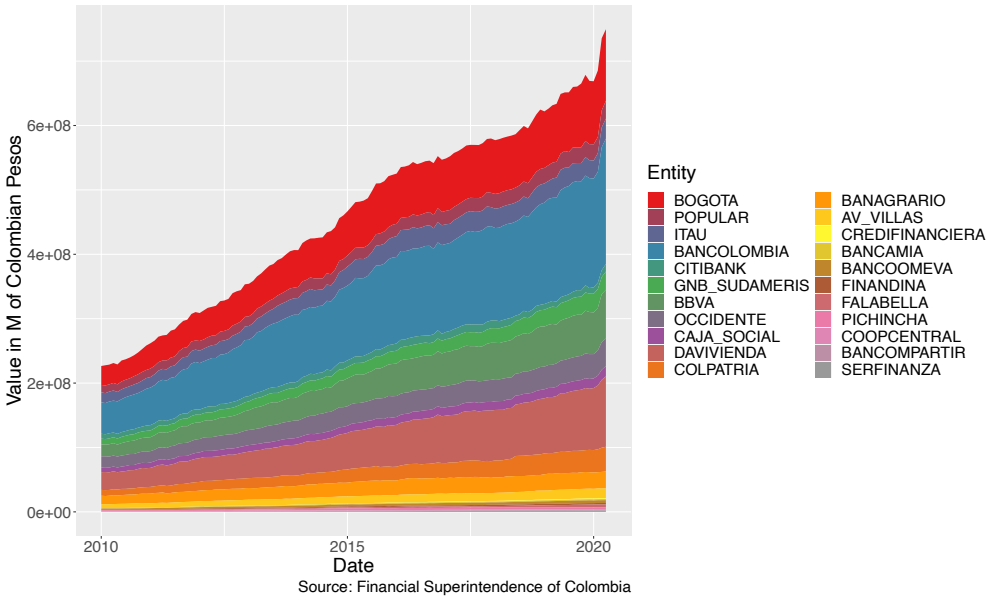


Figure 4.1: Total assets per banking establishment in Colombia from January 2010 to April 2020

with zeros in the data set.[16]

Hence, the data set to be used in the analysis consists of 22 banks and is summarized in Figure 4.1, which shows the total assets each banking establishment held. Figure 4.2 shows the relative proportion of the total assets (compositions) per entity. Three entities hold over 10.0% of the assets individually and can be considered as systemically important: Bancolombia, Banco de Bogota and Davivienda with 24.6%, 14.5% and 12.5%, respectively. Together, they account for more than 50.0% of the total assets in the banking system.

4.3.2 Concentration level

To analyze if there is a concentration of the assets within one or a small group of institutions, a benchmark is required for comparison. In this case, this is the uniform distribution of assets across all entities (or the neutral element of the perturbation operation for compositional data): $W = [w_1, ..., w_n]$ with $w_i = 1/n$. Therefore, the distance between this hypothetical distribution and the actual distribution of the assets among the entities should be measured. Thus, the Aitchison distance between the neutral composition and the observed compositions of the Colombian

[16]The concentration index was also calculated for the timeframes in which it was possible to have observations of each of these entities and the results are similar regarding the trend and the values, confirming the assumption that they can be neglected.

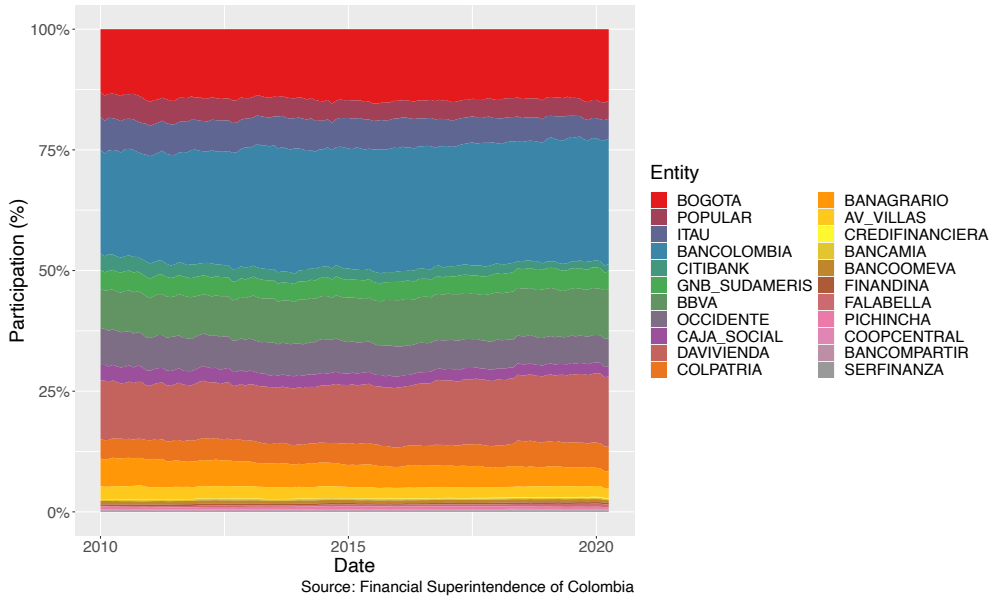


Figure 4.2: Assets composition per banking establishment in Colombia from January 2010 to April 2020

banking system's relative assets is computed.[17] Note that the aim is to evaluate the trend rather than the absolute values obtained because it is not possible to define what a high concentration would be regarding the value of the distance from an objective perspective. For instance, the indicator can virtually go to infinite as the composition approaches extreme values. Therefore, the explicit threshold to define excessive market concentration would depend on risk appetite of decision makers. The calculation of the monthly distances is shown in Figure 4.3. As observed, the distance has been decreasing throughout the period analyzed, which would signal a more equal distribution of assets among the banks, getting closer to the ideal uniform distribution.

Now, the next step is to obtain a model that can be used to predict the asset composition's behavior in the banking system, to assess the potential concentration of the market and use the outcome as an input for decision making on financial regulation, by creating an alert on potential TBTF institutions.

[17] The methodology proposed allows assessing the value of the distance between the observed composition and a composition used as benchmark. The neutral composition is chosen as benchmark, but other compositions may be chosen as benchmark based on the risk appetite of decision makers and the specificities of the system under analysis.

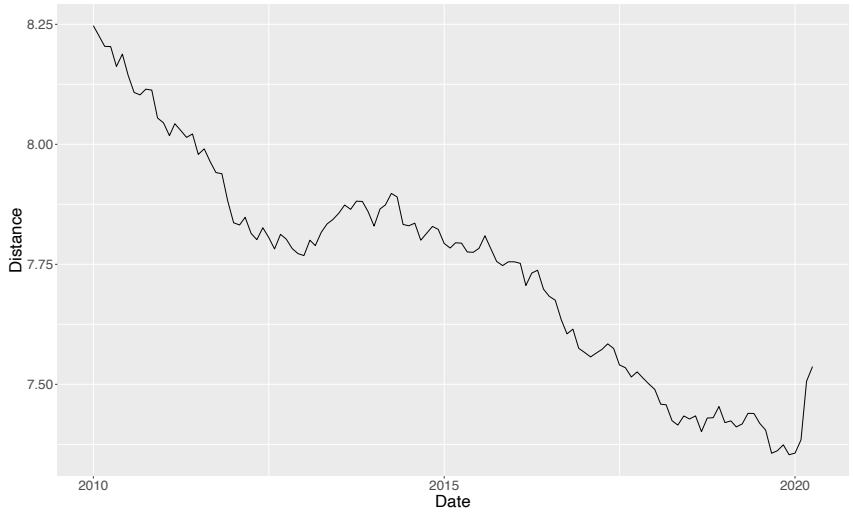


Figure 4.3: Observed concentration level index of the Colombian banking system: Aitchison distance between the actual composition of assets and the hypothetical composition in which assets are equally distributed among all entities

4.3.3 Data diagnosis

After defining the data set to be used, the time series needs to be diagnosed to determine which of the proposed models is more suitable to estimate. The first assumption in time series analysis refers to the stationarity of the series. A common test for stationarity is the augmented Dickey-Fuller test (Said and Dickey, 1984), mentioned in Chapter 2. In this case, the null hypothesis of unit roots in the characteristic equation is compared to the alternative of the stationarity of the process.

Hence, the augmented Dickey–Fuller test can be applied to each individual series in Z and X (i.e., the values and compositions of assets in the Colombian banking system, respectively). For the series of the value of assets banking institutions hold, the results show that there is no stationarity for 21 out of 22 entities at 5.0% of significance. This non-stationarity was expected because most time series showed an increasing trend in Figure 4.1. Similarly, for the series of compositions there is no stationarity in 19 out of 21 variables. Nevertheless, both sets are integrated of degree one, meaning that the first differences of each individual series of both Z and X are stationary.

