Trabajo Fin de Máster

Series Temporales con Variables Composicionales

Rosanny Ventura Lalondriz

Máster en Técnicas Estadísticas

Curso 2025-2026

Propuesta de Trabajo Fin de Máster

Título en galego: Series Temporais con Variables Composicionais
Título en español: Series Temporales con Variables Composicionales
English title: Time Series with Compositional Variables
Modalidad: Modalidad A
Autor/a: Rosanny Ventura Lalondriz, Universidade da Coruña
Director/a: Ana Perez González
Tutor/a:
Breve resumen del trabajo: El trabajo consiste en hacer una revisión de la metodología existente cuando los datos son de tipo composicional. Los datos composicionales han cobrado mucha importancia en los últimos años debido a sus múltiples aplicaciones en disciplinas aplicadas, como la geología o la química. Este trabajo tiene como objetivo explorar las distintas metodologías existentes, y una parte fundamental será realizar alguna aplicación práctica a conjuntos de datos, empleando las metodologías analizadas.
Recomendaciones: Se recomienda el manejo del software R y conocimientos matemáticos de nivel medio (cambios de base, transformaciones ortonormales, etc.)
Otras observaciones:

Doña Ana Perez González , informan que el Trabajo Fin de Máster titulado

Series Temporales con Variables Composicionales

fue realizado bajo su dirección por don/doña Rosanny Ventura Lalondriz para el Máster en Técnicas Estadísticas. Estimando que el trabajo esté terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, doña Ana Perez González, y doña Rosanny Ventura Lalondriz

√ sí □ no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En a Coruña, a 22 de enero de 2026.

PEREZ GONZALEZ ANA - 77594167X
Firmado digitalmente por PEREZ GONZALEZ ANA - 77594167X
Fecha: 2026.01.15 15:43:29 +01'00'

La directora:
doña Ana Perez González

La autora:
doña Rosanny Ventura Lalondriz

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Agradezco, en primer lugar, a Dios, por brindarme la fortaleza, claridad y perseverancia necesarias para culminar esta etapa académica. Su guía ha sido fundamental en cada decisión y desafío superado a lo largo de este proceso.

Asimismo, extiendo mi agradecimiento a todas las personas que, desde el ámbito académico y profesional, contribuyeron con sus conocimientos, orientaciones y críticas constructivas. Cada aporte ha enriquecido significativamente el desarrollo de este trabajo.

Finalmente, agradezco a quienes, de manera directa o indirecta, ofrecieron su apoyo moral, técnico o intelectual. Este logro es también resultado del compromiso colectivo por el conocimiento y el aprendizaje riguroso.

Índice general

Resumen

Resumen en español

Este estudio aborda el modelado de series temporales con variables composicionales, es decir, variables multidimensionales positivas cuyas partes suman una constante y mediante una transformación se pueden ver como valores en el espacio símplex. Se presenta el marco teórico del análisis composicional, con énfasis en las transformaciones que permiten proyectar los datos al espacio euclidiano. Se estudia el modelo compositional C-VARIMA como una extensión del modelo VARIMA clásico, adaptado para mantener la estructura composicional a lo largo del tiempo. Se realiza un estudio de simulación para evaluar el desempeño del modelo y, posteriormente, se aplica a un conjunto de datos reales. Los resultados evidencian la eficacia del modelo para capturar la dependencia temporal y preservar la coherencia composicional en series temporales multivariantes.

English abstract

This study addresses the modeling of time series with compositional variables, in other words, positive multivariate variables whose components sum to a constant and that, through a suitable transformation, can be viewed as values in the símplex space. The theoretical framework of compositional analysis is presented, with emphasis on the transformations that allow projecting the data into Euclidean space. The compositional C-VARIMA model is studied as an extension of the classical VARIMA model, adapted to preserve the compositional structure over time. A simulation study is carried out to evaluate the performance of the model and it is subsequently applied to a real data set. The results show the effectiveness of the model in capturing temporal dependence and preserving compositional coherence in multivariate time series.

Conceptos Básicos

Datos Composicionales

Los datos se consideran como composicionales cuando sus elementos son no negativos y suman una unidad, o en general a una constante fija para todos los elementos. Desde una perspectiva matemática, los datos composicionales pueden ser visualizados como puntos en el espacio símplex. Ejemplos de este tipo de datos incluyen probabilidades, proporciones y porcentajes. De igual forma, Aitchison (1986) [6] define un conjunto de datos composicionales en un espacio real de D dimensiones de la siguiente manera:

$$S^D = \{X = [x_1, x_2, ..., x_D] \in \mathbb{R}^D : x_i \geq 0,\ i = 1, 2, ..., D;\ \sum_{i=1}^{D} x_i = 1\} \tag{1}$$

Karl Pearson (1897) [33], advirtió sobre un problema fundamental al analizar datos compositivos con métodos tradicionales de la Estadística Multivariada: la dificultad de interpretar las correlaciones entre componentes cuyos valores están restringidos por una suma constante, lo que puede inducir correlaciones espurias. En particular, Pearson destacó que al trabajar con cocientes que comparten elementos comunes, las correlaciones resultantes pueden carecer de una base lógica o teórica. A pesar de la importancia de esta observación, su advertencia fue ignorada durante décadas, lo que llevó a una subestimación del problema en muchos análisis posteriores.

Fred Chayes (1960) [17], evidenció cómo la interpretación convencional de las correlaciones entre los componentes de una misma composición generaba correlaciones negativas artificiales. A pesar de este hallazgo, la comunidad analítica continuó aplicando métodos estadísticos clásicos, ignorando la restricción de suma fija inherente a los datos composicionales, lo que inevitablemente distorsionaba sus resultados.

El verdadero punto de inflexión en esta disciplina llegó en la década de 1980, en gran parte gracias a John Aitchison [4]. En 1982, su influyente artículo presentado a la Royal Statistical Society desmenuzó los desafíos que surgían al analizar datos que, por naturaleza, residen en el símplex (es decir, donde sus partes suman una constante).

El ejemplo más básico de datos composicionales involucra sólo dos componentes, lo que significa que la restricción de suma unitaria obliga a que el segundo componente sea simplemente uno menos el primer componente. Esta situación es similar a la que se presenta en las probabilidades de un evento binario. Cox y Snell (1989) [19] abordan este caso utilizando la transformación logit o logística de la probabilidad, lo que facilita la aplicación de modelos de regresión a las probabilidades transformadas mediante logit.

Principios Fundamentales de Datos Composicionales

Invariancia de Escala

Los vectores de componentes positivas que son proporcionales reflejan la misma composición.

Si una composición se multiplica por una constante, por ejemplo, al convertir partes por unidad a porcentajes, la información que se comunica permanece completamente equivalente. Por ende, los vectores de componentes positivas que son proporcionales constituyen una clase de equivalencia. Así, es conveniente seleccionar un representante de dicha clase para simplificar tanto el análisis como la interpretación. La forma convencional de realizar esta selección es normalizar el vector de tal manera que sus componentes sumen a una constante dada κ, que puede ser 1, 100, 1000, 10^6 o cualquier otra constante positiva. Esta elección se formaliza a través de la operación de clausura. Para $\mathbf{x} = (x_1, x_2, ..., x_D)$, un vector con D componentes positivas, su clausura se define como…

$$C\mathbf{x} = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \quad \frac{\kappa x_2}{\sum_{i=1}^{D} x_i}, ..., \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right) \tag{2}$$

Los componentes del vector cerrado se conocen como partes, en relación con un total k. El conjunto de vectores con D componentes positivas que suman una constante k forma el símplex de D partes, que se denota como S^D. Las composiciones equivalentes a $\mathbf{x}$ se representan como $C\mathbf{x}$.

Coherencia Subcomposicional

Los análisis que implican un subconjunto de partes no deben depender de otras partes no incluidas.

Una subcomposición se define como un subconjunto de componentes o partes de una composición. El análisis de una subcomposición requiere que los resultados no sean contradictorios con aquellos obtenidos de la composición completa. El principio de coherencia puede sintetizarse en dos criterios: (a) el principio de invariancia de escala debe aplicarse a cualquiera de las posibles subcomposiciones, lo cual implica la preservación de las proporciones de las partes; (b) cuando se utiliza una distancia o divergencia para comparar composiciones, esta debe ser mayor o igual a la que se obtiene al comparar las subcomposiciones correspondientes (dominancia subcomposicional).

La dominancia subcomposicional requiere una métrica para medir distancias entre composiciones y subcomposiciones que siga la regla de proyección: las distancias deben reducirse al proyectar. Surge la pregunta de si puede emplearse la distancia euclidiana ordinaria entre vectores reales. Este planteamiento no es válido, ya que ambos principios, el de invariancia de escala y el de dominancia subcomposicional, se verían comprometidos. Claramente, si se multiplican dos vectores con componentes positivas por una constante positiva c, la distancia euclidiana entre ellos aumenta en un factor c, infringiendo así el principio de invariancia de escala. Asimismo, la dominancia subcomposicional es violada por la distancia euclidiana ordinaria entre vectores composicionales.

Invariancia por Permutación

Las conclusiones de un análisis composicional no deberían depender del orden de las partes.

Por ejemplo, en composiciones geoquímicas, a menudo se ordenan las partes alfabéticamente. Un caso típico es la distribución de tamaños de grano en un sedimento: las partículas se clasifican, tras un tamizado, en categorías de tamaño. En un análisis composicional, la información relativa al orden de las diferentes clases no tiene relevancia.

Geometría de Aitchison

El desarrollo de los conceptos sugeridos por Aitchison (1986) [6] ha conducido a la geometría Aitchison del símplex. Esta geometría, siendo euclidiana, requiere definiciones específicas y una métrica particular. Consideremos las composiciones $\mathbf{x}, \mathbf{y} \in S^D$. La perturbación de $\mathbf{x}$ con $\mathbf{y}$ se define como la composición

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, x_2 y_2, \dots, x_D y_D) \tag{3}$$

y la potenciación de x por un número real α se define como la composición

$$\alpha \odot \mathbf{x} = C(x^{\alpha_1}, x^{\alpha_2}, \dots, x^{\alpha_D}). \tag{4}$$

Es fácil demostrar que para $\mathbf{n} = C(1, 1, \dots, 1)$ se cumple que $\mathbf{x} \oplus \mathbf{n} = x$. Así, la composición con todas las partes iguales es el elemento neutro de la perturbación. La perturbación y la potenciación, definidas en S^D, satisfacen los requisitos para operaciones de un espacio vectorial. Sin embargo, la principal ventaja de la perturbación es que, además de satisfacer los principios del análisis composicional, generalmente tiene una interpretación en el campo analizado.

El cambio de unidades en algunas o todas las partes de una composición también puede verse como una perturbación. Ejemplos típicos se encuentran en química, cuando las concentraciones en partes por millón (ppm) de peso se cambian a concentraciones molares (Buccianti y Pawlowsky-Glahn, 2005) [16]. Esto se realiza multiplicando cada componente por el inverso del peso molar. Cerrar la composición resultante puede ser innecesario en muchos casos. Aun así, conserva su carácter composicional.

La invariancia de escala requerida para un análisis composicional conduce al uso de razones entre partes, de modo que se cancelen las constantes de escala. Además, estas razones se interpretan en una escala relativa, y tomar sus logaritmos es entonces una elección natural. El análisis de CoDa (Datos composicionales) se basa esencialmente en el análisis estadístico de log-ratios entre partes. Los log-ratios más simples son aquellos que comparan dos partes.

Los log-ratios son útiles en el análisis, pero deben ser invariantes de escala,

$$\ln\left(\frac{x_i}{x_j}\right). \tag{5}$$

Frecuentemente, algunas preguntas pueden ser respondidas analizando un log-contraste apropiado, el cual generaliza el caso anterior al considerar una combinación lineal de logaritmos de las partes con coeficientes que suman cero:

$$\sum_{k=1}^{D} a_k \ln(x_k), \qquad \text{con } \sum_{k=1}^{D} a_k = 0. \tag{6}$$

La elección del log-contraste depende del problema planteado y de la interpretación de la composición. Nótese que el log-ratio en (5) es un caso particular de (6), al tomar $a_i = 1$, $a_j = -1$ y $a_k = 0$ para $k \neq i, j$.

