# "Too big to fail? An analysis of the Colombian banking system through compositional data."

Juan David Vega Baquero and Miguel Santolino

UBIREA | Institut de Recerca en Economia Aplicada Regional i Pública

UBIREA | Institut de Recerca en Economia Aplicada Regional i Pública
UNIVERSITAT DE BARCELONA

WEBSITE: www.ub.edu/irea/ • CONTACT: irea@ub.edu

The Research Institute of Applied Economics (IREA) in Barcelona was founded in 2005, as a research institute in applied economics. Three consolidated research groups make up the institute: AQR, RISK and GiM, and a large number of members are involved in the Institute. IREA focuses on four priority lines of investigation: (i) the quantitative study of regional and urban economic activity and analysis of regional and local economic policies, (ii) study of public economic activity in markets, particularly in the fields of empirical evaluation of privatization, the regulation and competition in the markets of public services using state of industrial economy, (iii) risk analysis in finance and insurance, and (iv) the development of micro and macro econometrics applied for the analysis of economic activity, particularly for quantitative evaluation of public policies.

## *Abstract*

Although still incipient in economics and finance, compositional data analysis (in which relative information is more important than absolute values) has become more relevant in statistical analysis in recent years. This article constructs a concentration index for financial/banking systems by means of 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 kind of institutions. The index has been applied to the Colombian banking system and assessed over time with a forecast to determine whether the system is becoming more concentrated or not. It was found that the concentration index has been decreasing in recent years and the model predicts that this trend will continue. In terms of the methodology used, compositional models were shown to be more stable and to lead to better prediction of the index than the classical multivariate methodologies.

***JEL classification:*** G17, G21, G28.

***Keywords:*** Simplex, Aitchison geometry, Systematically important banks, Vector autoregresion.

Juan David Vega Baquero: PhD student, University of Barcelona. Email: jvegabaq38@alumnes.ub.edu

Miguel Santolino (corresponding author): Riskcenter, Department of Econometrics, University of Barcelona, Diagonal 690, 08034-Barcelona, Spain. Tel.:+34-93-4020484; fax: +34-93-4021983. Email: msantolino@ub.edu

## *Acknowledgements*

This study was supported by the Spanish Ministry of Science and Innovation under grant PID2019-105986GB-C21.

# Too big to fail? An analysis of the Colombian banking system through compositional data.

Juan David Vega Baquero and Miguel Santolino*

University of Barcelona

April 28, 2021

**Abstract**

Although still incipient in economics and finance, compositional data analysis (in which relative information is more important than absolute values) has become more relevant in statistical analysis in recent years. This article constructs a concentration index for financial/banking systems by means of 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 kind of institutions. The index has been applied to the Colombian banking system and assessed over time with a forecast to determine whether the system is becoming more concentrated or not. It was found that the concentration index has been decreasing in recent years and the model predicts that this trend will continue. In terms of the methodology used, compositional models were shown to be more stable and to lead to better prediction of the index than the classical multivariate methodologies.

**JEL classification:** G17, G21, G28.
**Keywords:** simplex, aitchison geometry, systematically important banks, vector autoregresion.

## 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). Nevertheless, some authors did not consider that the concept of TBTF 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[1].

In recent years, the definition of TBTF institutions has changed to include not only the size but also other characteristics such as interconnectedness, substitutability, cross-national activity and

*Corresponding author. Riskcenter, Department of Econometrics, University of Barcelona, Diagonal 690, 08034-Barcelona, Spain. Tel.:+34-93-4020484; fax: +34-93-4021983; e-mail: *msantolino@ub.edu*

[1]Some literature on TBTF institutions after the crisis include Shull (2010) and Zhou (2010). An anecdotal explanation can be found in Sorkin (2010).

complexity of financial institutions to define them as systemically important banks - SIBs (Basel Committee on Banking Supervision, 2013). An interesting review of recent literature about SIBs was carried out by Moch (2018). Among the most important conclusions, the author notes that there is consensus on the fact that the relationship between the size of banks and systemic risk is 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 may lead to an amplified effect of size on systemic risk. A recent attempt to measure the impact of SIBs on systemic risk was developed by Li et al. (2020). They tried to determine the importance of an institution in the system by means of changes in the systemic risk measured with and without it (a leave-one-out method), using z-scores as the measure of risk. Another approach can be found in Bezrodna et al. (2019), who proposed a risk indicator that considers the importance of banks in terms of the size of their assets with respect to the total assets held by the banking system and connects this with the riskiness of the entity to determine whether a SIB needs greater supervision than others. This approach relies on the fact that highly concentrated, interconnected systems are more vulnerable to distress in one SIB.

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 article proposes a novel approach to evaluating concentration in financial systems by constructing a concentration index based on the relative size of each entity in terms of assets. Then, the main assumption of this article is that the importance of a financial entity on a system depends on its relative size, i.e., 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 degree of concentration of the system. The methodology presented uses the compositional data framework to estimate an indicator of concentration that can be tracked over time, to gain insight into the current and expected concentration of a defined system. In the compositional data framework, relative information is more relevant than absolute values. 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 article, the variable of interest is a bank's relative weight with respect to others, in this case in terms of assets. Therefore, the value of assets held by each bank 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.

This methodology will be 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 show the benefits of the use of compositional data framework in different fields.

The rest of the article is divided into four sections. The next section describes the methodology, emphasizing the compositional data framework and its application. Section three explores the Colombian banking system and the dataset used in the analysis, which is included in section four, together with the comparison and assessment of the proposed model. Finally, section five summarizes the findings and concludes.

# 2 Methodology

This section explains the methodology used for the analysis, starting with a brief explanation of the compositional data framework and the centered log-ratio (clr) and isometric log-ratio (ilr) transformations required to use regular statistical methods in compositional data. This is followed by the definition of the econometric models to be used. Finally, the criterion for the model selection and assessment of results are explained.