Because the series to be analyzed are not stationary and are integrated of degree one, it is worth testing for cointegration. In this case, considering the problem has more than two variables, the test to be used is the Johansen cointegration test (Johansen, 1991). This test uses the cointegration matrix of the data set (which

contains the linear relations of the variables) and assesses the null hypothesis of no cointegration (the rank r of the matrix is zero) against the alternative of rank $r > 0$, meaning there is a cointegrating relationship between at least two of the variables. Subsequently, it evaluates $r \leq 1$ against $r > 1$ and continues recurrently until $r \leq n - 1$ (where n is the number of variables tested). At this point, if the null hypothesis is rejected, meaning $r > n - 1$, it can be assumed that the matrix is of full rank ($r = n$), that is, all the variables are cointegrated. Due to the data set's high dimension (22 variables in the case of the values of assets and 21 in the case of the ilr-transformed compositions), confidence intervals for the tests cannot be computed. Nevertheless, the values of the statistics are high. Therefore, cointegration in the series can be assumed, although it is not possible to know the rank of the cointegration matrix.

The diagnosis showed that the data set is non-stationary and cointegrated, which allows estimating any of the models presented previously, each with its advantages and disadvantages. In the absence of stationarity, VAR in differences and VEC models are the usual approaches to obtain the assumed distribution of the residuals. However, the VAR model (and its extended version) can still be estimated with consistent results. Therefore, the models are compared using the MADPE described previously to determine which of the models performs better regarding forecasting.

4.4 Results

Now that the methodology and the data set have been defined, the methodology must be applied and the results assessed. First, the number of lags p to be used for each model needs to be defined via the AIC. Then, the four proposed models will be assessed through the MADPE. Afterwards, the models' specification assumptions will be assessed, to confirm their correct specification. Finally, the model with the best performance will be used to forecast the composition of the Colombian banking system and evaluate its concentration trend.

4.4.1 Model selection

The AIC selection criteria's results for the models are shown in Table 4.1. As seen, for the proposed specifications using compositional data, one lag should be used for the VAR models.[18] For the models using the value of the assets the banks held, three lags should be used in the VAR models.

Therefore, the initial VAR models from Equations 2.10 and 2.14 will correspond to:

[18] In the case of equality, the model with the lowest number of lags is to be chosen for parsimony.

Model	N. of lags	AIC	
		Compositional model	Model with value of assets
VAR	1	-12,993.35	72,096.36
	2	-12,993.35	71,487.74
	3	-12,836.35	70,434.82
Extended VAR	1	-13,008.18	72,051.00
	2	-13,008.18	71,450.68
	3	-12,885.30	70,360.87
VAR in differences	1	-12,461.58	71,923.12
	2	-12,461.58	71,923.12
	3	-12,461.58	70,442.79

Table 4.1: AIC results for different models and number of lags

$$Z_t = \sum_{i=1}^{3} B_i Z_{t-i} + \epsilon_t \tag{4.2}$$

$$ilr(X_t) = \varrho ilr(X_{t-1}) + \varepsilon_t \tag{4.3}$$

Similarly, the extended models to be estimated from general Equations 2.11 and 2.15 are:

$$Z_t = \sum_{i=1}^{3} B_i Z_{t-i} + \gamma \log(A_{t-1}) + \epsilon_t \tag{4.4}$$

$$ilr(X_t) = \varrho ilr(X_{t-1}) + \gamma \log(A_{t-1}) + \varepsilon_t \tag{4.5}$$

Likewise, for the models in differences, the equations are:

$$\Delta Z_t = \sum_{i=1}^{3} B_i \Delta Z_{t-i} + \epsilon_t \tag{4.6}$$

$$\Delta ilr(X_t) = \varrho \Delta ilr(X_{t-1}) + \varepsilon_t \tag{4.7}$$

For the VEC models, the program automatically selects the number of lags during the estimation[19] and is set to two in both cases. Therefore, following the models defined in Equations 2.13 and 2.17 the models to be estimated will be:

[19]The package *vars* in R estimated the VEC models (Pfaff, 2008).

$$\Delta Z_t = \eta + H \cdot Z_{t-1} + \sum_{i=1}^{2} B_i^* \Delta Z_{t-i} + \epsilon_t \tag{4.8}$$

$$\Delta ilr(X_t) = \mu + M \cdot ilr(X_{t-1}) + \sum_{i=1}^{2} \varrho_i^* \Delta ilr(X_{t-i}) + \epsilon_t \tag{4.9}$$

Now, starting from the base availability of data (124 observations), the MADPE is calculated for all models, taking $h = 36$, meaning the model will be used to predict up to 36 months ahead (i.e., 3 years).

Figure 4.4 compares the basic VAR models with the extended models. The first notable aspect is that including the first lag of the total value of the assets (in log scale) as a control variable does not improve their forecasting power regarding the MADPE. Furthermore, although similar when predicting up to 18 periods ahead, the compositional models show a more consistent performance, with the lowest value of the MADPE almost all of the time.

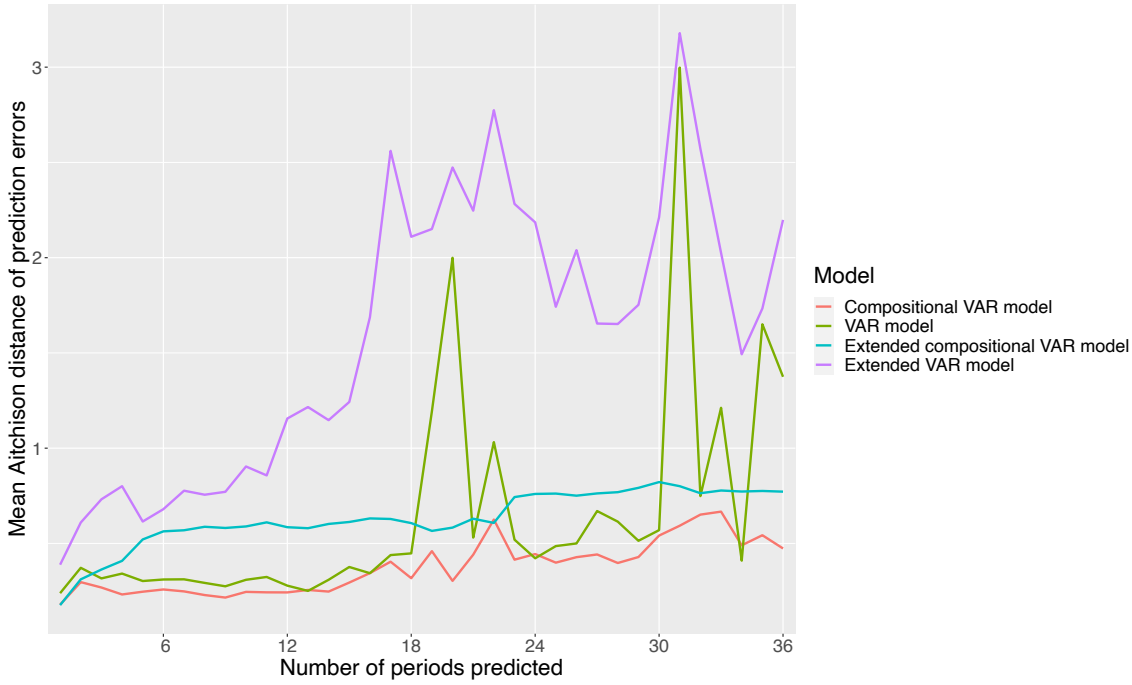


Figure 4.4: Mean Aitchison distance of prediction errors (MADPE): Compositional VAR model, VAR model, extended compositional VAR model and extended VAR model

Similarly, Figure 4.5 includes models in differences. As seen, the compositional model continues to outperform the model with the values of the assets. In addition, the model in differences seems to perform better compared to the basic composi-

tional VAR. However, the difference is very small, and it is not possible to conclude whether this difference is significant because the confidence intervals are not available.

The VEC models compared in Figure 4.6 show interesting results. VEC models for compositional data and for the value of the assets perform very similarly to the basic compositional VAR model. This could be due to that the modeled data is cointegrated. A model that considers this stylized fact would perform better. Nevertheless, the basic VAR model for compositional data shows very similar results to those of the other models for compositional data and to the VEC for the values of assets. An initial hypothesis would suggest that even if cointegration is not taken into account, the data expressed in compositional terms already considers the interaction between variables through explaining them in terms of relative importance with respect to the others. This, combined with the fact that VAR coefficients are consistent even in the absence of stationarity, could determine why compositional models perform similarly regardless of the specification. In addition, the basic compositional VAR model has other advantages, such as reduced manipulation of the data set and the lower number of parameters to be estimated. Thus, the compositional VAR model would require the estimation of $[(n-1)\times(n-1)\times p]+n-1$ coefficients and the VAR model for the value of assets $(n\times n\times p)+n$ coefficients. In the case of the VAR in differences models, the disadvantage comes from the fact that there are not T observations but $T-1$ observations. This reduces the estimation's degrees of freedom, while VEC estimations are even more complex, and the parameters cannot be easily interpreted.