Representaciones adecuadas y completas de una composición utilizando un conjunto de log-contrastes fueron propuestas en la década de 1980 (Aitchison, 1986) [6], de modo que toda la información de la composición se invierte en el conjunto de log-ratios.

Además, el siguiente producto interno, con su norma y distancia asociadas, puede ser utilizado para obtener una estructura de espacio de Hilbert finita (de $D-1$ dimensiones) (Billheimer et al., 2001; Pawlowsky-Glahn y Egozcue, 2001) [12] [31] :

El producto interno de Aitchison se define para dos composiciones $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ como

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right) \tag{7}$$

Este producto interno induce una estructura de espacio vectorial euclídeo sobre el símplex $\mathcal{S}^D$. A partir de él, se puede construir una norma y una distancia en el símplex:

$$\|\mathbf{x}\|_a^2 = \langle \mathbf{x}, \mathbf{x} \rangle_a, \quad d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a \tag{8}$$

El producto interno, la norma y la distancia Aitchison cumplen los principios del análisis composicional y, por lo tanto, son herramientas para un análisis composicional libre de inconsistencias. Junto con la perturbación y la potenciación, proporcionan una estructura euclidiana al símplex, llamada geometría simplicial Aitchison. Esto sugiere explotar las propiedades bien conocidas de los espacios euclidianos para analizar composiciones: base ortonormal, representación de coordenadas (ortonormales), proyecciones ortogonales, definiciones de ángulos, elipses, etc.

Finalmente, usamos $A_{D\times D}$ para denotar la familia de todas las matrices reales $D \times D$. Si $x \in S^D$ y $A \in A_{D\times D}$, definimos el producto $A \odot x$ como

$$A \odot x = C\left(\prod_{j=1}^{D} x_j^{a_{1j}}, ..., \prod_{j=1}^{D} x_j^{a_{Dj}}\right). \tag{9}$$

Por lo tanto, la función $x \to A \odot x$ es un endomorfismo del espacio vectorial $(S^D, \oplus, \odot)$. Además, cualquier endomorfismo de S^D puede escribirse de esta forma. La matriz asociada con el endomorfismo identidad es la bien conocida matriz de centrado $G_D = I_D - D^{-1}J_D$ de orden $D \times D$, donde I_D es la matriz identidad $D \times D$ (con unos en la diagonal principal y ceros en las demás entradas) y $J_D = 1_D 1_D^{\mathsf{T}}$ es la matriz de unos de dimensión $D \times D$.

Transformación alr

Una primera elección de estas representaciones fue la transformación de log-ratio aditivo (alr). Si x es una composición en el símplex de D partes S_D, se define como

$$alr_k(\mathbf{x}) = \ln\left(\frac{x_1}{x_D}, \frac{x_{k-1}}{x_D}, \frac{x_{k+1}}{x_D}, ..., \frac{x_{D-1}}{x_D}\right) \tag{10}$$

donde el logaritmo natural ln se aplica componente a componente. En consecuencia, el componente i es el simple log-ratio $alr_k(\mathbf{x}) = \ln\left(\frac{x_k}{x_D}\right)$. La transformación alr se invierte fácilmente para obtener la composición original a partir de los $D-1$ componentes alr y también reduce la perturbación y la potenciación a operaciones ordinarias en el espacio real de $D-1$ dimensiones:

$$alr((\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{y})) = \alpha \cdot alr(\mathbf{x}) + \beta \cdot alr(\mathbf{y}), \tag{11}$$

para cualquier par de vectores composicionales $(\mathbf{x}, \mathbf{y})$ y cualesquiera constantes reales α y β. Sin embargo, el alr tiene la desventaja de no ser invariante bajo la permutación de componentes, lo que puede causar fallos en algunos procedimientos estadísticos.

Transformación clr

Aitchison (1986) [6] introdujo la transformación log-ratio centrada (clr), que representa una composición de D-partes utilizando D coeficientes clr. Se define como

$$v = clr(x) = \ln\left(\frac{x_1}{gm(x)}, \frac{x_2}{gm(x)}, \dots, \frac{x_D}{gm(x)}\right) \tag{12}$$

donde $gm(x)$ es la media geométrica de las componentes

$$gm(x) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}, \tag{13}$$

los D coeficientes $clr_i(x) = \ln\left(\frac{x_i}{gm(x)}\right)$ son log-contrastes. A partir de $clr(x)$, se recupera la composición x con la transformación inversa de clr:

$$x = clr^{-1}(v) = C\exp(v), \tag{14}$$

donde la función exponencial se aplica componente a componente a $v = clr(x)$. De manera similar a la transformación alr, la perturbación y la potenciación en S_D corresponden a la suma y el producto en el espacio real R^D:

$$clr((\alpha \odot x) \oplus (\beta \odot y)) = \alpha \cdot clr(x) + \beta \cdot clr(y). \tag{15}$$

La desventaja de la transformación clr es que utiliza D coeficientes, que suman cero, para representar una composición que solo tiene $D-1$ componentes libres, la dimensión de S^D. Además, los componentes clr cambian al trabajar con una subcomposición.

Transformación ilr

Un paso importante para utilizar estos conceptos es construir bases ortonormales y sus correspondientes coordenadas. Una base ortonormal de S^D es un conjunto de composiciones $e_1, e_2, \dots, e_{D-1}$ tal que $\langle e_i, e_j\rangle_a = 0$ para $i \neq j$, y $\|e_i\|_a = 1$. Para una base fija, las coordenadas de una composición se obtienen mediante la función

$$\mathbf{x}^* = ilr(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1\rangle_a, \langle \mathbf{x}, \mathbf{e}_2\rangle_a \dots, \langle \mathbf{x}, \mathbf{e}_{D-1}\rangle_a), \tag{16}$$

con la inversa,

$$\mathbf{x} = ilr^{-1}(\mathbf{x}^*) = \bigoplus_{j=1}^{D-1} x_j^* \odot \mathbf{e}_j \tag{17}$$

La construcción de coordenadas ortonormales se ha denominado transformación log-ratio isométrica (ilr) (Egozcue et al. 2003) porque las coordenadas $x_j^* = ilr_j(\mathbf{x})$ son contrastes logarítmicos e isométricos:

$$ilr((\alpha \dots \mathbf{x}) \oplus (\beta \dots \mathbf{y})) = \alpha \cdot ilr(\mathbf{x}) + \beta \cdot ilr(\mathbf{y}), \tag{18}$$

Dada una base ortonormal específica, utilizada a modo de ejemplo y propuesta por Egozcue et al.[20],la transformación inversa clr^{-1} a un conjunto de vectores ortonormales definidos en un subespacio de $\mathbb{R}^D$ con suma cero. Cada vector base está formado por i valores iguales, un valor negativo en la posición $i+1$, y ceros en las posiciones restantes.Esta construcción corresponde a una de las múltiples bases posibles que pueden emplearse en la transformación ILR. En términos exponenciales:

- Los primeros i componentes son $\exp\left(\sqrt{\frac{1}{i(i+1)}}\right)$,
- El componente $i+1$ es $\exp\left(-\sqrt{\frac{i}{i+1}}\right)$,
- Los componentes restantes son $\exp(0) = 1$.

Después de aplicar la exponenciación, los vectores se normalizan dividiendo cada componente por la suma total, asegurando así que la composición resultante pertenezca al símplex. Este conjunto de composiciones forma una base ortonormal bajo el producto interno de Aitchison, preservando la estructura geométrica propia del espacio composicional.

Los componentes de la transformación $\mathbf{y} = \text{ilr}(\mathbf{x})$, calculados con respecto a una base ortonormal del símplex, se definen como:

$$y_i = \sqrt{\frac{i}{i+1}} \ln\left(\frac{g(x_1, \dots, x_i)}{x_{i+1}}\right), \quad \text{para } i = 1, 2, \dots, D-1, \tag{19}$$

donde $g(x_1, \dots, x_i)$ representa la media geométrica de los primeros i componentes de la composición $\mathbf{x}$. Estas coordenadas son log-ratios diseñados para capturar relaciones relativas entre partes, lo cual es una propiedad clave en el análisis composicional.

Ventajas y Desventajas de las Transformaciones

Cuadro 1: Resumen de las transformaciones

Transformaciones	**Ventajas**	**Desventajas**
alr (Logaritmo de razones aditivo): La transformación se basa en el logaritmo de razones, incorporando una única variable de referencia en el denominador. $alr_k(\mathbf{x}) = \ln\left(\frac{x_1}{x_D}, \frac{x_{k-1}}{x_D}, \frac{x_{k+1}}{x_D}, \dots, \frac{x_{D-1}}{x_D}\right)$	Transforma las operaciones de perturbación y potenciación en el símplex a operaciones equivalentes de adición y multiplicación por un escalar en el espacio euclidiano	No es isométrica. La transformación alr no preserva la distancia de Aitchison; la distancia euclidiana en las coordenadas alr no coincide con la del símplex. No es simétrica respecto al denominador. La componente usada en el denominador queda como referencia, de modo que el resultado depende de cuál parte se elija y no se cumple la invariancia por permutación.

(continuación)

Cuadro 1: Resumen de las transformaciones (continuación)

clr (Logaritmo centrado): transformación isométrica que se fundamenta en el logaritmo de razones en función de la media geométrica de las variables. $clr(x) = \ln\left(\frac{x_1}{gm(x)}, \frac{x_2}{gm(x)}, \dots, \frac{x_D}{gm(x)}\right)$	Evita la selección de una proporción variable y facilita la interpretación de las variables transformadas, permitiendo el análisis en función de las variables originales.	Los datos transformados muestran incoherencia subcomposicional, lo que da lugar a una matriz de datos singular, lo que dificulta la utilización de técnicas robustas para datos en esas coordenadas.
ilr (Logaritmo isométrico): transformación isométrica que se apoya en la elección de una base ortonormal $e_1, e_2, \dots, e_{D-1}$ dentro del hiperplano definido por las coordenadas transformadas de $e_1, i = 1, 2, \dots, D-1$. $ilr(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_i \rangle_a$,	Mantiene todas las propiedades favorables de la transformación y se adhiere a todos los principios del análisis composicional.	Las correlaciones entre sus coordenadas no pueden interpretarse directamente en términos de las variables composicionales originales. Esto se debe a que la relación entre las partes y las coordenadas ilr es no lineal, lo que dificulta la trazabilidad conceptual de estas dependencias en el espacio original del símplex.

Distribuciones Composicionales

Durante mucho tiempo, la distribución de Dirichlet fue la única opción analíticamente tratable para modelar este tipo de datos. Sin embargo, esta distribución presenta una limitación fundamental: asume independencia subcomposicional completa. Esto significa que, bajo cualquier partición de la composición, las subcomposiciones resultantes deben ser mutuamente independientes, lo cual rara vez se cumple en contextos empíricos. Esta restricción impide modelar estructuras de dependencia realistas entre componentes, haciendo que su aplicabilidad sea limitada en muchas situaciones prácticas.

Distribución Logística Normal

La distribución logística-normal fue definida por Aitchison y Shen (1980) [3] y estudiada en profundidad por Aitchison (1986) [6]. Posteriormente, Mateu-Figueras y Pawlowsky-Glahn (2008) [32] la reformulan en el marco de la geometría de Aitchison, interpretándola como una distribución normal en el símplex mediante coordenadas log-ratio isométricas (ilr). Se dice que un vector aleatorio Y de dimensión D sigue una distribución logística-normal $\mathcal{LN}(\mu, \Sigma)$, o alternativamente una distribución normal en el espacio S^D, si cualquier vector de coordenadas de razón logarítmica tiene una distribución normal conjunta de $D-1$ dimensiones. Esta definición puede adaptarse a una respuesta CoDa utilizando las coordenadas ALR, de la siguiente manera:

$$\mathbf{y}|\mu, \Sigma \sim \mathcal{LN}(\mu, \Sigma) \iff alr(\mathbf{y})|\mu, \Sigma \sim \mathcal{N}(\mu, \Sigma), \tag{20}$$

donde μ es un vector de dimensión $D-1$ y Σ es una matriz de covarianza de tamaño $(D-1) \times (D-1)$. Aunque aquí se utiliza la parametrización basada en las coordenadas ALR, una caracterización equivalente puede obtenerse empleando coordenadas ilr.