## 2.1 Compositional data and Aitchison geometry

A composition is defined as a row vector $\overrightarrow{x} = [x_1, ..., x_n]$ whose components only carry relative information (Pawlowsky-Glahn et al., 2011), which usually adds up to a constant $\kappa$. For simplicity, it is usually assumed that $\kappa = 1$ without loss of generality. Therefore, the sample space in which compositional data is defined as:

$$\mathcal{S}^n = \{\overrightarrow{x} \in \mathbb{R} | x_i \geq 0, i = 1, ..., n, \sum_{i=1}^{n} x_i = 1\}$$

For simplicity, throughout the article it will be assumed that $x_i > 0$, as the log transformations to be applied later do not allow for zeros. For further discussion of the case in which the composition has zeros, please refer to Aitchison (1986).

Therefore, the methods for analyzing the data are those defined for the algebraic-geometric structure of the simplex, which are commonly referred to as Aitchison geometry. Hence, following Aitchison (1986), the perturbation operation $\oplus$ within the simplex is defined as $\overrightarrow{x} \oplus \overrightarrow{y} = \left(\frac{x_1 \cdot y_1}{\sum_{i=1}^{n} x_i \cdot y_i}, ..., \frac{x_n \cdot y_n}{\sum_{i=1}^{n} x_i \cdot y_i}\right)$ for $\overrightarrow{x}, \overrightarrow{y} \in \mathcal{S}^n$. Moreover, the powering operation $\odot$ is defined as

$\lambda \odot \overrightarrow{x} = \left( \frac{x_1^\lambda}{\sum_{i=1}^n x_i^\lambda}, ..., \frac{x_n^\lambda}{\sum_{i=1}^n x_i^\lambda} \right)$ for $\lambda \in \mathbb{R}$.

With this in mind, Aitchison (1986) defines the compositional mean for $M$ compositions as $AM_\Delta(\overrightarrow{x}_1, ..., \overrightarrow{x}_M) = \frac{1}{M} \odot \bigoplus_{m=1}^M \overrightarrow{x}_m$. Additionally, according to Aitchison geometry the distance between two compositions $\overrightarrow{x}$ and $\overrightarrow{y}$ is defined by Equation 1:

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

where $\| \cdot \|_\Delta$ is the norm and $\overrightarrow{x} \ominus \overrightarrow{y} = \overrightarrow{x} \oplus [(-1) \odot \overrightarrow{y}]$.

## 2.2 The clr and ilr transformations

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

$$clr(\overrightarrow{x}) = \left[ \ln \frac{x_1}{g(\overrightarrow{x})}, ..., \ln \frac{x_i}{g(\overrightarrow{x})} \right] \text{ with } g(\overrightarrow{x}) = \left( \prod_{i=1}^n x_i \right)^{1/n}$$

where $g(\overrightarrow{x})$ is the component-wise geometric mean of the composition. As can be seen, since the clr is a component-wise transformation, it follows that $clr(\overrightarrow{x}) \in \mathbb{R}^n$. However, by definition, the elements of $clr(\overrightarrow{x})$ add up to zero, which imposes a new restriction[2].

To apply the usual econometric models, $\overrightarrow{x}$ 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 $clr(e_j)$. Ttherefore, $V$ 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:

$$ilr(\overrightarrow{x}) = clr(\overrightarrow{x}) \cdot V$$

[2]This implies that the covariance matrix of $clr(\overrightarrow{x})$ is singular, i.e. the determinant is zero (Pawlowsky-Glahn et al., 2011).

Another advantage of the ilr transformation is that it fulfills $\Delta(\overrightarrow{x}, \overrightarrow{y}) = d(ilr(\overrightarrow{x}), ilr(\overrightarrow{y}))$, 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 that proposed by Egozcue et al. (2003) through binary partitions. This method has the advantage of interpretability, as 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, a balance is generated between elements of one group and 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 $\overrightarrow{x}$ are now represented by coordinates in $\mathbb{R}^{n-1}$, which allows for the use of conventional statistical methods for multivariate analysis.

### 2.3 Compositional VAR model

Now that the compositional data has been transformed to use regular multivariate statistical methods, the model must be established to be used to analyze the Colombian banking system. In this case, the modeling strategy follows the approach in Boonen et al. (2019) who analyzed the market risk in a stock portfolio.

The vector $\overrightarrow{a}_t = [a_{1,t}, ..., a_{n,t}]$ contains the assets held by bank $i \in \{1, ..., n\}$ at time $t$. The first approach would be to analyze this series from the multivariate statistics framework. However, the analysis can be expanded through compositional data. To achieve this, the total assets held by the banking system at time $t$ as $A_t = \sum_{i=1}^{n} a_{i,t}$ must be defined. Now, $\overrightarrow{a}_t$ can be transformed into compositional data by defining the vector $\overrightarrow{x}_t = [x_{1,t}, ..., x_{n,t}]$ whose elements will be $x_{i,t} = a_{i,t}/A_t$, meaning $\overrightarrow{x}_t$ is the composition across all entities of the assets held by the banking system at time $t$.

Therefore, the model to be estimated corresponds to a vector autoregressive (VAR) model. In this case, the general form of the model is given by:

$$ilr(\overrightarrow{x}_t) = b + B_1 \cdot ilr(\overrightarrow{x}_{t-1}) + ... + B_p \cdot ilr(\overrightarrow{x}_{t-p}) + \epsilon_t \tag{2}$$

where $b$ is a vector of parameters in $\mathbb{R}^{n-1}$, $B_1, ..., B_p$ are each $n-1 \times n-1$ matrices of parameters in $\mathbb{R}$, $\epsilon$ is the error term that follows a multivariate normal distribution with a vector of means equal to zero and covariance matrix $\Sigma_\epsilon$, $N_{n-1}(0, \Sigma_\epsilon)$, and $p$ is the number of lags to be considered in the model. The model in (2) can be extended by adding the first lag of the total value of the assets held by banks in logarithms $\log(A_{t-1})$ as a control variable. Thus, the extended model is:

$$ilr(\overrightarrow{x}_t) = b + B_1 \cdot ilr(\overrightarrow{x}_{t-1}) + ... + B_p \cdot ilr(\overrightarrow{x}_{t-p}) + \gamma \cdot \log(A_{t-1}) + \epsilon_t \tag{3}$$