Furthermore, this leads to another finding: The classical multivariate time series models for the value of asset banking institutions held are very sensitive to the specification. Indeed, misspecification can lead to a considerable decrease in the model's performance. Moreover, these models seem to be more sensitive to shocks in the series. This can be seen in the peaks in the MADPE when forecasting 19/20, 31 and 35/36 periods ahead, which are especially notorious in the basic VAR models. For example, for the model forecasting 20 periods ahead, the data set used for the estimation runs from January 2010 to August 2018 (and September 2018 for the forecast of 19 periods ahead). However, there was a sharp increase in the assets the two main banking institutions held in October 2018, which might explain why there is a bigger error when trying to forecast with a model estimated without information on this specific period. Relevant news for this period informed about good results for the financial system (notably for these entities), but there was no mention of acquisitions or any other important change in the assets' compositions. Similar relationships can be found in the other peaks in the MADPE for the VAR models for the value of assets.

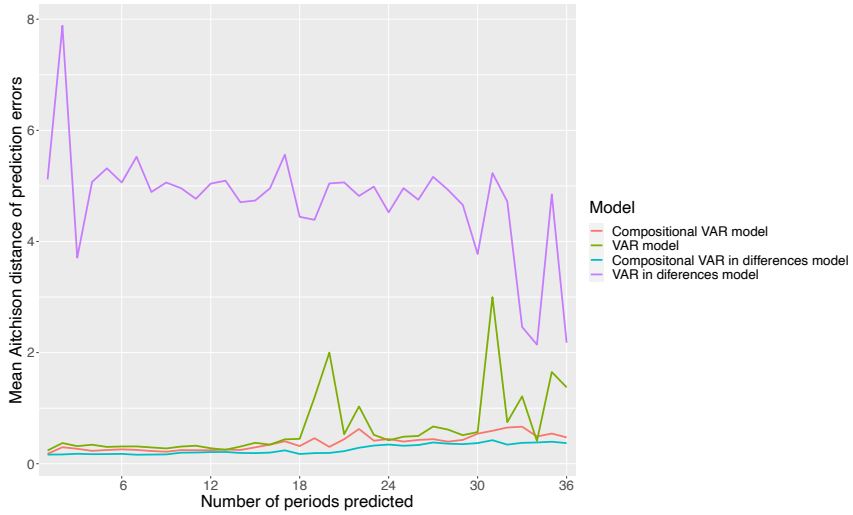


Figure 4.5: Mean Aitchison distance of prediction errors (MADPE): Compositional VAR model, VAR model, compositional VAR in differences model and VAR in differences model

Considering all these findings, the basic VAR model with compositional data (basic compositional VAR model) has more advantages compared to the other specifications. Therefore, the assumptions are assessed for this model. In addition, to maintain equivalence to the models using the value of assets, the same tests will be performed on the model in Equation 4.2 (basic VAR model).

4.4.2 Model diagnosis

Once the models have been defined, the assumptions of the models must be tested to determine whether they are correctly specified. For VAR models, the usual tests include Granger causality, autocorrelation, heteroscedasticity and normality of the residuals.

VAR models describe the joint generation process of numerous variables over time. Although in this chapter VAR models are used for forecasting, they can also be used for investigating relationships between variables, which the Granger causality test verifies (Granger, 1969). This test can be interpreted as a significance test for VAR models because it assesses whether adding lags of one variable improves the forecast of the other(s), that is, it contains valuable information for explaining the other variable(s) in the model. For the compositional VAR, the test shows that 5 out of the 21 series do not Granger cause the others at 5.0% significance (and only 2 when the significance is 10.0%). In the case of the VAR for the series of assets, only one of the 22 entities does not Granger cause the others.

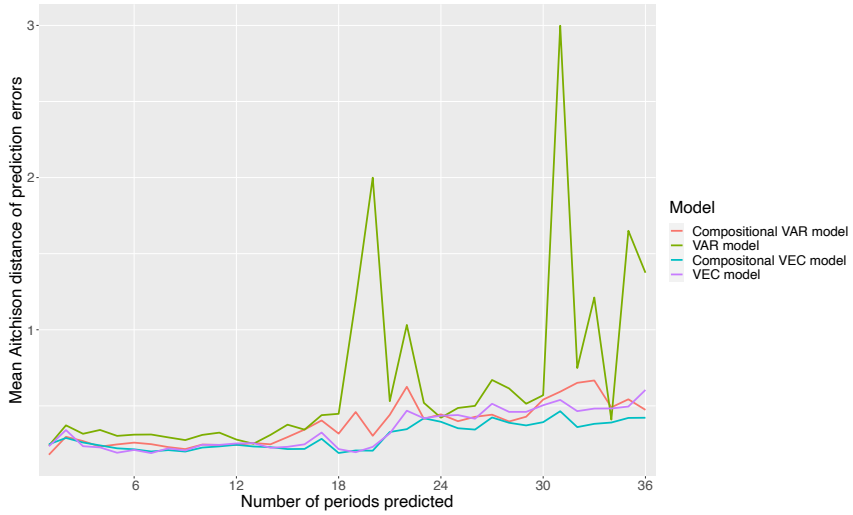


Figure 4.6: Mean Aitchison distance of prediction errors (MADPE): Compositional VAR model, VAR model, compositional VEC model and VEC model

The autocorrelation test determines whether the residuals are independently distributed, which is one of the main assumptions for the estimation. Following the results for the Portmanteau test, the residuals from the compositional VAR do not show signs of serial correlation, while those from the VAR for the value of assets are autocorrelated.

Regarding the homoscedasticity of the residuals, a multivariate test cannot be performed because of the data's high dimensionality. Therefore, individual tests for the residuals of each equations are performed. For each residual series, conditional heteroscedasticity models with a different number of lags (up to 20 in this case) were estimated. Meaning, that for the compositional model, there were a total of 420 models (21 series of residuals by 20 number of lags), while for the model for the value of assets, the figure is 440 (22 series of residuals by 20 number of lags). For both the compositional and the value of assets models, at 5.0% significance, there is evidence of heteroscedasticity for at least one of the tested number of lags in four series of residuals. To obtain overall insight into the results, the proportion of models in which there is evidence of heteroscedasticity with respect to the total can be estimated to construct a pseudo-statistic for heteroscedasticity for the VAR models. For the compositional VAR, this ratio corresponds to 8.1% (34 out of 420), while for the VAR for the value of assets, it is 7.3% (32 out of 440). In both cases, at 5.0% of significance, the hypothesis of heteroscedasticity cannot be rejected, but this is possible at 10.0%.

The residuals from the compositional model do not appear to be normally distributed, while those from the model for the value of assets are normally distributed.

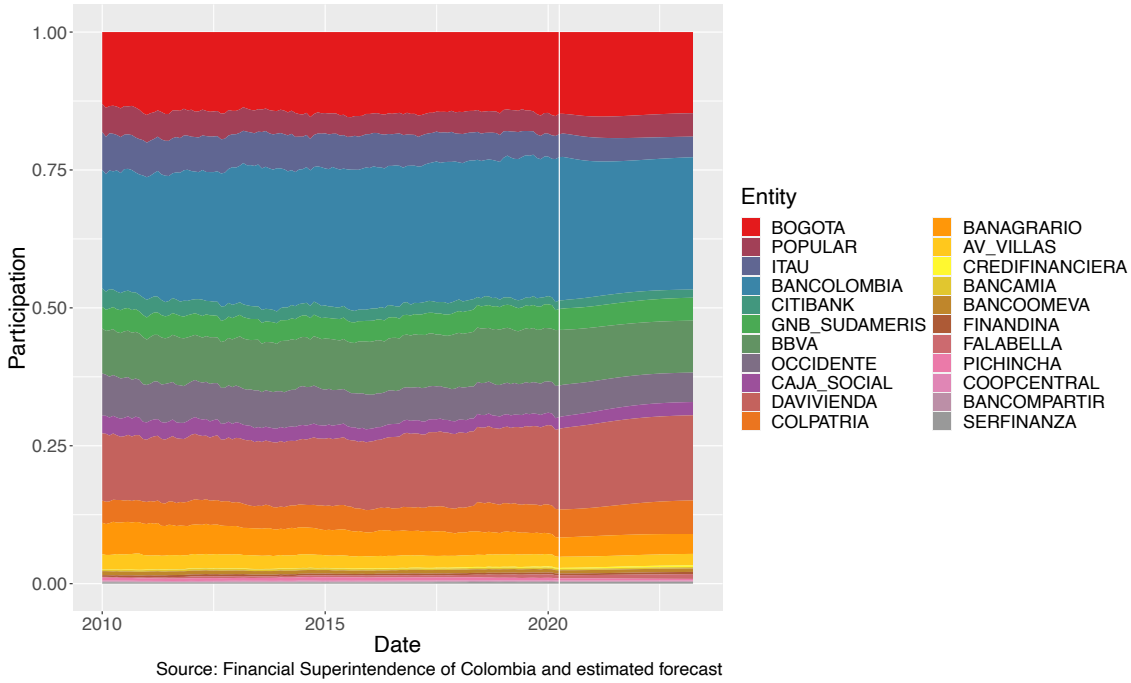


Figure 4.7: Assets composition per banking establishment from January 2010 to April 2020 and forecast for May 2020 to April 2023

Non-normal errors may violate some of the assumptions for the variance properties; therefore, some caution should be taken in coefficient inference. However, in this case, the model is used for forecasting and the normality of residuals is not required (Lütkepohl, 2007).