Distribución de Dirichlet

La distribución de Dirichlet fue introducida por Connor y Mosimann (1969) [18] y es una generalización de la conocida distribución Beta. Un vector aleatorio $\mathbf{Y}$ de dimensión D tiene una distribución de Dirichlet $D(\alpha)$ si su función de densidad de probabilidad es:

$$p(\mathbf{y}|\alpha) = \frac{1}{B(\alpha)} \prod_{d=1}^{D} y_d^{\alpha_d - 1}, \tag{21}$$

donde $\alpha = (\alpha_1, \dots, \alpha_D)$ es el vector de parámetros de forma para cada categoría ($\alpha_d > 0$ para todo d), $\sum_{d=1}^{D} y_d = 1$, y $B(\alpha)$ es la función multinomial Beta, que actúa como constante de normalización. La función multinomial Beta se define como:

$$B(\alpha) = \frac{\prod_{d=1}^{D} \Gamma(\alpha_d)}{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}, \tag{22}$$

donde $\alpha_0 = \sum_{d=1}^{D} \alpha_d$ es el parámetro de precisión. La distribución Beta es un caso particular cuando $D = 2$. Además, cada variable marginalmente sigue una distribución Beta con $\alpha = \alpha_d$ y $\beta = \alpha_0 - \alpha_d$. Si $\mathbf{Y} \sim D(\alpha)$, los valores esperados, varianzas y covarianzas son:

$$E(y_d) = \frac{\alpha_d}{\alpha_0}, \quad Var(y_d) = \frac{\alpha_d(\alpha_0 - \alpha_d)}{\alpha_0^2(\alpha_0 + 1)}, \quad Cov(y_d, y_k) = -\frac{\alpha_d \alpha_k}{\alpha_0^2(\alpha_0 + 1)}. \tag{23}$$

Comparativa de la Distribución Logística Normal y Dirichlet

Como señala Aitchison (1986, pp. 126-129) [6], las distribuciones logística-normal y de Dirichlet son distintas y no coinciden exactamente para ningún conjunto de parámetros. Sin embargo, a través de la divergencia de Kullback-Leibler (KL), que mide cuánto se desvía una aproximación q del objetivo p, se puede aproximar la distribución de Dirichlet mediante la distribución logística-normal.

El problema de minimizar KL:

$$K(p, q) = \int_{S_D} p(\mathbf{y}|\alpha) \log \frac{p(\mathbf{y}|\alpha)}{q(\mathbf{y}|\mu, \Sigma)} \, d\mathbf{y}, \tag{24}$$

donde $p(\mathbf{y}|\alpha)$ es la función de densidad de Dirichlet y $q(\mathbf{y}|\mu, \Sigma)$ es la densidad logística-normal, se resuelve con:

$$\mu = E\left[\log \frac{y_1}{y_D}, \dots, \log \frac{y_{D-1}}{y_D}\right] = E[alr(\mathbf{y})], \quad \Sigma = Var\left[\log \frac{y_1}{y_D}, \dots, \log \frac{y_{D-1}}{y_D}\right] = Var[alr(\mathbf{y})]. \tag{25}$$

La solución, expresada en términos de las funciones digamma (ψ) y trigamma (ψ')[1], es:

$$\mu_d = \psi(\alpha_d) - \psi(\alpha_D), \quad \Sigma_{dd} = \psi'(\alpha_d) + \psi'(\alpha_D), \quad \Sigma_{dk} = -\psi'(\alpha_D), \quad d \neq k. \tag{26}$$

[1] La función digamma $\psi(x)$ es la derivada del logaritmo de la función gamma, es decir, $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$. Su derivada, llamada función trigamma, se denota $\psi^{(1)}(x)$ y representa la segunda derivada logarítmica de $\Gamma(x)$. Véase Abramowitz y Stegun (1972) [1].

Modelado de Series Temporales Composicionales

Contexto Histórico y Problemática

Las series temporales multivariantes de proporciones, o composiciones, surgen en muchas áreas de aplicación. Estas series se caracterizan por D componentes no negativas $x_{1t}, ..., x_{Dt}$, que suman a una constante en cada tiempo t. Sin pérdida de generalidad, se puede suponer que la constante en cuestión es 1. Es habitual referirse a la serie $x_t : t = 1, ..., n$, donde $x_t = (x_{1t}, ..., x_{Dt})$, como una Serie Temporal Composicional o CTS por sus siglas en inglés. Los x_t son elementos del símplex S^D. Este tipo de datos aparece con frecuencia en disciplinas tan dispares como biología, demografía, ecología, economía, geología y política. Aunque una CTS constituye una serie temporal multivariante, las técnicas estándar, como las utilizadas en los modelos de media móvil autoregresiva integrada multivariante (VARIMA, por sus siglas en inglés), no son aplicables debido a la restricción de la suma constante (Barceló-Vidal et al. 2007) [8].

Históricamente, la modelización de CTS se ha basado casi exclusivamente en el enfoque de transformación, que consiste en la aplicación de una transformación inicial para romper la restricción de la suma unitaria, seguida del uso de técnicas estándar para modelar la serie temporal transformada. Así, se habilita la posibilidad de modelización VARIMA. En este contexto, una de las transformaciones más frecuentemente empleadas ha sido la transformación log-ratio aditiva (alr). Esta transformación depende de la elección del componente utilizado como denominador común en los log-ratios, por lo que existen tantas transformaciones alr posibles como partes, D, de los datos composicionales.

El enfoque de modelización basado en VARIMA para series temporales composicionales transformadas con alr ha sido empleado por Brunsdon (1987) [14], Smith y Brunsdon (1989) [39], y Brunsdon y Smith (1998) [15]. Ravishanker et al. (2001) [36] generalizaron el enfoque de Brunsdon y Smith (1998) [15] a una extensión de los modelos VARMA incorporando covariables. Aplicaciones recientes de este enfoque se encuentran en Mills (2009, 2010) [29] [28]. La mayoría de estas contribuciones concluyen que las predicciones son invariante a la elección del componente utilizado en el denominador común de la transformación alr. Esto es cierto si, como en las publicaciones citadas, sólo se contemplan modelos VARIMA completos, es decir, aquellos que incluyen todas las variables relevantes y posibles interacciones.

Estos modelos buscan capturar toda la dinámica estructural y estacional de las series de tiempo,proporcionando una representación más completa de los datos. Se contempla la posibilidad de simplificación post-estimación utilizando modelos restringidos, los cuales se construyen imponiendo restricciones específicas sobre los parámetros del modelo completo, como la eliminación de términos insignificantes o la fijación de ciertos coeficientes a valores predefinidos. Esta simplificación tiene como objetivo reducir el riesgo de sobreajuste y facilitar la interpretación, manteniendo un nivel aceptable de capacidad predictiva y explicativa. Otras contribuciones al análisis de CTS usando la

transformación alr son Silva (1996) [37] y Silva y Smith (2001) [38], quienes emplean un enfoque de modelización en el espacio de estados para la serie temporal transformada.

La transformación log-ratio centrada (o simétrica) (clr) fue utilizada por Quintana y West (1988) [35] para analizar datos de CTS mediante un modelo de regresión dinámica . Estos autores manejaron las singularidades de las matrices de covarianzas asociadas a la serie temporal transformada $clr(x_t) = z_t$ ignorando la restricción de la suma cero sobre los z_t. Modelaron la serie z_t asumiendo la no singularidad de las matrices de covarianzas e imponiendo a posteriori la restricción de la suma cero sobre el modelo estimado.

El modelado directo de la serie transformada mediante clr presenta serias dificultades técnicas debido a la singularidad inherente de las matrices de covarianzas. Las estrategias propuestas en la literatura para sortear este problema suelen ser, en general, poco compatibles con la estructura composicional, ya que a menudo ignoran la restricción de suma cero o tratan las matrices de covarianzas como si fueran no singulares. Bergman (2008) [10] utilizó la transformación log-ratio isométrica (ilr) para ajustar un modelo VAR a series temporales composicionales mensuales de la Encuesta de Fuerza Laboral de Suecia. La transformación ilr depende de la base ortonormal de S^D elegida en su definición. Bergman (2008) [10] empleó solo modelos completos para los datos transformados con ilr, es decir, aquellos que incluyen todas las variables relevantes y posibles interacciones, de modo que las predicciones de los modelos finales no dependen de la base ortonormal utilizada en la transformación ilr.

Por otra parte, Bhaumik et al. (2003) [11] utilizaron la conocida transformación Box-Cox aplicada a los cocientes de los componentes de una CTS como alternativa a la transformación alr.

$$\left(\frac{x_i}{x_D}\right)^{(\lambda)} = \begin{cases} \dfrac{\left(\frac{x_i}{x_D}\right)^{\lambda} - 1}{\lambda}, & \text{si } \lambda \neq 0 \\ \log\left(\frac{x_i}{x_D}\right), & \text{si } \lambda = 0 \end{cases} \tag{27}$$

Estos trabajos proponen modelar la dinámica temporal de composiciones trabajando en una representación de cocientes y aplicando una transformación Box–Cox antes del ajuste. Sobre las variables transformadas se emplean modelos lineales dinámicos, incorporando una clase amplia de distribuciones para los errores mediante mezclas de escalas de normales multivariadas [34]. La familia Box–Cox resulta atractiva porque contiene a la transformación logarítmica como caso particular (y, por tanto, puede recuperar el enfoque ALR); sin embargo, introduce parámetros adicionales que deben estimarse. Al igual que ocurre con ALR, este procedimiento depende de la parte seleccionada como denominador en la construcción de los cocientes.

Otra vía, conocida como enfoque basado en datos, parte directamente de la distribución inherente a los datos originales. Se trata de un método más intuitivo y, por lo general, más sencillo de interpretar. Para datos composicionales, la distribución más adecuada es la de Dirichlet. Aunque su marcada estructura de dependencia interna llevó a descartarla en contextos en los que se asumía independencia entre componentes Aitchison (1986) [6], ha demostrado ser muy valiosa cuando se emplea como distribución condicional.

En el caso de las series temporales composicionales, Grunwald et al. (1993) [21] plantean un enfoque que respeta directamente las restricciones del símplex al introducir un estado latente $\mathbf{x}_t$ que gobierna el comportamiento de la composición observada $\mathbf{y}_t$. En particular, se asume que $\mathbf{y}_t$ sigue una distribución Dirichlet condicionada a $\mathbf{x}_t$, mientras que $\mathbf{x}_t$ evoluciona en el tiempo con una dependencia Markoviana de primer orden, de modo que su distribución en t depende esencialmente de $\mathbf{x}_{t-1}$; esta transición puede describirse mediante una Dirichlet o una Dirichlet generalizada. En esta línea, Connor (1969) [18] amplía el planteamiento al proponer una generalización capaz de capturar estructuras de dependencia más flexibles entre los componentes, mitigando la rigidez de la Dirichlet estándar en la forma en que induce asociación entre las partes.

Ceros en Series de Tiempo Composicionales

La presencia de componentes exactamente nulos en series composicionales x_t representa una limitación fundamental para la aplicación directa de las transformaciones log-ratio convencionales, tales como la log-ratio aditiva (alr), la log-ratio centrada (clr) y la log-ratio isométrica (ilr). Para sortear esta indeterminación, la literatura ha propuesto históricamente la sustitución de ceros por constantes positivas pequeñas (e.g., Bell et al., 1986 [9] ; siguiendo el marco de Aitchison, 1986 [6]; y Martín-Fernández et al., 2003 [25]). Sin embargo, este procedimiento, si bien práctico, introduce un sesgo potencial significativo y compromete la robustez de los estimadores, especialmente en escenarios de escasez de datos o alta frecuencia de ceros.