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

These models are compared to the classical approach for multivariate series. As mentioned previously, the series of assets held by banks $\overrightarrow{a}_t$ can be analyzed directly through the multivariate VAR:

$$\overrightarrow{a}_t = c + C_1 \cdot \overrightarrow{a}_{t-1} + ... + C_p \cdot \overrightarrow{a}_{t-p} + \varepsilon_t \tag{4}$$

where $c$ will be a vector of parameters in $\mathbb{R}^n$, $C_1, ..., C_p$ will be $n \times n$ matrices of parameters defined in the real space and $\varepsilon_t \sim N_n(0, \Sigma_\varepsilon)$ is the error term. This model can also be extended by including $\log(A_{t-1})$:

$$\overrightarrow{a}_t = c + C_1 \cdot \overrightarrow{a}_{t-1} + ... + C_p \cdot \overrightarrow{a}_{t-p} + \varphi \cdot \log(A_{t-1}) + \varepsilon_t \tag{5}$$

where $\varphi$ will be the $n \times 1$ vector of coefficients associated with the control variable.

To define the number of lags to be used in every model, the Akaike's information criterion (AIC) is used. This is a likelihood criterion for measuring the fitness of a statistical model to a sample. It is defined as $AIC = -2\ell + 2K$ where $\ell$ is the log-likelihood of the estimation and $K$ corresponds to the number of parameters in the model. As can be seen, 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.

### 2.4 Stationarity

Even though the VAR model can be estimated in the absence of stationarity, as the estimators exist and are consistent (Sims et al., 1990), it should be considered that this is one of the main assumptions to determine the distribution of the estimators and be able to make inference. Therefore, the stationarity of the dataset is assessed to determine whether a VAR model is suitable and under which conditions.

In its strictest sense, stationarity refers to the property of a time series of having a time invariant distribution. This means that each realization of the variable follows the same distribution, independently of the moment $t$ (Fuller, 1996). However, in limited samples, this characteristic is hard to prove as the distribution of each realization is unknown. Therefore, weak stationarity is

usually a sufficient condition. In this case, only the first two moments are required to be constant over time (i.e. independent of $t$).

In the absence of 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 differentiated to obtain stationarity. Then, a VAR model is estimated with the new variables. This approach has a limitation in terms of interpretability, since 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. For the case of compositional data, the generic model would be given by:

$$\Delta ilr(\overrightarrow{x}_t) = d + D_1 \cdot \Delta ilr(\overrightarrow{x}_{t-1}) + ... + D_p \cdot \Delta ilr(\overrightarrow{x}_{t-p}) + \epsilon_t \tag{6}$$

where $\Delta$ is the difference operator defined as $\Delta ilr(\overrightarrow{x}_t) = ilr(\overrightarrow{x}_t) - ilr(\overrightarrow{x}_{t-1})$, $d$ is a vector of parameters in $\mathbb{R}^{n-1}$ and $D_1, ..., D_p$ are $n-1 \times n-1$ matrices of parameters in $\mathbb{R}$. Similarly the model can be estimated for the assets held by banks:

$$\Delta \overrightarrow{a}_t = f + F_1 \cdot \Delta \overrightarrow{a}_{t-1} + ... + F_p \cdot \Delta \overrightarrow{a}_{t-p} + \varepsilon_t \tag{7}$$

where $f$ is a vector of parameters in $\mathbb{R}^n$ and $F_1, ..., F_p$ will be $n \times n$ matrices of parameters in the real space.

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 said to be 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. For the compositional data, the model would be:

$$\Delta ilr(\overrightarrow{x}_t) = g + G \cdot ilr(\overrightarrow{x}_{t-1}) + B_1^* \cdot \Delta ilr(\overrightarrow{x}_{t-1}) + ... + B_{p-1}^* \cdot \Delta ilr(\overrightarrow{x}_{t-(p-1)}) + \epsilon_t \tag{8}$$

where $g$ is a vector of parameters in $\mathbb{R}^{n-1}$, $G = \sum_{s=1}^{p} B_s - I$ and $B_r^* = -\sum_{s=r+1}^{p} B_s$ ($B_s$ corresponds to the matrix of coefficients of the VAR model in Equation 2, meaning both $G$ and $B_r^*$ are of size $n-1 \times n-1$). Furthermore, $G = \alpha \cdot \beta'$ can be defined, where $\alpha$ will denote the size of the cointegration effects and $\beta$ the cointegration matrix leading to the stationarity of the series. The dimensions of both $\alpha$ and $\beta$ will be determined by the rank of $\beta$, which is explained below. Consequently, the VEC model can be interpreted as composed by the long-term dynamics (contained in $G$) and the short-term effects (modeled through $B_r^*$). Similarly, for the series of assets

held by banking institutions, the VEC model is defined by:

$$\Delta \overrightarrow{a}_t = h + H \cdot \overrightarrow{a}_{t-1} + C_1^* \cdot \Delta \overrightarrow{a}_{t-1} + ... + C_{p-1}^* \cdot \Delta \overrightarrow{a}_{t-(p-1)} + \varepsilon_t \tag{9}$$

where $h$ is a vector of parameters in $\mathbb{R}^n$, $H = \sum_{s=1}^{p} C_s - I$ and $C_r^* = -\sum_{s=r+1}^{p} C_s$.

Nevertheless, to estimate a VEC model, the order of integration of the variables must be known, meaning how many times is it necessary to differentiate each variable to obtain a stationary series, and whether the variables are cointegrated, i.e. 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 $\beta$.

Therefore, a diagnosis of the dataset will be diagnosed to test for the assumptions. With the information obtained, the models that suit the characteristics of the data will be estimated.