4.4.3 Forecast

After assessing the selected model (compositional model from Equation 4.3), the next step is to generate the forecast for 36 periods ahead (3 years from May 2020 to April 2023) to establish the expected composition of the Colombian banking system in the coming months and the concentration trend of the assets. Figure 4.7 shows the expected composition of the assets including the forecasted period to the right of the white line. As seen, the participation of each bank in the system is not expected to undergo major changes in the coming years. There is only a slight increase in the participation of Davivienda, the third entity, regarding participation, which might lead it to surpass Banco de Bogota (the second one). Apart from that, the financial system's composition seems to remain similar to that observed in previous years.

However, as can be seen in Figure 4.8, the decreasing trend of the distance between the benchmark and the actual composition of assets in the Colombian bank-

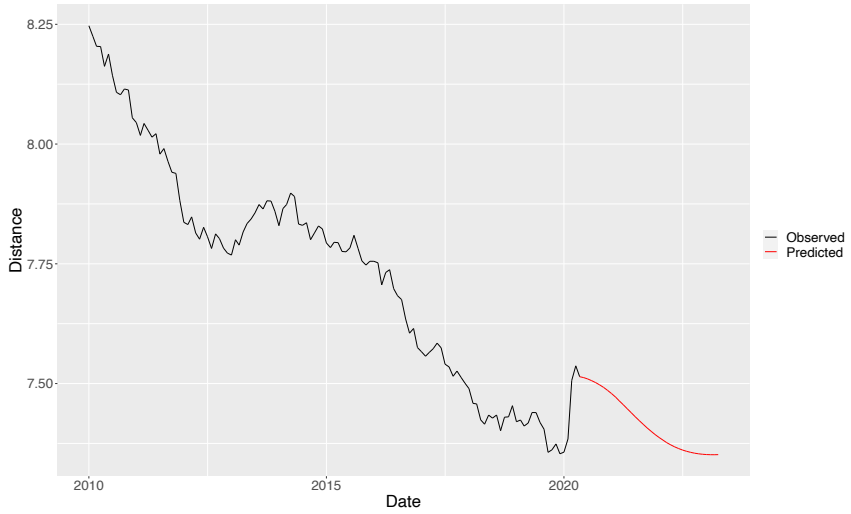


Figure 4.8: Predicted concentration level index of the Colombian banking system: Aitchison distance between the predicted composition of assets and the hypothetical uniform composition

ing system is expected to continue according to the model, despite the increase seen in the beginning of 2020. It tends to stabilize by the end of the forecasted period. This result is important regarding financial stability policy because, at least for the moment, the model does not predict that the overall system is becoming more concentrated, which would threaten the stability. Furthermore, the composition of the assets across banking institutions is expected to remain stable, without major changes in the coming months.

4.5 Conclusions

Compositional data methodologies have gained popularity in recent years as a new perspective to study and model phenomena in which relative information is more relevant than absolute values. In this study, this framework is used to propose an indicator for concentration in financial systems, as applied to the Colombian banking system. The main goal was to analyze whether the difference in the participation between large and small financial institutions has been decreasing or increasing in recent years and to predict their expected evolution in the future. The concentration of the banking system was estimated as the distance between the actual composition of the financial system and the hypothetical composition in which assets are equally distributed among all entities, which was monitored over time. Therefore, the larger the distance, the higher the degree of concentration of the system in a few entities.

This analysis is relevant regarding policy making, considering the lessons learnt during the 2008 financial crisis when the global financial system was at high risk because of some "too big to fail" institutions. Thus, the proposed indicator can be used as an additional early warning for regulators about this kind of institution, to reinforce monitoring them and to maintain the overall system's stability. Furthermore, the indicator can be extended to assess the effect of mergers and acquisitions between entities on the concentration of the system and, consequently, on its stability. Conversely, it is worth mentioning that the definition of the financial system and the banking system may vary depending on the context. There are multiple other actors that have an influence in the financial system and its overall risk. For instance, shadow banking and more recently, fintech companies offer services that should also be considered when assessing financial risk (Adrian and Ashcraft, 2016). Thus, the model can be extended to account for the specificities of these entities and their effect in the system's stability.

Compositional multivariate time series methods were used to predict the banking system's future composition. These methods have shown increased performance in the prediction, considering multiple stylized facts the data showed. One of the most remarkable results is that using compositional methods provides a more robust model with lower sensibility to outliers in the data set. Furthermore, the compositional framework appears to catch the intrinsic connections between all entities in the system, which would need to be modeled through cointegrated models (such as vector error correction models). Additionally, the model has shown that it fulfils the estimation assumptions, particularly regarding autocorrelation and homoscedasticity of errors. In terms of the expected future behavior of the Colombian banking system's composition and its concentration index trend, the forecast for the next 3 years shows little variation in the participation of each entity. Furthermore, the deconcentration trend that the banking system has shown in the last decade is expected to continue over the coming months.

To conclude, the methodology applied in this chapter opens opportunities for multiple applications in the context of financial risk and stability. This methodology is flexible enough to be adapted to other contexts, where a different number of entities (either much higher or much lower) or a different concentration pattern of assets among the entities could lead to interesting results. For instance, special official institutions were excluded from the study because their activity does not always follow the free market rules. Given that the literature has identified a potentially significant policy trade-off between competition and stability, the concentration level index was predicted including these entities to have an idea of the effect of these firms' omission on the results, and a more stable trend of the predicted concentration index was observed. Additionally, total assets were used in this study as a

measure of size. In general, most definitions of SIBs include assets as the metric for the size of the entities because total assets reflect loans and other components of bank's portfolios, which multiple other types of assets can form and are also considered in the risk metrics regulators require. Nevertheless, using alternative size metrics such as total loans can also be an interesting extension of the analysis and can also be considered by regulators, depending on the specific context of each country. Likewise, the methodology can be used to measure risk within entities. Some potential applications are analyzing portfolios as compositional data and assessing risk exposure to specific assets. Nevertheless, the proposed methodology does not consider the entrance and exit of actors in the system because this would require dealing with zeros in the compositions. Indeed, this study has excluded those banks entering and leaving the market, considering they had low participation. However, this limitation can affect the results, especially in longer time series when relatively important participants can enter or leave the market. This opens an opportunity for further developing the model, considering the existing literature on how to transform compositional data in the presence of zeros (Aitchison, 1986).

5 Proportionality between allocations in asset management[20]

Asset allocation refers to deciding the optimal participation of each asset within a portfolio. Therefore, these participations are a composition, and compositional methods should be used to treat the data and perform analysis over it. When trying to find relationships between parts of a composition, proportions have shown to be more suitable than correlations. In this chapter, the proportionality approach is used to analyze the asset allocation in a portfolio composed of five stocks from the IBEX 35 (the Spanish stock market index).

The chapter is structured as follows. Section 5.1 introduces the concepts of asset allocation and proportionality, while Section 5.2 explains in more detail the asset allocation method used to obtain the compositions. Then, Section 5.3 elaborates on the methodology applied and Section 5.4 shows the dataset. Results are shown in Section 5.5 and Section 5.6 concludes.

5.1 Introduction

Asset and capital allocation are important parts of portfolio and risk management for any business. From one side, asset allocation is related to the optimal portfolio selection, introduced by Markowitz (1952), and refers to the construction of portfolios that maximize the expected returns while minimizing risk. On the other hand, within Enterprise Risk Management (ERM), risk capital allocation is the partition of a specific capital (usually the capital requirements set by regulatory entities) among the different sources of risk (Dhaene et al., 2012). The set of guidelines to solve this kind of problems are commonly called capital allocation principles.

Despite being similar in the sense that both asset and capital allocation consider interconnectedness of the components and their risk to define participations, there are differences to consider. A very relevant one is that while capital allocation is derived from analyzing the contributions of individual risks to potential aggregate

[20]This chapter is based on a working paper in progress included in the IREA Working paper series (Vega Baquero and Santolino, 2025).

losses, asset allocation is an investment decision aiming at maximizing profit with minimum risk.

This chapter focuses on asset allocation. Following its construction, the Markowitz (1952, 1991) minimum variance portfolio principle has a compositional nature and, hence, any analysis needs to acknowledge this characteristic. However, up until now there has been no compositional analysis on asset allocation. Conversely, although still incipient, the compositional nature of capital allocation has been analyzed by Belles-Sampera et al. (2016), Boonen et al. (2019), Fiori and Porro (2023) and Fiori and Rosazza Gianin (2025), for example, showing how the compositional approach can be a relevant and desirable tool for this matter.