Como respuesta a esta problemática, se han explorado alternativas metodológicas. Una de ellas es la transformación hiperesférica (Nolan y Smith, 1995 [30]; Wang et al., 2007 [42]), que utiliza la función arcocoseno para proyectar los datos sobre una hipersfera, eludiendo así la dependencia de los logaritmos.

Modelos C-VARIMA: Teoría y Supuestos

Supuestos del Modelos C-VARIMA

Al analizar datos con modelos C-VARIMA para series de tiempo composicionales, es crucial que se cumplan ciertos supuestos para asegurar la validez, estacionariedad e interpretabilidad de los resultados. Estas condiciones son una adaptación de los requisitos de los modelos VARIMA euclidianos, pero están formuladas dentro del marco de la geometría composicional [Egozcue et al., 2003] [20].

1. Naturaleza de los Datos Composicionales:

La base para cualquier análisis composicional es la correcta caracterización de los datos de entrada:

- *Valores Estrictamente Positivos*: Cada componente de las observaciones de la serie de tiempo X_t debe ser estrictamente positiva ($x_t, i > 0$ para todo i,t). Esta condición es fundamental para que las transformaciones log-ratio (ilr, clr) y las operaciones composicionales (como la perturbación y el escalado) estén matemáticamente bien definidas, Aitchison (1986) [6] .
- *Suma Constante*: La suma de las componentes de cada observación composicional debe ser una constante predefinida (comúnmente 1, si se trata de proporciones o porcentajes). Esto asegura que los datos residan en el espacio símplex S^D, que es el dominio natural para los datos composicionales, Aitchison (1986) [6].

2. Estacionariedad:

En el contexto composicional, se dice que una serie temporal $\{X_t\}_{t\in\mathbb{Z}}$, con $X_t \in S^D$ para todo t, es un proceso *C-estacionario* si mantiene constantes a lo largo del tiempo tanto su media composicional como su estructura de dependencia de segundo orden, formulada en términos de covarianzas composicionales.

En particular, la *media composicional* (o C-media) se define como

$$\xi \;=\; \mathbb{E}_C[X_t] \;=\; \mathcal{C}(\exp(\mathbb{E}[\ln X_t])), \tag{28}$$

donde $\ln X_t$ denota el vector de logaritmos componente a componente, $\mathbb{E}[\ln X_t]$ es la esperanza clásica aplicada componente a componente, $\exp(\cdot)$ se aplica nuevamente componente a componente y $\mathcal{C}(\cdot)$ es el operador de *closure* que normaliza el vector resultante para que sus componentes sumen 1.

Decimos que la serie $\{X_t\}$ es C-estacionaria si esta media composicional

$$\xi = \mathbb{E}_C[\mathbf{X}_t] \tag{29}$$

permanece constante para todo t, y si además la estructura de dependencia de segundo orden puede describirse mediante covarianzas composicionales que sólo dependen del desfase (lag) entre observaciones, y no del tiempo absoluto. La C-media ξ actúa como un "centro de gravedad'' bajo la geometría del símplex y refleja el equilibrio relativo de las partes composicionales en el tiempo.

- *Autocovarianza composicional invariante en el tiempo:*
  La función de autocovarianza composicional se define como:

$$\Gamma_C(h) = \mathbb{E}\left[(\text{clr}(\mathbf{X}_{t+h}) - \text{clr}(\xi))\,(\text{clr}(\mathbf{X}_t) - \text{clr}(\xi))^\top\right], \tag{30}$$

y depende únicamente del rezago h, no del tiempo absoluto t. Esto implica que las relaciones de dependencia entre las partes de la composición son estables a lo largo del tiempo.

- *Autocorrelación composicional:* La correspondiente función de autocorrelación composicional está dada por:

$$R_C(h) = \left[\frac{\gamma_{C,ij}(h)}{\sqrt{\gamma_{C,ii}(0)\gamma_{C,jj}(0)}}\right]_{i,j=1}^{D}, \tag{31}$$

donde $\gamma_{C,ij}(h)$ representa la covarianza composicional entre las partes i y j con rezago h.

La propiedad de C-estacionariedad garantiza que el comportamiento conjunto y proporcional de las partes de la composición no varía con el tiempo. Además, si $\{\mathbf{X}_t\}$ es C-estacionario, entonces cualquier transformación lineal válida al espacio real (como clr, ilr o alr) genera un proceso estacionario en el sentido clásico.

3. Invertibilidad:

La invertibilidad hace referencia a la posibilidad de representar el modelo como un VAR de orden infinito, lo cual resulta crucial tanto para su estimación como para su interpretación:

- *Raíces del Polinomio de Medias Móviles Fuera del Círculo Unitario:* Al igual que en el caso de la estacionariedad, para que el componente de medias móviles del modelo C-VARMA transformado al espacio símplex sea invertible, es necesario que todas las raíces del polinomio característico se ubiquen fuera del círculo unitario

4. Propiedades del Término de Error (Ruido Blanco Composicional):

En el marco composicional, se denomina **ruido blanco composicional** a un proceso $\{\mathbf{W}_t\}$ que cumple propiedades específicas de primer y segundo orden en el símplex. Este tipo de proceso se denota como:

$$\{\mathbf{W}_t\} \sim WN_C(\mathbf{1}_C, \mathbf{C}), \tag{32}$$

donde $\mathbf{1}_C = \left(\frac{1}{D}, \ldots, \frac{1}{D}\right)$ es la media composicional uniforme y $\mathbf{C}$ es la matriz de covarianza en el rezago cero.

Las propiedades fundamentales que debe cumplir son:

- *Media composicional uniforme:*

$$\mathbb{E}_C[\mathbf{W}_t] = \mathbf{1}_C. \tag{33}$$

donde $\mathbf{1}_C = \left(\frac{1}{D}, ..., \frac{1}{D}\right)$ representa una composición uniforme de D partes. Esta media refleja que todas las partes de la composición tienen igual peso relativo.

Esta propiedad indica que las innovaciones, en promedio, no favorecen a ninguna parte específica de la composición, manteniendo una distribución proporcional equilibrada y simétrica en el símplex.

- *Autocovarianza nula para rezagos no nulos:*

$$\Gamma_C(0) = \mathbf{C}, \quad \Gamma_C(h) = \mathbf{0}_{D\times D} \quad \text{para todo } h \neq 0. \tag{34}$$

Esto asegura que no existe correlación serial entre errores en diferentes momentos del tiempo.

- *C-estacionariedad del proceso de error:*
  Un proceso de **ruido blanco composicional** es, por definición, **C-estacionario**, ya que mantiene constante su media composicional y su estructura de autocovarianza cumple con las condiciones de invarianza en el tiempo.
- *Normalidad composicional:*
  Si las transformaciones del proceso $\{\mathbf{W}_t\}$, como $\{\mathbf{Y}_t\} = \text{alr}(\mathbf{W}_t)$ o $\{\mathbf{U}_t\} = \text{clr}(\mathbf{W}_t)$, son independientes e idénticamente distribuidas según una normal multivariada, entonces $\{\mathbf{W}_t\}$ se denomina ruido blanco composicional gaussiano. En tal caso, las transformaciones $\{\mathbf{Z}_t\} = \text{ilr}(\mathbf{W}_t)$ siguen una distribución normal degenerada, lo cual respeta las restricciones del símplex.

Modelos C-VARIMA

Sea $\{x_t : t = 0, \pm1, \pm2, ...\}$ una serie temporal composicional formada por variables aleatorias de la forma $x_t = (x_{1t}, ..., x_{Dt})^\top$ definida en S^D (es decir, un proceso). Las propiedades de segundo orden de $\{x_t\}$ están especificadas por los vectores de C-media, $\xi_t = E_C\{x_t\} = (\xi_{t1}, ..., \xi_{tD})^\top$, la media composicional en el tiempo t, denotada como ξ_t, es el vector de medias de las D componentes de la serie de tiempo composicional x_t y matrices de C-autocovarianza.

$$\Gamma_C(t+h,t) = E(clr(\mathbf{x}_{t+h}) - clr(\xi_{t+h}))(clr(\mathbf{x}_t) - clr(\xi_t))^\top = [\Gamma_{C,ij}(t+h,t)_{i,j=1}^D] \tag{35}$$

Es importante notar que, en el contexto composicional, dado un proceso temporal composicional $\{\mathbf{x}_t\}$, no tiene sentido analizar ninguna de las partes individuales $\{x_{it}\}$ como una serie temporal univariante. Sin embargo, en algunos casos uno podría estar interesado en analizar el comportamiento relativo de dos partes i y j $(i \neq j)$, o, en general, de una serie temporal subcomposicional $\{\mathbf{x}_S\}$, donde S simboliza un subconjunto de dos o más de las partes $1, ..., D$ de $\mathbf{x}_t$.

Cuando se aplican las transformaciones clr, alr_k y ilr_V a un proceso composicional $\{\mathbf{x}_t\}$, inducen los procesos $\{\mathbf{z}_t\}$, $\{\mathbf{y}_t\}$ y $\{\mathbf{u}_t\}$, respectivamente. El primero, $\{\mathbf{z}_t\}$, definido en $\mathbb{R}^D$, está restringido al hiperplano V porque $\mathbf{z}_t^\top \mathbf{1}_D = 0$. Los otros dos procesos están definidos en $\mathbb{R}^{D-1}$, pero $\{\mathbf{y}_t\}$ depende del denominador utilizado en la transformación alr_k , y $\{\mathbf{u}_t\}$ depende de la matriz V utilizada en la transformación ilr_V. Denotamos por $\mu_{Z,t}$, $\mu_{Y,t}$ y $\mu_{U,t}$ los vectores de medias de $\{\mathbf{z}_t\}$, $\{\mathbf{y}_t\}$ y $\{\mathbf{u}_t\}$, respectivamente, y por $\Gamma_Z(t+h,t)$, $\Gamma_Y(t+h,t)$ y $\Gamma_U(t+h,t)$ las matrices de autocovarianza de estos procesos. Observa que $\mu_{Z,t} = clr(\xi_t)$ y, por definición, $\Gamma_Z(t+h,t) = \Gamma_C(t+h,t)$.

Modelo C-VAR(p)

Describe la dinámica de la composición actual $\mathbf{X}_t$ en función de sus propias composiciones pasadas y las de las demás variables composicionales en el sistema.

$$(\mathbf{X}_t \ominus \xi) \ominus (\Phi_{C,1} \odot (\mathbf{X}_{t-1} \ominus \xi)) \ominus \cdots \ominus (\Phi_{C,p} \odot (\mathbf{X}_{t-p} \ominus \xi)) = \mathbf{W}_t \tag{36}$$

Forma con operador de rezago composicional:

$$\Phi_C(L_C)(\mathbf{X}_t \ominus \xi) = \mathbf{W}_t \tag{37}$$

Donde:

- $\mathbf{X}_t \in \mathcal{S}^D$ es el vector composicional en el tiempo t.
- $\xi \in \mathcal{S}^D$ es el centro composicional.
- $\Phi_{C,i}$ son matrices de coeficientes composicionales de tamaño $D \times D$, para $i = 1, \ldots, p$.
- $\Phi_C(L_C) = I_D \ominus (\Phi_{C,1} \odot L_C) \ominus \cdots \ominus (\Phi_{C,p} \odot L_C^p)$ es un polinomio matricial en el operador de rezago composicional L_C y I_D es la matriz de identidad composicional.
- $\mathbf{W}_t \in \mathcal{S}^D$ es un ruido blanco composicional (WNC).