### 2.5 Model comparison

When the estimated models have been obtained, the next step is to compare them. In this case, the AIC cannot be used since the models to be compared now have different variables. Therefore, this comparison is achieved through the accuracy of the forecasts made by the models (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 $k \in \{T-h, ..., T-1\}$, the model can be estimated with the information up to $k$. Next, with the estimated model, forecast periods $\{k+1, ..., T\}$ and the forecasted values, the deviation from the observed values can be calculated through the mean Aitchison distance of prediction errors (MADPE):

$$MADPE(k, m) = \frac{1}{T-k} \sum_{t=k+1}^{T} AD_{\Delta}(\overrightarrow{x_t}, \hat{\overrightarrow{x_t}}^{k,m}) \tag{10}$$

where $AD_{\Delta}(\cdot, \cdot)$ is the Aitchison distance defined in Equation 1 and $\hat{\overrightarrow{x_t}}^{k,m}$ is the forecast for the period $t$ with the model $m$ estimated with information up to $k$. Again, as the MADPE measures the accuracy of the forecasts through the distance between the observed and predicted values, the model with the lowest MADPE will be best for forecasting.

# 3 Colombian financial system data

The Financial Superintendence of Colombia[3] (financial regulator) defines as credit establishments 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, that 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; 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 defined by the legal act creating each entity.

The aim of this study is to analyze the risk of concentration of the financial system with a low number of institutions managing most of the assets. Notably, special official institutions are funded (exclusively or in a high proportion) by the government. In some cases, they are created with the purpose stabilizing the system through liquidity in situations of disruption. Therefore, including these entities in the analysis could bias the results.

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 indicator of size. The total assets in credit establishments have been growing in recent decades with considerable participation of banking establishments, which had an average participation of nearly 96% during the last decade. The next type of entity is financial corporations with 2%, followed by commercial financing companies with 1%, and finally financial cooperatives with less than 1%. Therefore, the composition of credit establishments shows very high relevance of banking establishments, while the rest are irrelevant in terms of 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 is different in some sense from that of banking establishments. Consequently, the dataset to be used for the analysis will only contain banking establishments.

## 3.1 Colombian banking system

The dataset considered is from January 2010 to April 2020, which provides a base availability of 124 observations for each entity. The series consists of a total 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 impact on the outcome is negligible, and they have been excluded to avoid issues with zeros in the dataset.

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

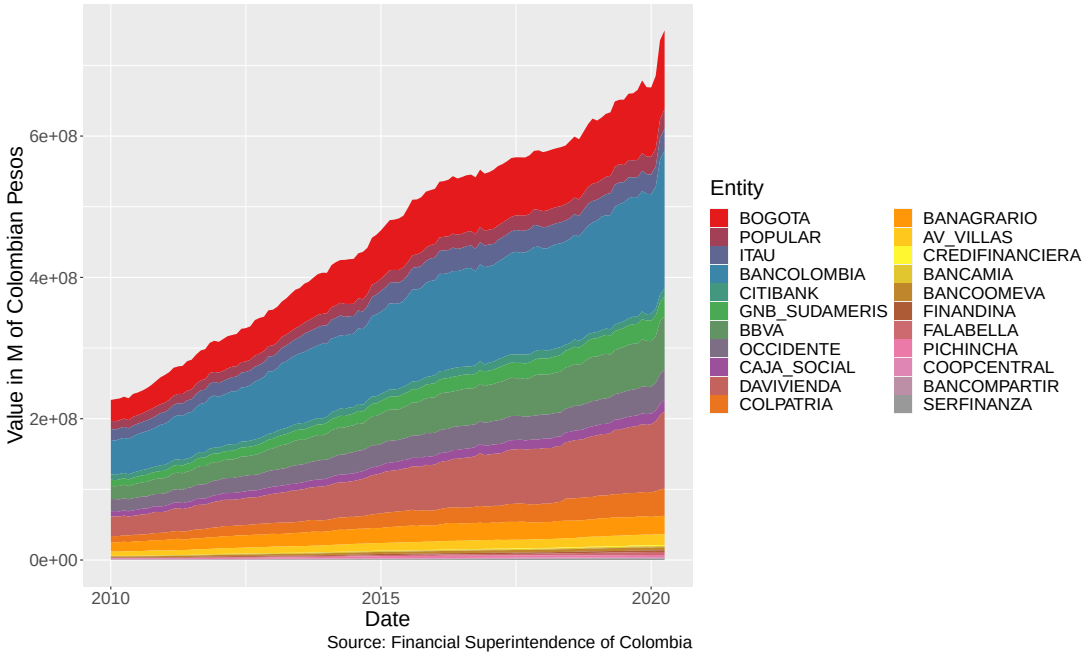


Figure 1: Total assets per banking establishment in Colombia from January 2010 to April 2020

Hence, the dataset to be used in the analysis consists of 22 banks and is summarized in Figure 1, which shows the total assets held by each banking establishment. Figure 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.

### 3.2 Concentration level

To analyze if there is 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): $\overrightarrow{z} = [z_1, ..., z_n]$ with $z_i = 1/n$. 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 relative assets of the Colombian banking system is computed. The calculation of the monthly distances is shown in Figure 3. As can be observed, the distance has been decreasing throughout the period analyzed, which would be a sign of 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 behavior of the asset composition in the banking system, to assess the potential concentration of the market and use the