Consequently, it is necessary to use what is known as the Aitchison geometry (Aitchison, 1986). In this framework, the logratio variance is defined as a measure of the proportionality between parts of a composition. Following this approach, Lovell et al. (2015) showed that proportions are more suitable than correlations to find relationships between parts of compositions (the assets in the portfolio, in this case), whilst Egozcue and Pawlowsky-Glahn (2023) assessed different functional forms to improve the interpretability of the logratio variance and formulated the Proportionality Index of Parts (PIP).

In this chapter, the logratio variance is revisited from the point of view of subcompositional coherence to propose a new proportionality index. This approach focuses on pairs of parts of the composition, which is the smallest subcomposition possible, to derive the maximum and minimum limits for the logratio variance and obtain an interpretable index. This proportionality methodology is applied to a portfolio composed by the top 5 stocks by market capitalization traded in the Spanish stock market. The optimal weights for each asset are calculated daily using the minimum variance portfolio principle, from January 1st, 2021 to June 30th, 2024.

As a result, the stocks showed low proportionalities. However, one of the assets showed consistently higher proportionalities to the others. A deeper look into this, showed that this stock exhibits the lowest volatility for most of the period under analysis, which would be consistent with the expectation that the allocation method would assign the less volatile asset a higher participation and assign the others using this one as benchmark.

As a limitation for this study, it is not possible to assess the behavior of the index for high proportionalities, since none of them appeared to be high. Hence, more work is required to better understand this matter.

5.2 Mean-Variance portfolio theory

Portfolio and risk management include a large set of components, from which asset and capital allocation are particularly relevant. The first to discuss asset allocation and introduce modern portfolio theory was Markowitz (1952), who discussed the trade-off between the two objectives of investors: maximize the expected returns and minimize risk. This is the main idea of the Mean-Variance portfolio theory, which tries to find all portfolios that satisfy the two objectives. Therefore, it can be formulated as the maximization of the returns given a maximum accepted risk or as the minimization of risk given a minimum expected return. For the sake of this analysis the latter formulation will be used, which is the minimum variance portfolio approach. Hence, the problem can be expressed as in Markowitz (1952, 1991):

$$\min_{\boldsymbol{w}} \quad \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$$

Subject to:

$$\boldsymbol{w}^\top \mathbf{1} = 1 \quad \text{(Fully invested)}$$

where $\boldsymbol{w}$ is the vector of asset weights and $\boldsymbol{\Sigma}$ is the covariance matrix of asset returns[21]. The solution of this problem is a set of weights $\boldsymbol{w}^*$ that contains the participation of each asset within the total of the allocation. Therefore, $\boldsymbol{w}^*$ has a compositional nature, as explained in more detail in the next section. As such, when assessing the existence of relationships between the components of the allocation, correlation is not the most suitable tool and proportions are a better approach, as concluded by Lovell et al. (2015).

5.3 Compositional data and proportionality

Ever since Pearson (1897) mentioned the issue of spurious correlation, followed by Chayes (1960) deepening on the closure problem for certain variables, the measurement of the relationship between parts of a composition has remained an ongoing discussion. As part of his lifelong interest in compositional data, Aitchison (1986) proposed the logratio variance, defined in Equation 2.8.

As can be seen, this logratio variance is zero when the two variables are exactly proportional. Nevertheless, it is noticeable that there is no upper bound for this indicator, which means that the magnitude itself does not provide much information about the proportionality of two variables. However, following Lovell et al.

[21] It is possible to add the constraint $\boldsymbol{w}^\top \boldsymbol{\mu} = \mu$ for target return μ.

(2015), it is possible to compare the pairwise proportionalities of several variables and define which of these pairs are more proportional than others.

To overcome this, Egozcue and Pawlowsky-Glahn (2023) proposed a set of measures of association based on the logratio variance, aiming at improving the interpretability issue, and assess the desirable characteristics of them to find the one that behaves best. The result is the Proportionality Index of Parts (PIP), which has the form:

$$PIP(i,j) = \frac{1}{1+\sqrt{\tau_{ij}}} \tag{5.1}$$

These limits indicate that the PIP takes the value of 1 if $\tau_{ij} = 0$, i.e., when logratios $log(x_i/x_j)$ are constant. For the PIP to take the value of zero, it would be necessary that τ_{ij} goes to infinite. Hence, following the properties of logarithms and variances, it is possible to consider the following finite upper and lower limits for τ_{ij} by decomposing it as follows:

$$\begin{aligned}
\tau_{ij} &= Var\left(log(x_i/x_j)\right) \\
\tau_{ij} &= Var\left(log(x_i) - log(x_j)\right) \\
\tau_{ij} &= Var\left(log(x_i)\right) + Var\left(log(x_j)\right) - 2Cov\left(log(x_i), log(x_j)\right)
\end{aligned} \tag{5.2}$$

Now, since the interest is in the relationship between pairs of parts of the composition, it is possible to consider the subcomposition formed only by x_i and x_j, which is the minimal subcomposition possible with two elements, and apply the closure operation, following the definitions in Equations 2.6 and 2.7 to have:

$$\begin{aligned}
\tilde{x}_{ij} &= \left(\frac{x_i}{x_i + x_j}\right) \\
\tilde{x}_{ji} &= \left(\frac{x_j}{x_j + x_i}\right)
\end{aligned} \tag{5.3}$$

Now, replacing x_i and x_j for $\tilde{x}_{ij}$ and $\tilde{x}_{ji}$, respectively, in Equation 5.2[22]

$$\tau_{ij} = Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right) - 2Cov\left(log(\tilde{x}_{ij}), log(\tilde{x}_{ji})\right) \tag{5.4}$$

[22]This will not change the relationship τ_{ij}, since it is invariant under the subcomposition X_m (Egozcue and Pawlowsky-Glahn, 2023). It is straightforward to see that since $log(\tilde{x}_{ij}/\tilde{x}_{ji}) = log\left(\frac{x_i/(x_i+x_j)}{x_j/(x_j+x_i)}\right) = log(x_i/x_j)$, the value of τ_{ij} is the same, regardless of the use of x_i and x_j or $\tilde{x}_{ij}$ and $\tilde{x}_{ji}$

Since the covariance in Equation 5.4 is always negative[23], this new relationship has a lower bound which is attained when the $Cov\left(log(\tilde{x}_{ij}), log(\tilde{x}_{ji})\right) = 0$, implying the highest possible relationship between the variables. Consequently, it is possible to define:

$$\tau_{ij}^{MIN} = Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right) \tag{5.5}$$

On the other hand, regarding the maximum value τ_{ij} can take, following the Cauchy–Schwarz inequality [24], the limit for the covariance is given by the covariance inequality (Keener, 2010) $|Cov(A,B)| \leq \sqrt{Var(A)} \cdot \sqrt{Var(B)} = \sigma_A \cdot \sigma_B$, where $\sigma_A = \sqrt{Var(A)}$ and $\sigma_B = \sqrt{Var(B)}$ are the standard deviations of A and B, respectively. This implies that:

$$\tau_{ij}^{MAX} = Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right) + 2\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})} \tag{5.6}$$

Setting these limits for τ_{ij} allows for a new definition of the proportionality as the difference between its maximum and the observed values, normalized by its full range:

$$Prop_{ij} = \frac{\tau_{ij}^{MAX} - \tau_{ij}}{\tau_{ij}^{MAX} - \tau_{ij}^{MIN}}$$

$$Prop_{ij} = \frac{Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right) + 2\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})} - \left[Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right) - 2Cov\left(log(\tilde{x}_{ij}), log(\tilde{x}_{ji})\right)\right]}{Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right) + 2\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})} - \left[Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right)\right]}$$

$$Prop_{ij} = \frac{Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right) + 2\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})} - Var\left(log(\tilde{x}_{ij})\right) - Var\left(log(\tilde{x}_{ji})\right) + 2Cov\left(log(\tilde{x}_{ij}), log(\tilde{x}_{ji})\right)}{Var\left(log(\tilde{x}_{ij})\right) + Var\left(log(\tilde{x}_{ji})\right) + 2\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})} - Var\left(log(\tilde{x}_{ij})\right) - Var\left(log(\tilde{x}_{ji})\right)}$$

$$Prop_{ij} = \frac{2\left(\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})} + Cov\left(log(\tilde{x}_{ij}), log(\tilde{x}_{ji})\right)\right)}{2\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})}} = \frac{\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})}}{\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})}} + \frac{Cov\left(log(\tilde{x}_{ij}), log(\tilde{x}_{ji})\right)}{\sigma_{log(\tilde{x}_{ij})}\sigma_{log(\tilde{x}_{ji})}}$$

$$Prop_{ij} = 1 + \rho_{log(\tilde{x}_{ij}), log(\tilde{x}_{ji})} = 1 + \rho_{log(\tilde{x}_{ij}), log(1-\tilde{x}_{ij})}$$

[23] Because of the closure operation, $\tilde{x}_{ji} = 1 - \tilde{x}_{ij}$, which implies that whenever one of the components increases its participation, the other one will decrease and, therefore, the covariance will be negative.