Modelo C-VMA(q)

Describe la composición actual $\mathbf{X}_t$ como una función de un término constante y errores pasados del proceso:

$$\mathbf{X}_t \ominus \xi = \mathbf{W}_t \ominus (\Theta_{C,1} \odot \mathbf{W}_{t-1}) \ominus \cdots \ominus (\Theta_{C,q} \odot \mathbf{W}_{t-q}) \tag{38}$$

Forma con operador de rezago composicional:

$$\mathbf{X}_t \ominus \xi = \Theta_C(L_C)\mathbf{W}_t \tag{39}$$

donde:

- $\mathbf{X}_t \in \mathcal{S}^D$ es el vector composicional en el tiempo t.
- $\xi \in \mathcal{S}^D$ es el centro composicional.
- $\Theta_{C,j}$ son matrices de coeficientes composicionales, $j = 1, \ldots, q$. -$\Theta_C(L_C) = I_D \ominus (\Theta_{C,1} \odot L_C) \ominus \cdots \ominus (\Theta_{C,q} \odot L_C^q)$ es un polinomio matricial composicional y I_D es la matriz de identidad composicional.
- $\mathbf{W}_t \in \mathcal{S}^D$ es un ruido blanco composicional (WNC).
- $\Theta_{C,j}$ son matrices de coeficientes composicionales, $j = 1, \ldots, q$.

Modelo C-VARMA(p,q)

Este modelo combina las características autorregresivas y de medias móviles en el espacio símplex. Es una combinación de los dos modelos anteriores.

$$(\mathbf{X}_t \ominus \xi) \ominus \sum_{i=1}^{p} (\Phi_{C,i} \odot (\mathbf{X}_{t-i} \ominus \xi)) = \mathbf{W}_t \ominus \sum_{j=1}^{q} (\Theta_{C,j} \odot \mathbf{W}_{t-j}) \tag{40}$$

(Aquí la suma se refiere a la aplicación de forma iterativa de la operación, es decir : $A \ominus B \ominus C = (A \ominus B) \ominus C$.

Forma con operador de rezago:

$$\Phi_C(L_C)(\mathbf{X}_t \ominus \xi) = \Theta_C(L_C)\mathbf{W}_t \tag{41}$$

donde:

- $\mathbf{X}_t \in \mathcal{S}^D$ es el vector composicional en el tiempo t.
- $\xi \in \mathcal{S}^D$ es el centro composicional.
- $\Phi_{C,i}$ son matrices de coeficientes composicionales de tamaño $D \times D$, para $i = 1, \dots, p$.
- $\Theta_{C,j}$ son matrices de coeficientes composicionales, $j = 1, \dots, q$.
- $\mathbf{W}_t \in \mathcal{S}^D$ es un ruido blanco composicional (WNC).
- $\Phi_C(L_C)$ y $\Theta_C(L_C)$ son los polinomios matriciales composicionales definidos anteriormente para C-VAR(p) y C-VMA(q), respectivamente.

Modelo C-VARIMA(p,d,q)

Extiende el C-VARMA para datos no estacionarios incluyendo diferenciación composicional.

$$\Phi_C(L_C)(1 - L_C)^d(\mathbf{X}_t \ominus \xi) = \Theta_C(L_C)\mathbf{W}_t \tag{42}$$

donde:

- $\mathbf{X}_t \in \mathcal{S}^D$ es el vector composicional en el tiempo t.
- $\xi \in \mathcal{S}^D$ es el centro composicional.
- $\Phi_{C,i}$ son matrices de coeficientes composicionales de tamaño $D \times D$, para $i = 1, \dots, p$.
- $\Theta_{C,j}$ son matrices de coeficientes composicionales, $j = 1, \dots, q$.
- $\mathbf{W}_t \in \mathcal{S}^D$ es un ruido blanco composicional (WNC).
- $\Phi_C(L_C)$ y $\Theta_C(L_C)$ son los polinomios matriciales composicionales definidos anteriormente para C-VAR(p) y C-VMA(q), respectivamente.
- $(1 - L_C)^d$ indica aplicar d veces la diferencia composicional: $\mathbf{X}_t \ominus \mathbf{X}_{t-1}$.Este operador transforma una serie no estacionaria en una estacionaria en el espacio símplex.

Estudio de Simulación

Introducción

El objetivo de este capítulo es llevar a cabo la simulación de una serie de tiempo en el contexto de datos composicionales. Para ello, se genera un conjunto de observaciones dentro del espacio del símplex utilizando la distribución Dirichlet, que garantiza que la suma de los componentes de cada vector de datos sea igual a uno. Esta característica es fundamental en el análisis composicional, donde las proporciones relativas y no los valores absolutos son de interés.

La distribución Dirichlet depende por un vector de parámetros α^3, cuya configuración tiene un impacto directo en la forma y dispersión de los datos generados. Dependiendo de los valores asignados de α, es posible obtener diferentes estructuras de dispersión dentro del símplex, como se observa en la Figura 1:

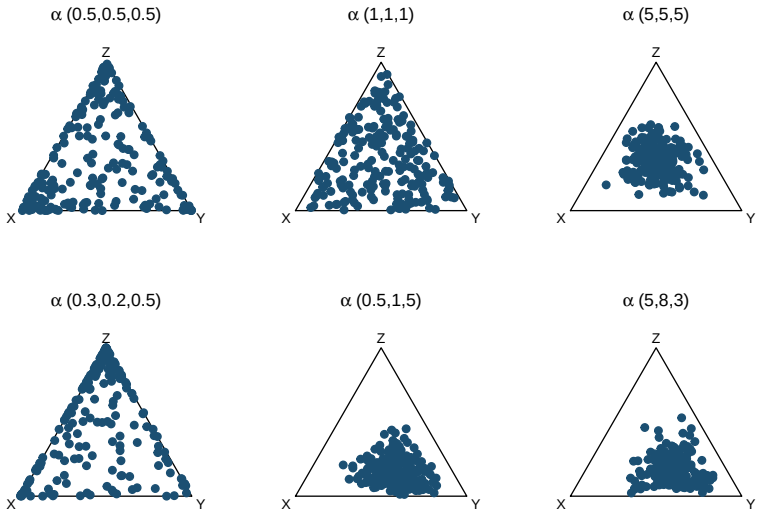

Figura 1: Comportamiento de los datos en una distribución Dirichlet con diferentes valores de Alpha

- Si α es un vector, son todos sus componentes > 1, la distribución tiende a concentrarse en el centro del símplex. Los vectores generados presentan una alta mezcla entre componentes, lo que da lugar a datos más balanceados y agrupados.
- Si α es un vector, son todos sus componentes $= 1$, la distribución se vuelve uniforme. En este caso, todas las combinaciones posibles de proporciones tienen la misma probabilidad de ocurrencia, lo que permite observar una dispersión homogénea dentro del espacio.

- En cambio, cuando el vector de $\alpha < 1$, la distribución se concentra en los vértices del símplex. Los datos tienden a ser extremos, mostrando combinaciones en las que uno o dos componentes predominan sobre el resto.

Cuando los parámetros α de una distribución Dirichlet son distintos entre sí, el comportamiento de los datos reflejará una asignación desigual de probabilidades entre los componentes. Si todos los valores son menores que 1, aunque diferentes, los datos tienden a ubicarse cerca de los vértices del símplex, lo que indica combinaciones extremas en las que un componente domina, pero no siempre el mismo. Esta configuración genera alta variabilidad y una distribución asimétrica. En cambio, si todos los α son mayores que 1 pero distintos, la distribución favorece composiciones más balanceadas, aunque no completamente uniformes: algunos componentes tienden a tener mayor presencia debido a los valores más altos de α. Finalmente, si la distribución combina valores mayores y menores que 1, se obtiene un comportamiento mixto. Los componentes con $\alpha > 1$ tienden a ser estables y con proporciones moderadas, mientras que los que tienen $\alpha < 1$ muestran mayor variabilidad y extremidad. Esto da lugar a composiciones con mezcla parcial, donde algunos elementos destacan por su presencia consistente y otros por su comportamiento más disperso o extremo.

Generación de Datos

Con el objetivo de evaluar el desempeño de los modelos aplicables al análisis de series de tiempo composicionales, se procedió a la generación de datos simulados controlando distintos factores estructurales. Los datos generados representan composiciones de tres partes (dimensión composicional $D = 3$) a lo largo de $T = 100$ unidades temporales, replicadas en $n_{\text{series}} = 100$ series independientes. Para garantizar la reproducibilidad del experimento, se fijó una semilla aleatoria mediante `(seed = 100 + i)`.

Especificación del Modelo

Las series generadas siguen una estructura autorregresiva con posibles componentes de media móvil en el espacio composicional. El modelo base se expresa como

$$\Phi_C(L_C)\,(1 - L_C)^d (X_t \ominus \xi) \;=\; \Theta_C(L_C)\,W_t, \tag{43}$$

donde:

- $X_t \in \mathbb{S}^D$ representa la composición observada en el tiempo t (símplex de dimensión 3, con componentes X, Y y Z).
- ξ es el centro composicional alrededor del cual se describe la dinámica; en este estudio se fija como la composición uniforme
$$\xi = \mathbf{1}_C = \left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right) \in \mathbb{S}^D.$$
- L_C es el operador de rezago en álgebra composicional.
- $\Phi_C(L_C)$ y $\Theta_C(L_C)$ son polinomios composicionales en L_C que recogen, respectivamente, los efectos autorregresivos y de media móvil; en la aplicación considerada se trabaja con polinomios de orden 1 (modelos C-VAR(1), C-VMA(1) y C-VARMA(1,1)).
- d es el orden de diferenciación composicional (en este estudio se fija $d = 0$).

- $\{W_t\}$ es un proceso de ruido blanco composicional en el símplex. En la simulación se especifica de forma paramétrica mediante composiciones independientes

$$W_t \sim \text{Dirichlet}(\alpha_W), \qquad \alpha_W = (1,1,1), \qquad t = 1, \dots, T,$$

de modo que

$$\mathbb{E}[W_t] = \left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right) = \xi,$$

Parámetros de simulación

Los valores utilizados para simular las series son los siguientes:

- **Número de series**: $n_{\text{series}} = 100$; se generan 100 trayectorias independientes por escenario.
- **Horizonte temporal**: $T = 100$; cada serie contiene 100 observaciones posteriores al periodo de *burn-in*.
- **Dimensión composicional**: $D = 3$; cada vector composicional tiene tres componentes (X, Y, Z) que representan partes de un total unitario.

$$W_t \sim \text{Dirichlet}(1,1,1),$$

con media igual al centro composicional $\xi = (1/3, 1/3, 1/3)$.

- **Modelo dinámico base**: según el escenario, la estructura temporal corresponde a un modelo C-VAR(1), C-VMA(1) o C-VARMA(1,1) composicional, con matrices de coeficientes autorregresivos y de media móvil diagonales.

Escenarios de simulación

Con el objetivo de evaluar el impacto conjunto de la dinámica temporal y de la estructura de dispersión composicional, se consideran nueve escenarios experimentales que combinan:

1. Tres niveles de concentración de las series X_t, representados mediante un parámetro de forma

$$\alpha^X \in \{(0.5, 0.5, 0.5),\ (1,1,1),\ (5,5,5)\},$$

que generan, respectivamente, composiciones más dispersas, de dispersión intermedia y más concentradas alrededor del centro composicional. En la implementación computacional, estos niveles se inducen sobre las trayectorias de X_t mediante una transformación de potencia composicional, manteniendo inalterado el proceso de ruido composicional $\{W_t\}$.

2. Tres estructuras dinámicas:
   - Modelos C-VAR(1), controlados por una matriz diagonal de coeficientes autorregresivos

$$\Phi_1 = \text{diag}(0.8,\ 0.5,\ 0.2),$$

   - Modelos C-VMA(1), con matriz diagonal de coeficientes de media móvil

$$\Theta_1^{(1)} = \text{diag}(0.8,\ 0.5,\ 0.2),$$

- Modelos C-VARMA(1,1), que combinan la matriz autorregresiva anterior con una matriz de media móvil alternativa

$$\Theta_1^{(2)} = \text{diag}(0.2,\, 0.5,\, 0.8).$$

La combinación de estos dos factores da lugar a los nueve escenarios resumidos en el Cuadro~2, donde se indica el nivel de concentración composicional asociado a α^X y los valores de los parámetros autorregresivos Φ y de media móvil Θ para cada caso.