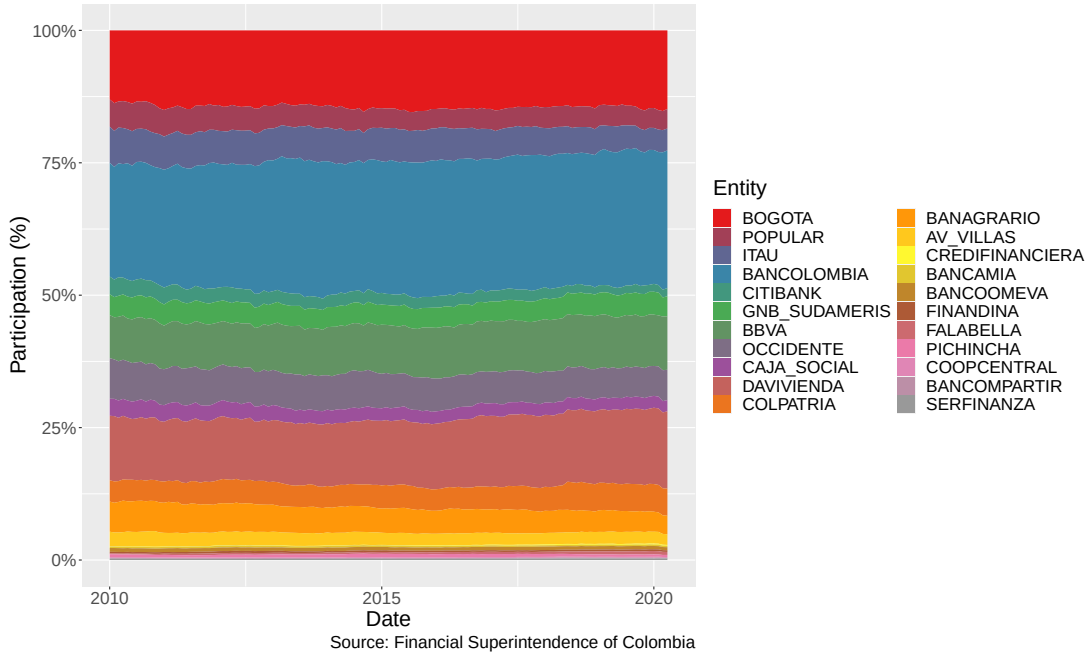


Figure 2: Assets composition per banking establishment in Colombia from January 2010 to April 2020

outcome as an input for decision making on financial regulation, by creating an alert on potential too big to fail institutions.

### 3.3 Data diagnosis

After defining the dataset 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. To test for stationarity, unit root tests are commonly used. These tests rely on the fact that, for a stationary series, all the roots of the characteristic equation lie inside the unit circle. If the process has at least one root equal to one, then the process is not stationary. A common test for stationarity is the augmented Dickey-Fuller test (Said and Dickey, 1984). 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 $\overrightarrow{a}$ and $\overrightarrow{x}$ (i.e. the values and compositions of assets in the Colombian banking system, respectively). For the series of the value of assets held by banking institutions, the results show that there is no stationarity for 21 out of 22 entities at 5.0% of significance. Similarly, for the series of compositions there is no stationarity in 19 out of 21 variables. Nevertheless, both sets are found to be integrated of degree one, meaning that the first differences of each individual series of both $\overrightarrow{a}$ and $\overrightarrow{x}$ are stationary.

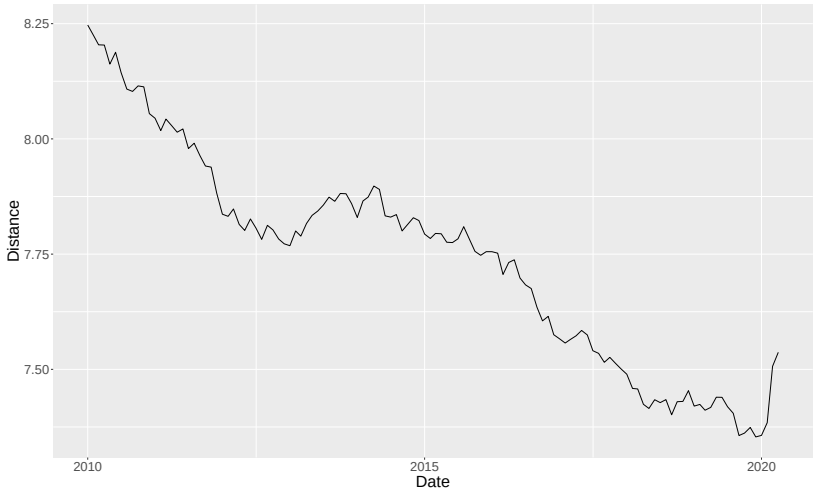


Figure 3: Observed concentration level index of the Colombian banking system: Aitchison distance between the actual composition of assets and the hypothetical composition in which assets are equally distributed among all entities

Since 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 dataset (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 high dimension of the dataset (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, it can be assumed that there is cointegration in the series, although it is not possible to know the rank of the cointegration matrix.

The diagnosis showed that the dataset is non-stationary and cointegrated, which allows to estimate 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 approach 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 to be compared using the MADPE described previously, to determine which of the models performs better in terms of forecasting.

| Model | N. of lags | AIC | |
|---|---|---|---|
| | | Compositional model | Model with value of assets |
| VAR | 1 | -12993.35 | 72096.36 |
| | 2 | -12993.35 | 71487.74 |
| | 3 | -12836.35 | 70434.82 |
| Extended VAR | 1 | -13008.18 | 72051.00 |
| | 2 | -13008.18 | 71450.68 |
| | 3 | -12885.30 | 70360.87 |
| VAR in differences | 1 | -12461.58 | 71923.12 |
| | 2 | -12461.58 | 71923.12 |
| | 3 | -12461.58 | 70442.79 |

Table 1: AIC results for different models and number of lags

# 4 Results

Now that the methodology and the dataset 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 by means of the AIC. Then, the four proposed models will be assessed through the mean Aitchison distance of prediction errors (MADPE). Afterwards, the specification assumptions of the models 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.

## 4.1 Model selection

The results for the AIC selection criteria for the models are shown in Table 1. As can be seen, for the proposed specifications using compositional data, one lag should be used for the VAR models[4]. For the models using the value of the assets held by banks, three lags should be used in the VAR models.