[24] Expressed in its general form as: $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|$, where $\mathbf{u}$ and $\mathbf{v}$ are two vectors, $\langle \cdot, \cdot \rangle$ is the inner product and $\| \cdot \|$ is the norm (Ford, 2015).

where $\rho_{log(\tilde{x}_{ij}),log(\tilde{x}_{ji})}$ is the Pearson's correlation coefficient, defined as: $\rho_{A,B} = \frac{Cov(A,B)}{\sigma_A \cdot \sigma_B}$. Since this correlation is always negative, it will be in the interval $[-1, 0]$, with 0 being the highest proportionality. Thus, $Prop_{ij}$ will be in the interval $[0, 1]$ with 0 meaning no proportionality and 1 meaning complete proportionality. As can be seen, this index depends only on the relative importance of the two elements in question within the subcomposition formed by them, meaning that it is not affected by the other components.

Nevertheless, it is worth mentioning the case $i = j$. As can be seen, in this case $\tilde{x}_{ij} = \tilde{x}_{ji} = \tilde{x}_{ii} = 0.5$, which is a constant. Consequently, $Var\left(log(\tilde{x}_{ii})\right) = 0$ and $\tau_{ii} = \tau_{ii}^{MAX} = \tau_{ii}^{MIN} = 0$, so $Prop_{ii}$ is not defined. Therefore, since this case is the maximum proportionality possible, it is defined as $Prop_{ii} = 1$. Then, the full definition of the proportionality index would be:

$$Prop_{ij} = \begin{cases} 1 + \rho_{log(\tilde{x}_{ij}),log(1-\tilde{x}_{ij})} & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases} \tag{5.7}$$

5.4 Data and empirical approach

When measuring relationships between assets in a portfolio, subcompositional coherence is relevant because of the nature of portfolios. Depending on the policies defined by each investor, the composition of portfolios is updated frequently, and having a measure of dependence that is not affected by the entrance and exit of assets in the portfolio can be useful and desirable.

As an example, the methodology proposed is used to assess the proportionalities between stocks in a portfolio composed by the five stocks with the highest market capitalization of the IBEX35, the main reference index of the Spanish stock market. The participations in the portfolios are calculated daily from January 1st, 2021, to June 30th, 2024, following the Markowitz (1952) capital allocation method and using the previous month of observations to estimate the volatility. For the selected stocks, the price used is the daily closing price adjusted for splits, dividends, and capital gain distributions.

Hence, following the definition in 5.3, the optimal weights $\boldsymbol{w}^*$ for any pair of assets i and j can be expressed as:

$$\begin{aligned} \tilde{w}^*_{ij} &= \left(\frac{w^*_i}{w^*_i + w^*_j} \right) \\ \tilde{w}^*_{ji} &= \left(\frac{w^*_j}{w^*_i + w^*_j} \right) \end{aligned} \tag{5.8}$$

Then, following equation 5.7 the pairwise proporcionality between assets' allocations in risk management will have the form:

$$Prop_{ij} = \begin{cases} 1 + \rho_{log(\tilde{w}^*_{ij}),log(1-\tilde{w}^*_{ij})} & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases} \tag{5.9}$$

5.5 Results

Table 5.1 shows the proportionality matrix for the stocks in the proposed portfolio: Banco Santander (SAN), Iberdrola (IBE), Industria de Diseño Textil – INDITEX (ITX), Banco Bilbao Vizcaya Argentaria (BBVA), and Caixa Bank (CABK). Although the proportionalities are rather low in general, it is worth noticing that those of Iberdrola with the other stocks are the highest ones.

	SAN	IBE	ITX	BBVA	CABK
SAN	1	0.33	0.25	0.25	0.24
IBE	0.33	1	0.32	0.29	0.33
ITX	0.25	0.32	1	0.25	0.27
BBVA	0.25	0.29	0.25	1	0.24
CABK	0.24	0.33	0.27	0.24	1

Table 5.1: Proportionality matrix for the proposed portfolio.

Enquiring into the potential causes for this, one can look at the allocation method used to obtain the optimal weights $\boldsymbol{w}^*$. As mentioned in section 5.2, it is the minimization of the variance what defines the optimum. Figure 5.1 shows the 30-day volatility of the five stocks. It is possible to see that, for most of the period, Iberdrola exhibits the lowest volatility. This makes the allocation method put a higher weight on this stock and it is possible that the participation of the other stocks in the portfolio keeps a closer relation with this stock than to others.

Nevertheless, it is necessary to look deeper into the proportionalities and compare different scenarios to understand better the behavior of the indicator.

5.6 Conclusions

The logratio variance has shown interesting properties when analyzing compositional data, particularly the proportionality between its parts. Recent work aiming at analyzing proportionality has proposed indexes that overcome the lack of scale of the logratio variance. This chapter proposed a different approach, by using properties of the variance and re-expressing subcompositions to assess proportionality using a different index. With the new index, it is possible to define the maximum and minimum possible values of the logratio variance, and express the index in relative terms to these limits.

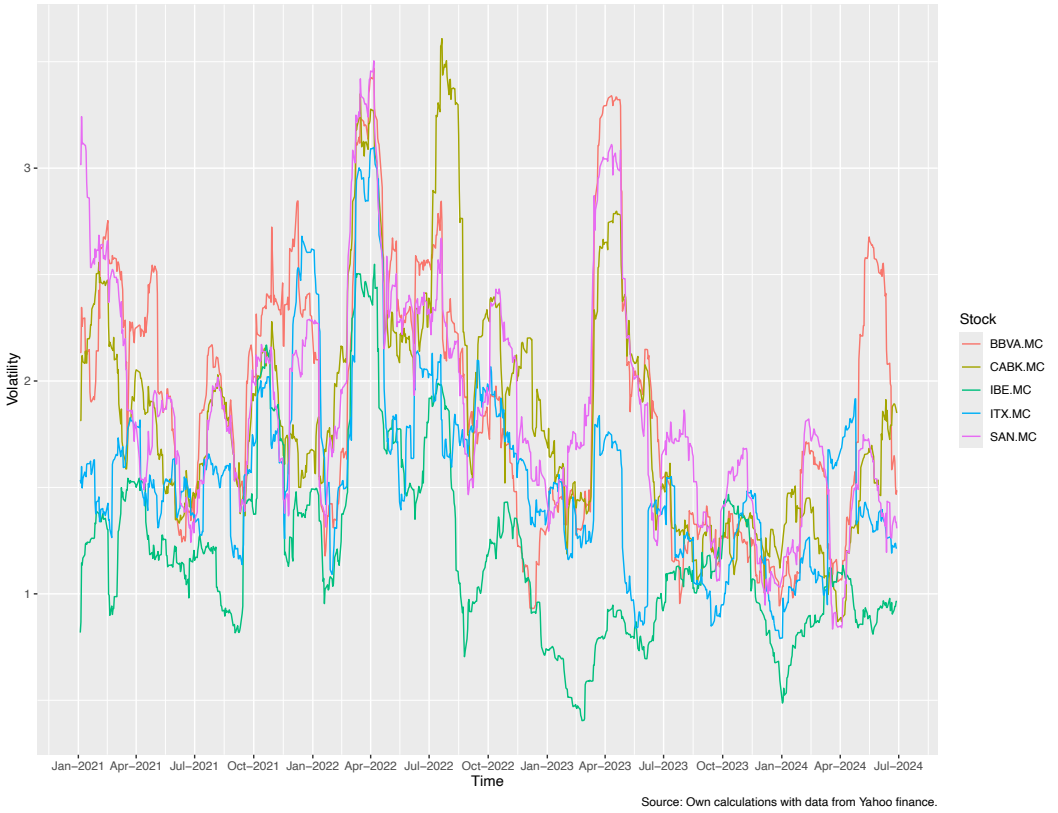


Figure 5.1: 30-day volatility of the selected stocks.

The proposed methodology was applied in a scenario in portfolio analysis in which subcompositional coherence is relevant, making proportionality a measure with desirable characteristics in this case. As a result, the stocks showed low proportionalities. However, one of the assets showed consistently higher proportionalities to the others. A deeper look into this, showed that this stock exhibits the lowest volatility for most of the period under analysis, which would be consistent with the expectation that the allocation method would assign the less volatile asset a higher

participation and assign the others using this one as benchmark.

Summarizing, although the participations of the assets show little proportionality, the index showed consistency in the different scenarios. Additionally, comparing the values of proportionalities, it is possible to understand the how the minimum variance method provides allocations, offering evidence on the advantages of using the compositional framework for this analysis. Further work is required to analyze more deeply the behavior of the proposed index when proportionalities are high, since the index has a theoretical bound, but the application proposed did not show any proportionality close to it.

6 Concluding remarks

Throughout this thesis it has been shown how compositional methods are a powerful tool in multivariate analysis in finance. The initial interest in financial stability plus the acquired one in compositional methods, led to find three applications in relevant financial topics in which the relative information is more relevant than the actual values.