Cuadro 2: Escenarios considerados para la simulación de series de tiempo composicionales. La columna α^X representa el nivel de concentración de las trayectorias de X_t, mientras que Φ y Θ recogen los coeficientes diagonales de los términos autorregresivos y de media móvil, respectivamente.

Escenario	α^X	Φ	Θ
1	$(0.5, 0.5, 0.5)$	$(0.8, 0.5, 0.2)$	$(0, 0, 0)$
2	$(0.5, 0.5, 0.5)$	$(0, 0, 0)$	$(0.8, 0.5, 0.2)$
3	$(0.5, 0.5, 0.5)$	$(0.8, 0.5, 0.2)$	$(0.2, 0.5, 0.8)$
4	$(1, 1, 1)$	$(0.8, 0.5, 0.2)$	$(0, 0, 0)$
5	$(1, 1, 1)$	$(0, 0, 0)$	$(0.8, 0.5, 0.2)$
6	$(1, 1, 1)$	$(0.8, 0.5, 0.2)$	$(0.2, 0.5, 0.8)$
7	$(5, 5, 5)$	$(0.8, 0.5, 0.2)$	$(0, 0, 0)$
8	$(5, 5, 5)$	$(0, 0, 0)$	$(0.8, 0.5, 0.2)$
9	$(5, 5, 5)$	$(0.8, 0.5, 0.2)$	$(0.2, 0.5, 0.8)$

En consecuencia, los escenarios 1, 4 y 7 corresponden a modelos C-VAR(1) puros, los escenarios 2, 5 y 8 a modelos C-VMA(1), y los escenarios 3, 6 y 9 a modelos C-VARMA(1,1), manteniendo siempre el mismo proceso de innovaciones $\{W_t\} \sim \text{Dirichlet}(1, 1, 1)$ y variando únicamente la estructura dinámica y el nivel de concentración composicional de las series simuladas X_t.

Transformaciones Composicionales

El análisis de datos composicionales ha dado lugar al desarrollo de un conjunto específico de herramientas y técnicas estadísticas que permiten tratar adecuadamente este tipo particular de datos. Estas composiciones, que representan proporciones o partes de un todo, requieren métodos analíticos distintos a los utilizados en estadística clásica, debido a la restricción inherente de la suma constante. Aplicar técnicas convencionales directamente sobre datos composicionales puede conducir a resultados erróneos o interpretaciones inadecuadas, ya que no se respetan las propiedades geométricas del espacio símplex en el que residen estos datos.

Para sortear estas limitaciones, se han propuesto diversas transformaciones que permiten mapear los datos composicionales desde el espacio símplex a un espacio euclídeo, en el cual las herramientas estadísticas tradicionales pueden aplicarse de forma válida. Entre estas transformaciones se encuentran la transformación log-ratio centrada (`clr`), la log-ratio aditiva (`alr`) y la log-ratio isométrica (`ilr`). La transformación `ilr` es una de las más utilizadas por sus propiedades geométricas deseables y su capacidad para preservar la distancia euclídea en el espacio transformado. Por otro lado, la transformación `alr` resulta especialmente útil en contextos donde se desea comparar cada componente con una parte de referencia fija, facilitando su interpretación en ciertos análisis.

En este capítulo se aplicaron dos transformaciones: la log-ratio isométrica (`ilr`) y la log-ratio aditiva (`alr`), ambas implementadas mediante las funciones `ilr()` y `alr()` del paquete **compositions** en `R`. La transformación `ilr` proyecta los datos composicionales en un espacio euclídeo de dimensión $D-1$, donde D representa el número de componentes de la composición, eliminando la redundancia y respetando la estructura geométrica del símplex. Las nuevas coordenadas generadas son ortogonales, lo que permite aplicar herramientas estadísticas multivariantes de forma rigurosa y coherente.

La transformación `alr`, en cambio, transforma las composiciones usando log-ratios respecto a una parte de referencia (usualmente la última componente), generando coordenadas interpretables como relaciones directas con dicha parte. Esta transformación es útil en casos donde una parte de la composición actúa como denominador natural o base de comparación.

No se empleó la transformación log-ratio centrada (`clr`) debido a que esta genera un conjunto de coordenadas que presentan una dependencia lineal intrínseca: la suma de todas las coordenadas transformadas es siempre cero. Esto implica que los datos `clr` no pertenecen a un subespacio euclídeo completo, y por tanto, no pueden ser utilizados directamente en modelos estadísticos multivariantes convencionales, como modelos c-VAR o C-VARMA. Además, en el contexto de series de tiempo, esta dependencia contamina las estructuras de autocorrelación y covarianza entre componentes, lo cual dificulta la especificación, interpretación y estimación de modelos dinámicos válidos.

El uso conjunto de las transformaciones `ilr` y `alr` en este estudio permitió comparar sus efectos en el modelado de series de tiempo composicionales simuladas, destacando las ventajas prácticas y analíticas de cada una en función del tipo de análisis, la interpretabilidad de los resultados y las propiedades geométricas preservadas.

Escenarios

Con el objetivo de evaluar el comportamiento dinámico de series de tiempo composicionales bajo diferentes condiciones de dependencia temporal y variabilidad composicional, se diseñó un estudio de simulación que combina procesos autorregresivos Φ, de medias móviles Θ y mixtos Φ y Θ con innovaciones provenientes de una distribución Dirichlet. En total, se definieron nueve escenarios experimentales, variando tanto los parámetros de autocorrelación como la concentración de la distribución Dirichlet. Esto permite observar el impacto conjunto de la dinámica temporal y la estructura composicional en el rendimiento de distintos modelos estadísticos aplicados a transformaciones log-ratio.

Cada escenario explora una combinación específica de intensidad de correlación temporal (baja, media y alta) y estructura de dispersión composicional (alta, media y baja concentración), replicando condiciones realistas que podrían encontrarse en datos composicionales.

Escenario 1

Este escenario simula composiciones altamente dispersas, generadas a partir de una distribución Dirichlet con parámetros bajos $\alpha = (0.5, 0.5, 0.5)$, lo que produce una fuerte variabilidad entre componentes.

A esta variabilidad se le incorpora una dinámica temporal autoregresiva (C-VAR) de orden 1, con distintos niveles de autocorrelación para cada componente: 0.8, 0.5 y 0.2.

Cuadro 3: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (0.5, 0.5, 0.5)$ y estructura C-VAR(1) con coeficientes $\Phi = 0.2$, 0.5 y 0.8.

	$\Phi = 0.2$		$\Phi = 0.5$		$\Phi = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.347334	0.347334	0.341191	0.341191	0.339508	0.339508
Y	0.345888	0.345888	0.340257	0.340257	0.335651	0.335651
Z	0.354964	0.354964	0.348135	0.348135	0.345376	0.345376

Cuadro 4: AIC por transformación, distribución Dirichlet con $\alpha = (0.5, 0.5, 0.5)$ y estructura C-VAR(1) con coeficientes $\Phi = 0.2$, 0.5 y 0.8.

Transformación	$\Phi = 0.2$	$\Phi = 0.5$	$\Phi = 0.8$
ILR	-3.34	-4.49	-5.33
ALR	-2.24	-3.39	-4.23

De acuerdo con el Cuadro 3, los valores de RMSE por componente (X, Y, Z) muestran una disminución sistemática a medida que aumenta Φ (por ejemplo, para X: $0.347334 \to 0.341191 \to 0.339508$), lo cual sugiere que, en este diseño, una mayor persistencia temporal permite una reconstrucción/predicción más precisa de las componentes composicionales. Además, la componente Z presenta consistentemente los RMSE más altos, indicando que es la fracción más difícil de recuperar bajo esta configuración.

Un resultado relevante es que los RMSE reportados para ILR y ALR son idénticos en todas las componentes y niveles de Φ (Cuadro 3), lo que indica que la precisión predictiva en el símplex se mantiene al cambiar la parametrización log-ratio en este escenario. Sin embargo, al comparar el ajuste con penalización por complejidad mediante AIC (Cuadro 4), ILR es sistemáticamente preferido, al mostrar valores más bajos en los tres casos $(-3.34, -4.49, -5.33)$ frente a ALR $(-2.24, -3.39, -4.23)$. En conjunto, esto implica que, aunque ambas transformaciones ofrecen un desempeño predictivo equivalente (RMSE), ILR logra un mejor equilibrio entre ajuste y simplicidad del modelo, especialmente cuando la dependencia temporal aumenta (AIC más favorable al pasar de $\Phi = 0.2$ a $\Phi = 0.8$).

Escenario 2

Se mantiene la distribución Dirichlet $\alpha = (0.5, 0.5, 0.5)$, con alta dispersión composicional, pero se reemplaza la dependencia autoregresiva por un proceso de medias móviles (C-VMA) de orden 1. Los coeficientes Θ para las componentes son 0.8, 0.5 y 0.2.

Cuadro 5: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (0.5, 0.5, 0.5)$ y estructura C-VMA(1) con coeficientes $\Theta = 0.2$, 0.5 y 0.8.

	$\Theta = 0.2$		$\Theta = 0.5$		$\Theta = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.348021	0.348021	0.343211	0.343178	0.340386	0.340388
Y	0.349964	0.349964	0.343950	0.343962	0.343537	0.343546
Z	0.349316	0.349316	0.346734	0.346753	0.345764	0.345757

Cuadro 6: AIC por transformación, distribución Dirichlet con $\alpha = (0.5, 0.5, 0.5)$ y estructura C-VMA(1) con coeficientes $\Theta = 0.2$, 0.5 y 0.8.

Transformación	$\Theta = 0.2$	$\Theta = 0.5$	$\Theta = 0.8$
ILR	-3.32	-4.19	-4.62
ALR	-2.22	-3.09	-3.53

Según el Cuadro 5, los RMSE por componente tienden a disminuir conforme aumenta Θ, reflejando que una mayor contribución del término MA favorece la capacidad del modelo para capturar la estructura temporal en el símplex. Por ejemplo, para X el error baja de 0.348021 a 0.343211 y luego a 0.340386; para Z la reducción también es sostenida ($0.349316 \to 0.346734 \to 0.345764$). En cuanto a la dificultad por componente, con $\Theta = 0.2$ la mayor discrepancia se observa en Y (0.349964), mientras que al aumentar Θ la componente Z pasa a concentrar los *RMSE* más elevados, sugiriendo que su reconstrucción sigue siendo relativamente más exigente cuando la dependencia de tipo MA es más intensa.

Al comparar transformaciones, los RMSE de ILR} y ALR son prácticamente iguales en todos los casos (las diferencias aparecen sólo a nivel de la cuarta o quinta cifra decimal), por lo que, en términos predictivos, el desempeño es esencialmente equivalente bajo esta estructura C-VMA(1) (Cuadro 5). No obstante, el AIC del Cuadro 6 favorece de manera consistente a ILR, con valores más bajos para $\Theta = 0.2, 0.5, 0.8$ ($-3.32, -4.19, -4.62$) frente a ALR ($-2.22, -3.09, -3.53$). En conjunto, este escenario confirma el mismo patrón: ILR ofrece un mejor compromiso entre ajuste y complejidad, aun cuando el error predictivo (RMSE) sea prácticamente indistinguible respecto a ALR.

Escenario 3

Este escenario combina dinámicas autoregresivas y de medias móviles, formando un proceso C-VARMA(1,1) con parámetros $\Phi = (0.8, 0.5, 0.2)$ y $\Theta = (0.2, 0.5, 0.8)$, aplicados a composiciones generadas desde una distribución Dirichlet $\alpha = (0.5, 0.5, 0.5)$. Permite observar cómo interactúan la memoria y el ruido en composiciones dispersas.

Cuadro 7: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (0.5, 0.5, 0.5)$ y estructura C-VARMA(1,1) con coeficientes $\Theta = 0.8 - \Phi = 0.2$, $\Theta = 0.5 - \Phi = 0.5$ y $\Theta = 0.2 - \Phi = 0.8$.