Therefore, the initial VAR models from Equations 2 and 4 will correspond to:

$$ilr(\overrightarrow{x}_t) = b + B \cdot ilr(\overrightarrow{x}_{t-1}) + \epsilon_t \tag{11}$$

$$\overrightarrow{a}_t = c + C_1 \cdot \overrightarrow{a}_{t-1} + C_2 \cdot \overrightarrow{a}_{t-2} + C_3 \cdot \overrightarrow{a}_{t-3} + \varepsilon_t \tag{12}$$

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

Similarly, the extended models to be estimated from general Equations 3 and 5 are:

$$ilr(\overrightarrow{x}_t) = b + B_1 \cdot ilr(\overrightarrow{x}_{t-1}) + \gamma \cdot \log(A_{t-1}) + \epsilon_t \tag{13}$$

$$\overrightarrow{a}_t = c + C_1 \cdot \overrightarrow{a}_{t-1} + C_2 \cdot \overrightarrow{a}_{t-2} + C_3 \cdot \overrightarrow{a}_{t-3} + \varphi \cdot \log(A_{t-1}) + \varepsilon_t \tag{14}$$

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

$$\Delta ilr(\overrightarrow{x}_t) = d + D \cdot \Delta ilr(\overrightarrow{x}_{t-1}) + \epsilon_t \tag{15}$$

$$\Delta \overrightarrow{a}_t = f + F_1 \cdot \Delta \overrightarrow{a}_{t-1} + F_2 \cdot \Delta \overrightarrow{a}_{t-2} + F_3 \cdot \Delta \overrightarrow{a}_{t-3} + \varepsilon_t \tag{16}$$

For the VEC models, the number of lags is selected automatically by the program during the estimation[5] and is set to two in both cases. Therefore, the models to be estimated will be:

$$\Delta ilr(\overrightarrow{x}_t) = g + G \cdot ilr(\overrightarrow{x}_{t-1}) + B_1^* \cdot \Delta ilr(\overrightarrow{x}_{t-1}) + B_2^* \cdot \Delta ilr(\overrightarrow{x}_{t-2}) + \epsilon_t \tag{17}$$

$$\Delta \overrightarrow{a}_t = h + H_1 \cdot \overrightarrow{a}_{t-1} + C_1^* \cdot \Delta \overrightarrow{a}_{t-1} + C_2^* \cdot \Delta \overrightarrow{a}_{t-2} + \varepsilon_t \tag{18}$$

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. three years).

Figure 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 in terms of 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 the time.

[5]VEC models were estimated by the package *vars* in R.

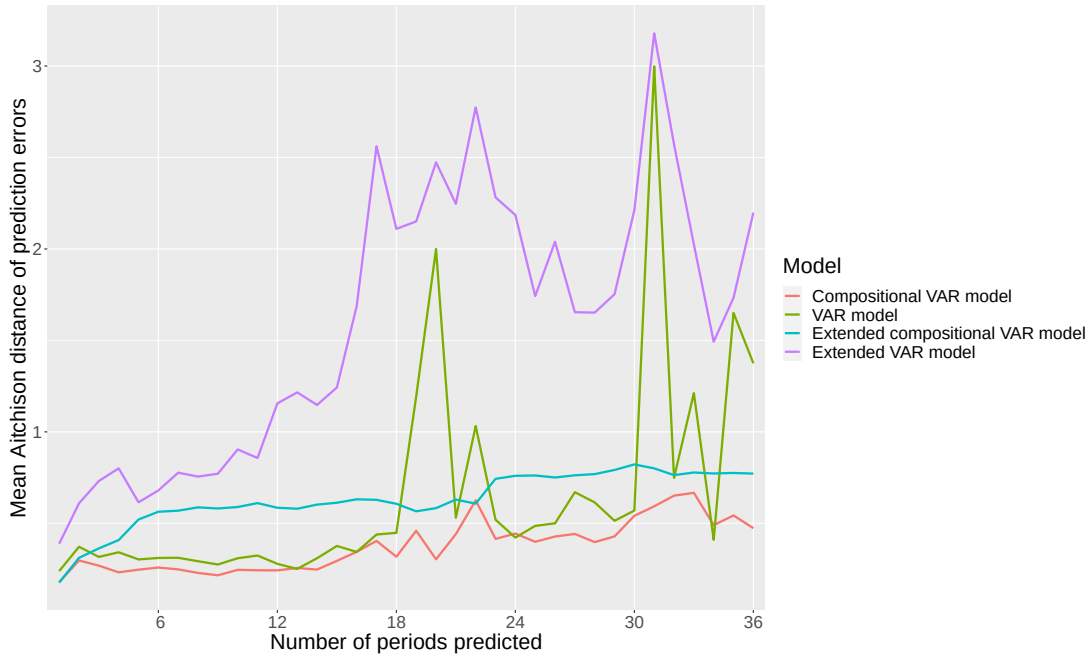


Figure 4: Mean Aitchison distance of prediction errors (MADPE): Compositional VAR model, VAR model, extended compositional VAR model and extended VAR model

Similarly, Figure 5 includes models in differences. As can be 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 than the basic compositional VAR. However, the difference is very small, and it is not possible to conclude whether this difference is significant or not, since the confidence intervals are not available.

The VEC models compared in Figure 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 the fact 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 by 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, like reduced manipulation of the dataset 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 $T-1$. This reduces the degrees of freedom of the estimation, while VEC estimations are even more complex, and the