Firstly, the Feldstein-Horioka puzzle was analyzed in a particular setting. The integration of the stock markets of Chile, Colombia, Mexico and Peru into the Latin American Integrated Market (MILA) would have been expected to generate flows of capitals from one market to the other, as investors would seek for more profitable opportunities after facilitating the investment. The study aimed at assessing whether a change in the composition of the MILA occurred because of the agreement. However, and in line with the Feldstein-Horioka puzzle, the compositional methods used found that there was no change in the composition of the market. This means that, despite being able to invest in any market, investors continue to prefer local assets, which is also known as the "home bias" of investment.

Secondly, compositional methods were used to create an index of concentration, applied to the Colombian banking system. The index measures the difference between the actual composition of the system and a theoretical composition in which all entities have the same participation (which would mean a complete de-concentration of the system). For the period analyzed, the Colombian banking system showed a de-concentration trend, meaning that the actual composition of the market is getting closer to the ideal one. Furthermore, time series models were used to forecast the behavior of the index. In this case, the compositional version of the models showed a good performance, compared to the traditional multivariate approach. Finally, the compositional time series model was used to predict the value of the index in the future, showing that the system is expected to continue its trend towards a lower concentration.

The third study deepened into a topic that has been key for compositional data analysis since its very foundations: alternatives to spurious correlations in compositional data. Indeed, this issue gave birth to the log-ratio variance as the main tool for measuring proportionality between parts of a composition and recent developments in literature have used it as the building block for constructing indexes to measure

such associations. These tools were used to analyze the connection between optimal allocations in a portfolio composed by assets from the IBEX35 (the Spanish stock exchange reference index), coming from the Mean-Variance portfolio (MVP) theory. Besides, by using the statistical properties of the log-ratio variance, a new index for proportionality was proposed. Further than showing consistency in terms of subcompositional coherence, the index has shown to be useful to understand the process behind the optimization of the MVP method.

As can be seen, in the three analyses it was possible to obtain interesting findings that have gained recognition among researchers from related fields. These results have also been shared with the academic community in several events, promoting fascinating debates around them. Moreover, the analyses made also have policy implications in terms of financial stability, including early warning tools for "too big to fail" institutions or frameworks to analyze the investment and risk allocation decisions of investors.

The expertise acquired in compositional methods and recent debates in financial stability through the process of this thesis opens the door to numerous analyses that can be done in the future. To start with, current work focuses on the proportionality index proposed in the third analysis, because even if the results obtained so far are relevant, there is still much to say about its properties and the interest of fellow researchers in the application of compositional methods in asset and capital allocation can lead to enriching synergies and valuable results. Furthermore, the Feldstein-Horioka puzzle analysis can be expanded to different settings, to be analyzed using compositional methods, while the recent developments in asset and capital allocation from a compositional perspective open new research possibilities in the design and evaluation of such methods.

Bibliography

Adedeji, O. and Thornton, J. (2007). Saving, investment and capital mobility in African countries. *Journal of African Economies*, 16(3):393–405. cited By 18.

Adrian, T. and Ashcraft, A. B. (2016). *Shadow banking: a review of the literature*, pages 282–315. Palgrave Macmillan UK, London.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70(1):57–65.

Aitchison, J. (1984). Reducing the dimensionality of compositional data sets. *Journal of the International Association for Mathematical Geology*, 16(6):617–635.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Monographs on statistics and applied probability. Chapman & Hall, London.

Alzoubi, M., Alsmadi, A. A., and kasasbeh, H. (2022). Systemically important bank: A bibliometric analysis for the period of 2002 to 2022. *SAGE Open*, 12(4):21582440221141259.

Arimany Serrat, N., Farreras, M. A., and Coenders, G. (2022). New developments in financial statement analysis. liquidity in the winery sector. *Accounting*, 8:355–366.

Arimany-Serrat, N., Farreras-Noguer, M. A., and Coenders, G. (2023). Financial resilience of spanish wineries during the covid-19 lockdown. *International Journal of Wine Business Research*, 35(2):346–364.

Barth, J. R. and Wihlborg, C. (2017). Too big to fail: Measures, remedies, and consequences for efficiency and stability. *Financial Markets, Institutions & Instruments*, 26(4):175–245.

Basel Committee on Banking Supervision (2013). Global systemically important banks: updated assessment methodology and the higher loss absorbency requirement. *Bank for International Settlements*.

Basel Committee on Banking Supervision (2022). The Basel Framework. *Bank for International Settlements*.

Belles-Sampera, J., Guillén, M., and Santolino, M. (2016). Compositional methods applied to capital allocation problems. *The Journal of Risk*, 19(2):1–15.

Bellod-Redondo, J. F. (1996). Ahorro e inversión en el largo plazo: El caso de la América Latina. *El trimestre económico*, 43(251):1113–1137.

Bezrodna, O., Ivanova, Z., Onyshchenko, Y., Lypchanskyi, V., and Rymar, S. (2019). Systemic risk in the banking system: Measuring and interpreting the results. *Banks and Bank Systems*, 14(3):34–47.

Bikker, J. and Spierdijk, L., editors (2019). *Handbook of competition in banking and finance*. Edward Elgar Publishing.

Boonen, T. J., Guillen, M., and Santolino, M. (2019). Forecasting compositional risk allocations. *Insurance: Mathematics and Economics*, 84:79–86.

Bouzid, S. and Kervrann, C. (2019). Handling zeros in compositional data analysis. *Computational and Mathematical Methods in Medicine*, 2019:5015172.

Carreras-Simó, M. and Coenders, G. (2021). The relationship between asset and capital structure: a compositional approach with panel vector autoregressive models. *Quantitative Finance and Economics*, 5(4):571–590.

Cetorelli, N., Hirtle, B., Morgan, D., Peristiani, S., and Santos, J. (2007). Trends in financial market concentration and their implications for market stability. *Economic Policy Review*, 13(1):33–51.

Cetorelli, N. and Traina, J. (2021). Resolving "Too Big To Fail". *Journal of Financial Services Research*, 60:1–23.

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research (1896-1977)*, 65(12):4185–4193.

Chen, S. and Dong, H. (202). Dynamic network connectedness of bitcoin markets: Evidence from realized volatility. *Frontiers in Physics*.

Coakley, J., Kulasi, F., and Smith, R. (1996). Current account solvency and the Feldstein-Horioka puzzle. *Economic Journal*, 106(436):620–627. cited By 170.

Coenders, G. and Arimany Serrat, N. (2025). Accounting statement analysis at industry level. a gentle introduction to the compositional approach.

Comas-Cufí, M., Martín-Fernández, J. A., and Mateu-Figueras, G. (2016). Log-ratio methods in mixture models for compositional data sets. *SORT-Statistics and Operations Research Transactions*, 1(2):349–374.

Dhaene, J., Tsanakas, A., Valdez, E., and Vanduffel, S. (2012). Optimal capital allocation principles. *Journal of Risk and Insurance*, 79:1 – 28.

Dong, H, e. a. (2020). The asymmetric effect of volatility spillover in global virtual financial asset markets: The case of bitcoin. *Emerging Markets Finance and Trade*.

Drakos, A., Kouretas, G., Stavroyiannis, S., and Zarangas, L. (2017). Is the Feldstein-Horioka puzzle still with us? national saving-investment dynamics and international capital mobility: A panel data analysis across EU member countries. *Journal of International Financial Markets, Institutions and Money*, 47:76–88. cited By 12.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2023). Subcompositional coherence and a novel proportionality index of parts. *SORT (Statistics and Operations Research Transactions)*, 2:229–244.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Feldstein, M. and Horioka, C. (1980). Domestic saving and international capital flows. *Economic Journal*, 90(358):314–329.

Fiori, A. M. and Porro, F. (2023). A compositional analysis of systemic risk in european financial institutions. *Annals of Finance*, 19(3):325–354.

Fiori, A. M. and Rosazza Gianin, E. (2025). Compositional risk capital allocations. *Statistical Methods & Applications*, 34(2):261–290.

Ford, N. and Horioka, C. (2017). The 'real' explanation of the Feldstein-Horioka puzzle. *Applied Economics Letters*, 24(2):95–97.

Ford, W. (2015). *Numerical Linear Algebra with Applications*. Academic Press, Boston.

Fouquau, J., Hurlin, C., and Rabaud, I. (2008). The Feldstein-Horioka puzzle: A panel smooth transition regression approach. *Economic Modelling*, 25(2):284–299. cited By 137.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*. John Wiley and Sons, New York, second ed. edition.

Glassman, D. A. and Riddick, L. A. (1996). Why empirical international portfolio models fail: evidence that model misspecification creates home asset bias. *Journal of International Money and Finance*, 15(2):275–312.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Hernandez-Romero, M. and Coenders, G. (2025). Financial resilience of agricultural and food production companies in spain: A compositional cluster analysis of the impact of the ukraine-russia war (2021-2023).

Horioka, C. Y., Terada-Hagiwara, A., and Nomoto, T. (2016). Explaining foreign holdings of asia's debt securities: The feldstein–horioka paradox revisited*. *Asian Economic Journal*, 30(1):3–24.