	$\Theta = 0.8 - \Phi = 0.2$		$\Theta = 0.5 - \Phi = 0.5$		$\Theta = 0.2 - \Phi = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.343253	0.342839	0.339871	0.339848	0.338404	0.338382
Y	0.341895	0.342317	0.341272	0.341305	0.339191	0.339232
Z	0.341637	0.341623	0.340438	0.340432	0.340624	0.340598

Cuadro 8: AIC por transformación, distribución Dirichlet con $\alpha = (0.5, 0.5, 0.5)$ y estructura C-VARMA(1,1) con coeficientes $\Theta = 0.8 - \Phi = 0.2$, $\Theta = 0.5 - \Phi = 0.5$ y $\Theta = 0.2 - \Phi = 0.8$.

Transformación	$\Theta = 0.8 - \Phi = 0.2$	$\Theta = 0.5 - \Phi = 0.5$	$\Theta = 0.2 - \Phi = 0.8$
ILR	-5.17	-5.52	-5.71
ALR	-4.07	-4.42	-4.61

De acuerdo con el Cuadro 7, los RMSE tienden a reducirse al pasar desde la combinación con menor persistencia autorregresiva hacia la de mayor Φ. En particular, la componente X muestra una caída clara del error ($\approx 0.3433 \to 0.3399 \to 0.3384$), y la componente Y también mejora de forma sostenida ($\approx 0.3419 \to 0.3413 \to 0.3392$), lo cual sugiere que, en este escenario, el incremento en la dependencia de tipo AR aporta una ganancia neta en precisión aun cuando Θ disminuya. Para Z, el error disminuye de la primera a la segunda combinación ($\approx 0.3416 \to 0.3404$) y luego presenta un leve repunte en la tercera (≈ 0.3406), aunque se mantiene por debajo del caso $\Theta = 0.8, \Phi = 0.2$; esto indica que la respuesta de Z es algo más sensible al balance entre los términos AR y MA.

En cuanto a la comparación entre transformaciones, los RMSE de ILR y ALR son muy similares en las tres combinaciones (diferencias sólo en cifras decimales finales), por lo que el desempeño predictivo es esencialmente equivalente (Cuadro 7). Sin embargo, el criterio AIC del Cuadro 8 favorece de manera consistente a ILR, con valores más bajos para $\Theta = 0.8, \Phi = 0.2$, $\Theta = 0.5, \Phi = 0.5$ y $\Theta = 0.2, \Phi = 0.8$

$(-5.17, -5.52, -5.71)$ frente a ALR $(-4.07, -4.42, -4.61)$. En consecuencia, aunque ambas transformaciones alcanzan prácticamente el mismo nivel de error (RMSE), ILR proporciona un ajuste global más favorable bajo penalización por complejidad, especialmente en la configuración con mayor persistencia autorregresiva ($\Phi = 0.8$).

Escenario 4

La distribución Dirichlet se fija en $\alpha = (1, 1, 1)$, lo que corresponde a una dispersión composicional media. Se mantiene una estructura autoregresiva C-VAR(1) con coeficientes $(0.8, 0.5, 0.2)$, permitiendo comparar los efectos del cambio en la concentración de la composición frente al Escenario 1.

Cuadro 9: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (1, 1, 1)$ y estructura C-VAR(1) con coeficientes $\Phi = 0.2$, 0.5 y 0.8.

	$\Phi = 0.2$		$\Phi = 0.5$		$\Phi = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.384613	0.384613	0.368069	0.368069	0.357556	0.357556
Y	0.376392	0.376392	0.359688	0.359688	0.350051	0.350051
Z	0.381517	0.381517	0.368236	0.368236	0.361980	0.361980

Cuadro 10: AIC por transformación, distribución Dirichlet con $\alpha = (1, 1, 1)$ y estructura C-VAR(1) con coeficientes $\Phi = 0.2$, 0.5 y 0.8.

Transformación	$\Phi = 0.2$	$\Phi = 0.5$	$\Phi = 0.8$
ILR	-0.6	-1.75	-2.58
ALR	0.5	-0.65	-1.48

Con base en el Cuadro 9, los RMSE por componente (X, Y, Z) presentan una disminución clara a medida que aumenta Φ, lo que sugiere que una mayor persistencia autorregresiva contribuye a estabilizar la trayectoria y mejorar la precisión del ajuste/predicción en el símplex. Por ejemplo, para X el RMSE baja de 0.384613 a 0.368069 y luego a 0.357556; para Y de 0.376392 a 0.359688 y a 0.350051; y para Z de 0.381517 a 0.368236 y a 0.361980. Además, en la mayoría de los casos la componente X exhibe los mayores errores, indicando que, bajo esta configuración, resulta la parte más difícil de reconstruir con precisión.

Al comparar estos resultados con los escenarios anteriores (con $\alpha = (0.5, 0.5, 0.5)$), se observa que los RMSE del presente escenario son más elevados en general (Cuadro 9), evidenciando que el cambio en

la estructura de variabilidad composicional inducida por la Dirichlet también impacta el desempeño del modelado dinámico, aun manteniendo la misma estructura C-VAR(1). En cuanto a las transformaciones, nuevamente se aprecia que ILR y ALR producen RMSE idénticos en cada componente y nivel de Φ, lo que indica equivalencia práctica en precisión predictiva bajo este escenario.

No obstante, el AIC del Cuadro 10 favorece sistemáticamente a ILR, con valores más bajos $(-0.6, -1.75, -2.58)$ frente a ALR $(0.5, -0.65, -1.48)$ para $\Phi = 0.2, 0.5, 0.8$, respectivamente. En conjunto, el Escenario 4 reafirma el patrón observado: aunque ILR y ALR muestran precisión comparable (RMSE), ILR ofrece un mejor desempeño global al considerar el criterio de información, especialmente cuando la dependencia temporal aumenta.

Escenario 5

Con la misma distribución Dirichlet $\alpha = (1, 1, 1)$, este escenario implementa un modelo C-VMA(1) con coeficientes $(0.8, 0.5, 0.2)$. Analiza cómo las fluctuaciones de corto plazo afectan composiciones más equilibradas en comparación con el Escenario 2.

Cuadro 11: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (1, 1, 1)$ y estructura C-VMA(1) con coeficientes $\Theta = 0.2$, 0.5 y 0.8.

	$\Theta = 0.2$		$\Theta = 0.5$		$\Theta = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.380640	0.380640	0.369610	0.369612	0.363295	0.362984
Y	0.377645	0.377645	0.369348	0.369353	0.370654	0.370793
Z	0.387118	0.387118	0.371822	0.371820	0.365652	0.365888

Cuadro 12: AIC por transformación, distribución Dirichlet con $\alpha = (1, 1, 1)$ y estructura C-VMA(1) con coeficientes $\Theta = 0.2$, 0.5 y 0.8.

Transformación	$\Theta = 0.2$	$\Theta = 0.5$	$\Theta = 0.8$
ILR	-0.55	-1.42	-1.86
ALR	0.55	-0.32	-0.76

Conforme al Cuadro 11, los RMSE por componente muestran, en general, una reducción cuando Θ aumenta, lo cual indica que una estructura MA más intensa contribuye a capturar mejor la dependencia temporal y a disminuir el error de estimación/predicción en el símplex. En particular, la componente

X mejora de forma sostenida ($0.380640 \to 0.369610 \to 0.363295$), mientras que Z también presenta un descenso marcado ($0.387118 \to 0.371822 \to 0.365652$). En cambio, para Y la disminución no es estrictamente monótona: cae de 0.377645 a 0.369348 cuando Θ pasa a 0.5, pero aumenta ligeramente en $\Theta = 0.8$ (0.370654), sugiriendo una sensibilidad diferencial de esta componente frente a incrementos altos del efecto MA.

En cuanto a la comparación entre transformaciones, los RMSE de ILR y ALR son prácticamente idénticos en todos los niveles de Θ, con diferencias únicamente en los últimos decimales (Cuadro 11). Esto reafirma que, desde el punto de vista de precisión (RMSE), el rendimiento es esencialmente equivalente al cambiar la parametrización log-ratio. Sin embargo, el criterio AIC del Cuadro 12 favorece sistemáticamente a ILR: para $\Theta = 0.2, 0.5, 0.8$ se obtienen $-0.55, -1.42, -1.86$, frente a ALR con $0.55, -0.32, -0.76$. Por tanto, aunque ambas transformaciones entregan errores comparables, ILR mantiene un mejor desempeño global al considerar el balance entre ajuste y complejidad del modelo bajo la dinámica C-VMA(1).

Escenario 6

Este escenario considera una distribución Dirichlet $(1, 1, 1)$ junto con una estructura C-VARMA(1,1) con parámetros $\Phi = (0.8, 0.5, 0.2)$ y $\Theta = (0.2, 0.5, 0.8)$. Representa una situación intermedia tanto en complejidad temporal como en dispersión composicional.

Cuadro 13: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (1, 1, 1)$ y estructura C-VARMA(1,1) con coeficientes $\Theta = 0.8 - \Phi = 0.2$, $\Theta = 0.5 - \Phi = 0.5$ y $\Theta = 0.2 - \Phi = 0.8$.

	$\Theta = 0.8 - \Phi = 0.2$		$\Theta = 0.5 - \Phi = 0.5$		$\Theta = 0.2 - \Phi = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.357649	0.357979	0.350357	0.350409	0.348071	0.348046
Y	0.371843	0.371760	0.370316	0.370327	0.369458	0.369398
Z	0.355584	0.355444	0.352597	0.352501	0.347496	0.347545

Cuadro 14: AIC por transformación, distribución Dirichlet con $\alpha = (1, 1, 1)$ y estructura C-VARMA(1,1) con coeficientes $\Theta = 0.8 - \Phi = 0.2$, $\Theta = 0.5 - \Phi = 0.5$ y $\Theta = 0.2 - \Phi = 0.8$.

Transformación	$\Theta = 0.8 - \Phi = 0.2$	$\Theta = 0.5 - \Phi = 0.5$	$\Theta = 0.2 - \Phi = 0.8$
ILR	-2.38	-2.74	-2.93
ALR	-1.28	-1.64	-1.83

De acuerdo con el Cuadro 13, los RMSE tienden a disminuir al pasar hacia configuraciones con mayor persistencia autorregresiva Φ, aun cuando Θ se reduzca. En particular, la componente X mejora de manera sostenida ($0.357649 \rightarrow 0.350357 \rightarrow 0.348071$), y la componente Y también presenta una reducción progresiva ($0.371843 \rightarrow 0.370316 \rightarrow 0.369458$). Para Z, el error es menor en la tercera combinación ($\Theta = 0.2, \Phi = 0.8$), al pasar de 0.355584 a 0.352597 y luego a 0.347496, lo que sugiere que, en este escenario, el incremento de Φ aporta una ganancia clara en precisión para todas las componentes.

Al comparar transformaciones, los resultados de ILR y ALR son muy similares en términos de RMSE, con diferencias pequeñas en los últimos decimales (Cuadro 13), por lo que la precisión predictiva se mantiene prácticamente inalterada al cambiar la parametrización log-ratio. Sin embargo, el AIC del Cuadro 14 favorece consistentemente a ILR, con valores más bajos en las tres combinaciones ($-2.38, -2.74, -2.93$) frente a ALR ($-1.28, -1.64, -1.83$). En conjunto, este escenario confirma que, aunque el desempeño predictivo (RMSE) entre transformaciones es casi equivalente, ILR ofrece un mejor balance global entre ajuste y complejidad, especialmente en la combinación con mayor dependencia AR ($\Phi = 0.8$).

Escenario 7

Aquí se modelan composiciones con baja dispersión, generadas desde una distribución Dirichlet $\alpha = (5, 5, 5)$. Se aplica un proceso C-VAR(1) con coeficientes $(0.8, 0.5, 0.2)$. Este diseño permite observar cómo la homogeneidad composicional modula los efectos de la autocorrelación en series de tiempo.

Cuadro 15: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (5, 5, 5)$ y estructura C-VAR(1) con coeficientes $\Phi = 0.2$, 0.5 y 0.8.