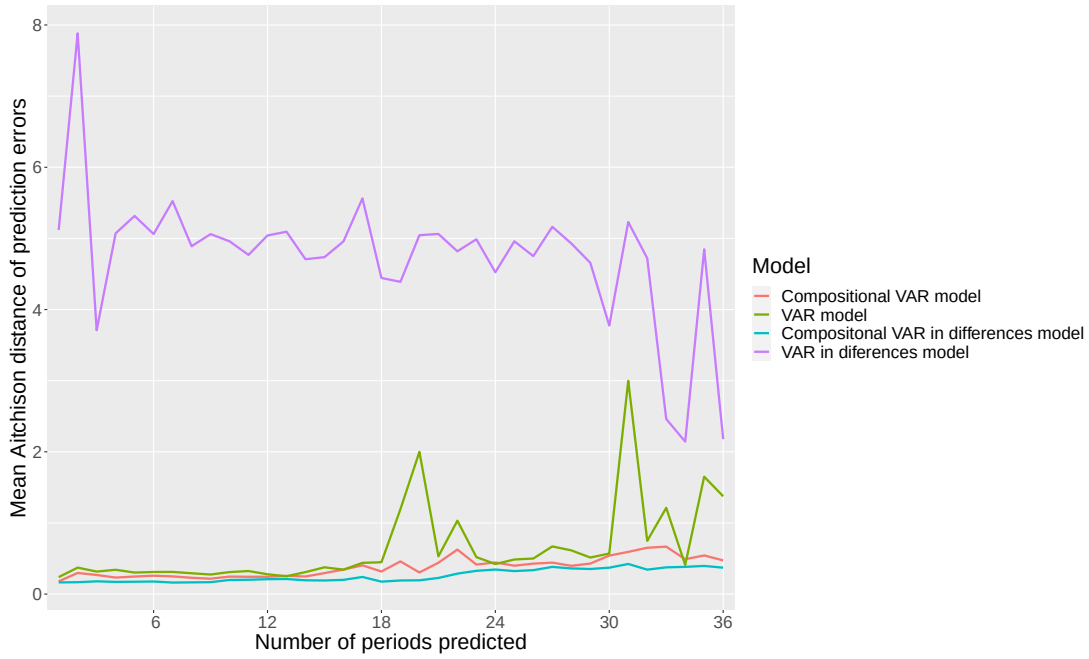


Figure 5: Mean Aitchison distance of prediction errors (MADPE): Compositional VAR model, VAR model, compositional VAR in differences model and VAR in differences model

parameters cannot be easily interpreted.

Furthermore, this leads to another finding: the classical multivariate time series models for the value of assets held by banking institutions are very sensitive to the specification. Indeed, misspecification may lead to a considerable decrease in performance of the model. 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 dataset used for the estimation runs from January 2010 to August 2018 (and September 2018 for the forecast 19 periods ahead). However, there was a sharp increase in the assets held by the two main banking institutions 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. Similar relationships can be found in the other peaks in the MADPE for the VAR models for the value of assets.

Considering all these findings it can be concluded that the basic VAR model with compositional data (basic compositional VAR model) has more advantages than 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 12 (basic VAR model).

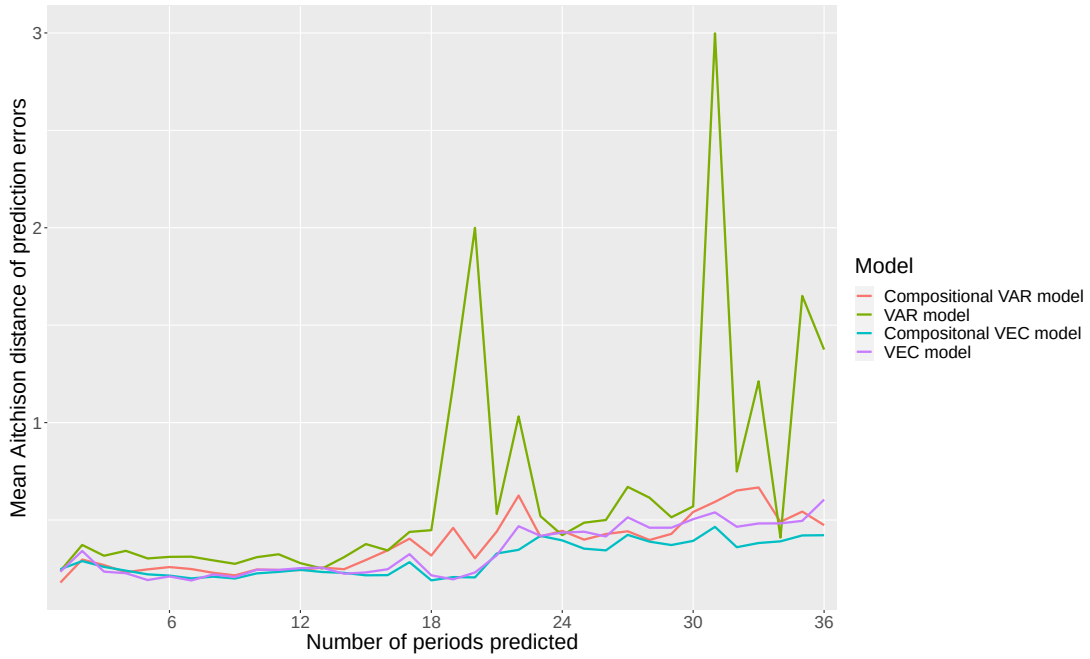


Figure 6: Mean Aitchison distance of prediction errors (MADPE): Compositional VAR model, VAR model, compositional VEC model and VEC model

## 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 a number of variables over time. Although in this article VAR models are used for forecasting, they can also be used for investigating relationships between variables, which are verified by the Granger causality test (Granger, 1969). This test can be interpreted as a significance test for VAR models, as it assesses whether adding lags of one variable improves the forecast of the other(s), i.e. 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. The autocorrelation test intends to determine 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, whilst those from the VAR for the value of assets are autocorrelated.

Regarding the homoscedasticity of the residuals, a multivariate tests cannot be performed because of the high dimensionality of the data. 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 twenty in this case) were estimated. This means 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), whilst 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, whilst those from the model for the value of assets are normally distributed. 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).

### 4.3 Forecast

After the assessment of the selected model (compositional model from Equation 11), the next step is to generate the forecast for 36 periods ahead (three 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 7 shows the expected composition of the assets including the forecasted period to the right of the white line. As can be seen, the participation of each bank in the system is not expected to undergo major changes in coming years. There is only a slight increase in the participation of Davivienda, the third entity in terms of participation, which might lead it to surpass Banco de Bogota (the second one). Apart from that, the composition of the financial system seems to remain similar to that observed in previous years.