Ibarra-Yunez, A. (2008). Capital mobility and mexico's challenge toward financial integration. *Latin American Business Review*, 9(3-4):256–279. cited By 0.

Ioannou, S., Wójcik, D., and Dymski, G. (2019). Too-Big-To-Fail: Why Megabanks Have Not Become Smaller Since the Global Financial Crisis? *Review of Political Economy*, 31(3):356–381.

Jia, S, e. a. (2021). Asymmetric risk spillover of the international crude oil market in the perspective of crude oil dual attributes. *Frontiers in Environmental Science*.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59(6):1551–1580.

Keener, R. W. (2010). *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York.

Kynčlová, P., Filzmoser, P., and Hron, K. (2015). Modeling compositional time series with vector autoregressive models. *Journal of Forecasting*, 34:303–314.

Li, X., Tripe, D., Malone, C., and Smith, D. (2020). Measuring systemic risk contribution: The leave-one-out z-score method. *Finance Research Letters*, 36 (C).

Linares-Mustarós, S., Coenders, G., and Vives-Mestres, M. (2018). Financial performance and distress profiles. from classification according to financial ratios to compositional classification. *Advances in Accounting*, 40:1–10.

Linares-Mustarós, S., Farreras-Noguer, M. A., Arimany-Serrat, N., and Coenders, G. (2022). New financial ratios based on the compositional data methodology. *Axioms*, 11(12).

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., and Bähler, J. (2015). Proportionality: A valid alternative to correlation for relative data. *PLOS Computational Biology*, 11(3):1–12.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag Berlin Heidelberg.

Magrini, A. (2025). Bankruptcy risk prediction: A new approach based on compositional analysis of financial statements. *Big Data Research*, 41:100537.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Markowitz, H. M. (1991). Foundations of portfolio theory. *The Journal of Finance*, 46(2):469–477.

Martín-Fernández, J. A., Bren, M., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. (ND). A measure of difference for compositional data based on measures of divergence. Available at: https://ima.udg.edu/~barcelo/index_archivos/A_mesure_of_difference.pdf. Consulted on September 19, 2025.

Martín-Fernández, J. A. and Thió-Henestrosa, S. (2013). *Compositional Data Analysis in Practice*. Springer Series in Statistics.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley.

Mills, T. (2010). Forecasting compositional time series. *Quality & Quantity: International Journal of Methodology*, 44(4):673–690.

Mishkin, F. S., Stern, G., and Feldma, R. (2006). How big a problem is too big to fail? a review of Gary Stern and Ron Feldman's "too big to fail: The hazards of bank bailouts". *Journal of Economic Literature*, 44(4):988–1004.

Moch, N. (2018). The contribution of large banking institutions to systemic risk: What do we know? a literature review. *Review of Economics*, 69(3):231–257.

Molas-Colomer, X., Linares-Mustaros, S., ÀNGELS, F.-N., and CARLES FERRER-COMALAT, J. (2024). A new methodological proposal for classifying firms according to the similarity of their financial structures based on combining

compositional data with fuzzy clustering. *Journal of Multiple-Valued Logic & Soft Computing*, 43.

Narayan, P. (2005). The saving and investment nexus for China: Evidence from cointegration tests. *Applied Economics*, 37(17):1979–1990. cited By 1532.

Olkin, I. and Sampson, A. (2001). Multivariate analysis: Overview. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 10240–10247. Pergamon, Oxford.

Omarova, S. T. (2019). The "Too Big To Fail" Problem. *Minnesota Law Review*, 103:2495–2541.

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2011). Lecture notes on compositional data analysis. *University of Girona*.

Pearson, K. (1897). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60(359-367):489–498.

Pfaff, B. (2008). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, 27(4).

Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607.

Shull, B. (2010). Too big to fail in financial crisis: Motives, countermeasures, and prospects. *Levy Economics Institute Working Paper No. 601*.

Simone, G. (2014). Dealing with zeros in compositional data. *Mathematical Geosciences*, 46(4):411–424.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica*, 58(1):113–144.

Sinha, D. and Sinha, T. (1998). An exploration of the long-run relationship between saving and investment in the developing economies: A tale of latin american countries. *Journal of Post Keynesian Economics*, 20(3):435 – 443.

Sorkin, A. R. (2010). *Too Big to Fail: The Inside Story of How Wall Street and Washington Fought to Save the Financial System–and Themselves*. Penguin Books.

Stern, G. and Feldman, R. (2004). *Too big to fail: The hazards of bank bailouts*. Washington, D.C.: Brookings Institution Press.

Templ, M., Hron, K., and Filzmoser, P. (2011). Sparse log-ratio analysis for compositional data. *Mathematics*, 3(1):219–235.

Thomas, P. and Lovell, D. (2014). Compositional data analysis (CoDA) approaches to distance in information retrieval. In Geva, S., Trotman, A., Bruza, P., Clarke, C. L. A., and Järvelin, K., editors, *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 991–994. ACM.

van den Boogaart, K. and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Use R! Springer Berlin Heidelberg.

Vega Baquero, J. D. and Santolino, M. (2022a). Capital flows in integrated capital markets: Mila case. *Quantitative Finance and Economics*, 6(4):622–639.

Vega Baquero, J. D. and Santolino, M. (2022b). Too big to fail? An analysis of the Colombian banking system through compositional data. *Latin American Journal of Central Banking*, 3(2):100060.

Vega Baquero, J. D. and Santolino, M. (2025). Proportionality between allocations in asset management. *IREA Working paper, in progress*.

Verbelen, R., Antonio, K., and Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1275–1304.

Vieira, I. (2003). Evaluating capital mobility in the EU: A new approach using swaps data. *European Journal of Finance*, 9(5):514–532. cited By 3.

Vives, X. (2016). *Competition and stability in banking: The role of regulation and competition policy*. Princeton University Press.

Zheng, T. and Chen, R. (2017). Dirichlet arma models for compositional time series. *Journal of Multivariate Analysis*, 158:31–46.

Zhou, C. (2010). Are banks too big to fail? measuring systemic importance of financial institutions. *International Journal of Central Banking*, 6(4):205–250.

Dades composicionals per a l'anàlisi en economia i finances
Resum

A molts problemes multivariants a finances i economia els valors relatius de les variables són més rellevants que els seus valors absoluts, sent aquesta la base de l'anàlisi de dades composicionals. Aquesta tesi té com a objectiu contribuir a la integració del marc composicional en un àmbit rellevant d'anàlisi en finances: l'estabilitat financera. La tesi s'organitza en tres parts diferents, cadascuna centrada en una anàlisi financera concreta.

La primera es refereix al puzle Feldstein-Horioka (F-H), que afirma que la liberalització dels mercats de capitals no condueix necessàriament a un moviment de capital en busca d'una millor assignació de recursos, com suggereix la teoria clàssica. En els últims anys, Xile, Colòmbia, Mèxic i Perú s'han incorporat al Mercat Integrat Llatinoamericà a través d'un acord que permet als inversors de qualsevol dels mercats participants invertir en els altres. Mètodes composicionals, tant transversals com de sèries temporals, es van utilitzar per a avaluar si la creació del mercat conjunt va conduir a un flux de capital entre mercats. Com a resultat, no va ser possible rebutjar la hipòtesi de F-H, recolzant la idea que la liberalització dels mercats de capitals no és suficient per generar fluxos de capital entre mercats.

En segon lloc, es crea un índex de concentració de sistemes financers i bancaris mitjançant mètodes composicionals per establir l'existència potencial d'entitats financeres "Too big to fail" (massa grans per fer fallida), proporcionant així als reguladors d'una eina d'alerta davant d'aquest tipus d'institucions. L'índex es va aplicar al sistema bancari colombià i es va monitoritzar en el temps per a avaluar si el sistema financer estava cada vegada més concentrat. Els resultats van trobar que l'índex de concentració anava disminuint i que aquesta tendència continuaria en el futur. Des del punt de vista metodològic, els models composicionals van mostrar ser més estables i amb una millor capacitat predictiva en comparació amb les metodologies clàssiques.

Finalment, les relacions entre actius d'una cartera s'avaluen des d'una perspectiva composicional. Durant molts anys, el tema de les correlacions espúries entre variables expressades en termes relatius, com les dades composicionals, ha rebut un gran interès. Com alternativa a la correlació, la tesi proposa un índex de proporcionalitat per les parts d'una composició que es basa en la variància dels log-ràtios, una mesura àmpliament utilitzada per analitzar la proporcionalitat. L'índex es va calcular per a una cartera hipotètica d'accions del mercat borsari espanyol per a avaluar les connexions entre les assignacions generades pel mètode Mean-Variance. L'índex proporciona informació rellevant sobre com es generen les assignacions òptimes amb aquest mètode.

Classificatió JEL: C01, E22, F32, G11, G17, G21, G28.
Paraules clau: Geometria d'Aitchison, puzle Feldstein-Horioka, bancs amb importància sistèmica, concentració bancària, cartera Mean-Variance, proporcionalitat.