However, as can be seen in Figure 8, the decreasing trend of the distance between the benchmark and the actual composition of assets in the Colombian banking 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 in terms of financial stability policy, as, 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 near future.

## 5 Conclusions

Compositional data methodologies have been gaining 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 article, this framework is used to propose an indicator for concentration in financial

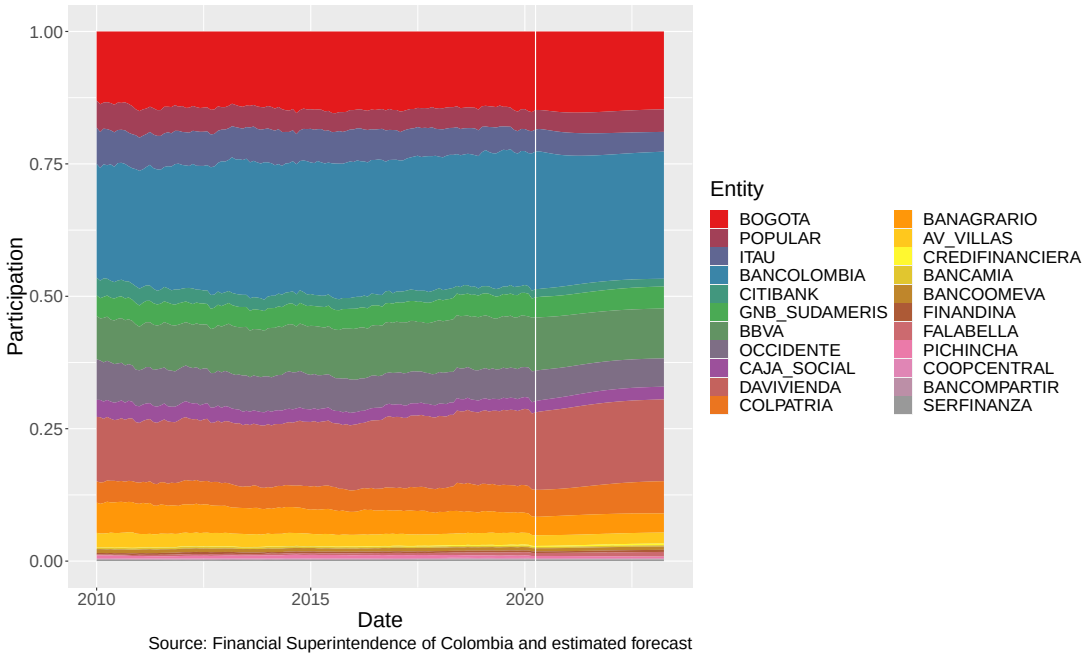


Figure 7: Assets composition per banking establishment from January 2010 to April 2020 and forecast for May 2020 to April 2023

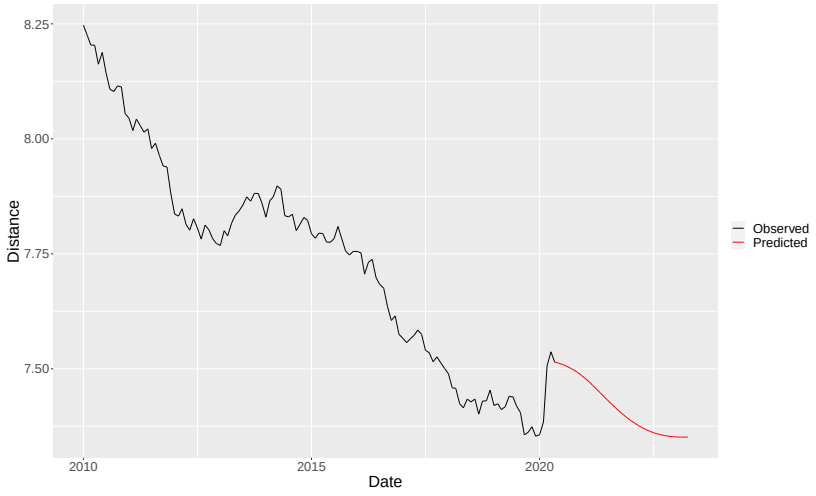


Figure 8: Predicted concentration level index of the Colombian banking system: Aitchison distance between the predicted composition of assets and the hypothetical composition in which assets are equally distributed among all entities

systems, 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 in terms of policy making, considering the lessons learnt during the financial crisis of 2008 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 institutions, to reinforce monitoring of them and maintain the stability of the overall system. Furthermore, the indicator can be extended to assess the impact of mergers and acquisitions between entities on the concentration of the system and, consequently, on its stability.

Compositional multivariate time series methods were used to predict the future composition of the banking system. These methods have shown increased performance in the prediction, considering multiple stylized facts shown by the data. One of the most remarkable results is that the use of compositional methods provides a more robust model with lower sensibility to outliers in the dataset. 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 in terms of autocorrelation and homoscedasticity of errors. In terms of the expected future behavior of the composition of the Colombian banking system and its concentration index trend, the forecast for the next three years show 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 article 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. 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, since this would require to deal with zeros in the compositions. Indeed, this study has excluded those banks entering and leaving the market, considering that they had low participation. However, this limitation can impact the results, especially in longer time series, when relatively important participants can enter or leave the market. This opens an opportunity for further development of the model, considering the existing literature on how to transform compositional data in the presence of zeros (Aitchison, 1986).

## Acknowledgements

This study was supported by the Spanish Ministry of Science and Innovation under grant PID2019-105986GB-C21.

## References

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

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

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.

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

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.

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

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

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

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).

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

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.

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

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*.

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.

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.

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

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.

UBIREA
Institut de Recerca en Economia Aplicada Regional i Pública
*Research Institute of Applied Economics*
**Universitat de Barcelona**
Av. Diagonal, 690 • 08034 Barcelona

**WEBSITE:** www.ub.edu/irea/ • **CONTACT:** irea@ub.edu