	$\Phi = 0.2$		$\Phi = 0.5$		$\Phi = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.522103	0.522102	0.505824	0.505824	0.483532	0.483531
Y	0.485132	0.485132	0.475796	0.475796	0.467401	0.467404
Z	0.488645	0.488645	0.466301	0.466301	0.448524	0.448522

Cuadro 16: AIC por transformación, distribución Dirichlet con $\alpha = (5, 5, 5)$ y estructura C-VAR(1) con coeficientes $\Phi = 0.2$, 0.5 y 0.8.

Transformación	$\Phi = 0.2$	$\Phi = 0.5$	$\Phi = 0.8$
ILR	5.84	4.68	3.85
ALR	6.94	5.78	4.95

Según el Cuadro 15, los RMSE por componente presentan una disminución marcada al incrementar Φ, evidenciando que una mayor persistencia autorregresiva contribuye a mejorar la precisión del ajuste/predicción. Por ejemplo, para X el error baja de 0.522103 a 0.505824 y luego a 0.483532; para Y desciende de 0.485132 a 0.475796 y a 0.467401; y para Z se reduce de 0.488645 a 0.466301 y a 0.448524. En términos comparativos, X muestra consistentemente los *RMSE* más altos, sugiriendo que esta componente continúa siendo la más difícil de reconstruir bajo la estructura C-VAR(1)} y esta configuración de variabilidad composicional.

Un aspecto importante de este escenario es que los RMSE son más elevados que en los escenarios con $\alpha = (0.5, 0.5, 0.5)$ y $\alpha = (1, 1, 1)$ (Cuadro 15), lo que indica que, aun con una Dirichlet más concentrada, el error global aumenta bajo esta calibración del experimento (posiblemente por la menor amplitud efectiva de variación y la forma en que el error se está midiendo por componente). En cuanto a la comparación entre transformaciones, ILR y ALR producen RMSE prácticamente idénticos para cada componente y nivel de Φ, manteniéndose la equivalencia en precisión predictiva.

Sin embargo, el criterio AIC del Cuadro 16 favorece de forma consistente a ILR, con valores inferiores en los tres niveles de Φ: 5.84, 4.68, 3.85 frente a 6.94, 5.78, 4.95 para ALR. Además, el AIC mejora (disminuye) conforme aumenta Φ, reforzando que una mayor dependencia temporal facilita un ajuste más eficiente del modelo. En conjunto, el Escenario 7 confirma que, aunque el desempeño predictivo entre ILR y ALR es muy similar en términos de RMSE, ILR mantiene ventaja al considerar criterios de información.

Escenario 8

Bajo la misma distribución Dirichlet $\alpha = (5, 5, 5)$, se implementa un modelo C-VMA(1) con coeficientes $(0.8, 0.5, 0.2)$. Representa un entorno con perturbaciones breves aplicadas sobre composiciones estables y poco variables.

Cuadro 17: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (5, 5, 5)$ y estructura C-VMA(1) con coeficientes $\Theta = 0.2$, 0.5 y 0.8.

	$\Theta = 0.2$		$\Theta = 0.5$		$\Theta = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.508688	0.508688	0.495047	0.495200	0.491102	0.490888
Y	0.508390	0.508390	0.488898	0.488774	0.469367	0.469978
Z	0.482718	0.482718	0.468612	0.468394	0.457941	0.457933

Cuadro 18: AIC por transformación, distribución Dirichlet con $\alpha = (5, 5, 5)$ y estructura C-VMA(1) con coeficientes $\Theta = 0.2$, 0.5 y 0.8.

Transformación	$\Theta = 0.2$	$\Theta = 0.5$	$\Theta = 0.8$
ILR	5.93	5.04	4.61
ALR	7.03	6.14	5.70

Los resultados del Cuadro 17 evidencian que, al incrementar Θ, el RMSE tiende a disminuir en las tres componentes, lo que sugiere que una dependencia MA más marcada ayuda a capturar mejor la dinámica temporal y reduce el error. Este patrón es especialmente notorio en Y ($0.508390 \rightarrow 0.488898 \rightarrow 0.469367$) y en Z ($0.482718 \rightarrow 0.468612 \rightarrow 0.457941$), mientras que en X la reducción es más suave ($0.508688 \rightarrow 0.495047 \rightarrow 0.491102$), manteniéndose como la componente con mayor dificultad relativa en buena parte de los casos.

Respecto a la elección de transformación, ILR y ALR generan errores muy similares y las diferencias observadas son pequeñas (Cuadro 17). De hecho, en $\Theta = 0.8$ se aprecia un leve intercambio: ALR resulta ligeramente menor en X y Z, mientras que ILR es apenas menor en Y, sin que esto implique cambios sustantivos en el desempeño predictivo.

En contraste, al evaluar el ajuste penalizado por complejidad mediante AIC (Cuadro 18), se mantiene una preferencia clara por ILR, con valores consistentemente más bajos para $\Theta = 0.2, 0.5, 0.8$ (5.93, 5.04, 4.61) frente a ALR (7.03, 6.14, 5.70). Asimismo, el AIC mejora al aumentar Θ, lo que refuerza la idea de que una mayor dependencia MA se traduce en modelos más eficientes en términos de ajuste global.

Escenario 9

Finalmente, se simulan composiciones con baja dispersión composicional $\alpha = (5, 5, 5)$, bajo un modelo ARMA(1,1) con parámetros $\Phi = (0.8, 0.5, 0.2)$ y $\Theta =(0.2, 0.5, 0.8)$. Este escenario refleja una dinámica compleja en condiciones composicionales altamente homogéneas.

Cuadro 19: Medidas de error RMSE por componente (X, Y, Z), distribución Dirichlet con $\alpha = (5, 5, 5)$ y estructura C-VARMA(1,1) con coeficientes $\Theta = 0.8 - \Phi = 0.2$, $\Theta = 0.5 - \Phi = 0.5$ y $\Theta = 0.2 - \Phi = 0.8$.

	$\Theta = 0.8 - \Phi = 0.2$		$\Theta = 0.5 - \Phi = 0.5$		$\Theta = 0.2 - \Phi = 0.8$	
Componente	ILR	ALR	ILR	ALR	ILR	ALR
X	0.462786	0.463182	0.467962	0.467595	0.468831	0.469085
Y	0.444210	0.443500	0.433767	0.434693	0.434485	0.434853
Z	0.489331	0.489176	0.485013	0.484263	0.478125	0.477636

Cuadro 20: AIC por transformación, distribución Dirichlet con $\alpha = (5, 5, 5)$ y estructura C-VARMA(1,1) con coeficientes $\Theta = 0.8 - \Phi = 0.2$, $\Theta = 0.5 - \Phi = 0.5$ y $\Theta = 0.2 - \Phi = 0.8$.

Transformación	$\Theta = 0.8 - \Phi = 0.2$	$\Theta = 0.5 - \Phi = 0.5$	$\Theta = 0.2 - \Phi = 0.8$
ILR	4.08	3.7	3.51
ALR	5.18	4.8	4.60

De acuerdo con el Cuadro 19, el patrón de error no es completamente uniforme entre componentes: mientras que Z mejora de forma clara al desplazarse hacia mayor persistencia autorregresiva ($0.489331 \to 0.485013 \to 0.478125$ en ILR), la componente X muestra un leve incremento del RMSE al pasar de $(0.8, 0.2)$ hacia $(0.2, 0.8)$ ($0.462786 \to 0.467962 \to 0.468831$ en ILR). Para Y, el error disminuye notablemente en la combinación intermedia $(\Theta, \Phi) = (0.5, 0.5)$ ($0.444210 \to 0.433767$) y luego presenta un aumento marginal en $(0.2, 0.8)$ (0.434485), lo que sugiere que, para esta componente, el equilibrio entre términos AR y MA puede resultar más favorable que un predominio marcado de Φ.

Al comparar transformaciones, ILR y ALR exhiben RMSE muy cercanos en las tres combinaciones (Cuadro 19), sin diferencias que cambien la lectura sustantiva del desempeño predictivo. No obstante, el criterio AIC del Cuadro 20 favorece consistentemente a ILR, con valores inferiores en todos los casos (4.08, 3.70, 3.51) frente a ALR (5.18, 4.80, 4.60). Además, el AIC mejora al avanzar hacia mayor Φ, lo que respalda que, en términos de ajuste penalizado, la configuración con mayor persistencia autorregresiva resulta más eficiente bajo esta estructura C-VARMA(1,1).

Resultados clave

En conjunto, los nueve escenarios muestran que el incremento de la dependencia temporal tiende a mejorar el desempeño: tanto en estructuras C-VAR(1) como C-VMA(1), al aumentar Φ o Θ se observa, en general, una reducción del RMSE y una mejora del AIC (Cuadros 3-6, 9-12 y 15-18). En los escenarios mixtos C-VARMA(1,1) (Cuadros 7-8, 13-14 y 19-20), el desempeño suele ser competitivo y frecuentemente superior al de los modelos puros; en particular, las combinaciones con mayor Φ tienden a presentar mejores criterios de información, aunque el efecto en RMSE puede variar por componente (como se aprecia en el Escenario 9 para X y Y).

Respecto a la estructura de dispersión composicional inducida por la Dirichlet, se observa un cambio sistemático en los niveles de error: al pasar de $\alpha = (0.5, 0.5, 0.5)$ a $\alpha = (1, 1, 1)$ y luego a $\alpha = (5, 5, 5)$, los RMSE reportados tienden a ser mayores en los escenarios con α más alto (por ejemplo, los escenarios 7-9 exhiben RMSE claramente superiores a 1-3), lo que indica que la variabilidad composicional asociada a cada configuración de α influye de manera directa en la dificultad del ajuste y la predicción. Paralelamente, el AIC también refleja este patrón: los escenarios con $\alpha = (0.5, 0.5, 0.5)$ presentan valores más favorables (más bajos) que los escenarios con $\alpha = (1, 1, 1)$, y estos a su vez son más favorables que los de $\alpha = (5, 5, 5)$.

Finalmente, al comparar ILR y ALR, se mantiene una regularidad importante: los RMSE son idénticos o prácticamente indistinguibles en todos los escenarios y configuraciones, lo que sugiere que la precisión predictiva en el símplex es robusta a la elección de la transformación bajo este diseño de simulación. Sin embargo, el AIC favorece de manera consistente a ILR en los nueve escenarios, indicando un mejor

balance entre calidad de ajuste y complejidad del modelo. En síntesis, los resultados apoyan que (i) mayor dependencia temporal suele mejorar el desempeño, (ii) las estructuras C-VARMA tienden a ofrecer ventajas frente a modelos puramente AR o MA, y (iii) aunque ILR y ALR rinden de forma muy similar en RMSE, ILR es preferible cuando se prioriza el ajuste penalizado (AIC).

Estudio de Caso Real

Análisis Exploratorio

Este apartado presenta un análisis descriptivo de la estructura de la Formación Bruta de Capital Fijo (FBCF) en España durante el periodo 1850–2023. Para ello, se utilizan las series históricas elaboradas por Prados de la Escosura (2017), que permiten examinar la evolución de los principales componentes de la inversión: viviendas, maquinaria y equipos, equipos de transporte y otras construcciones.

Dado que estos datos representan partes de un todo es decir, proporciones que componen el 100 % de la inversión total en cada año se adopta un enfoque basado en el análisis de datos composicionales (CoDa). Esta metodología resulta especialmente adecuada para este tipo de información, ya que evita errores comunes derivados de aplicar técnicas estadísticas tradicionales a datos con restricción de suma constante, como las correlaciones espurias.

Cada observación anual puede interpretarse como una composición $\mathbf{C}_t$ de $D = 4$ partes, que describen cómo se distribuye la inversión fija bruta en ese año. El vector composicional para un año t se expresa como:

$$\mathbf{C}_t = (c_{1t}, c_{2t}, c_{3t}, c_{4t}), \quad \text{donde} \quad \sum_{i=1}^{4} c_{it} = 1. \tag{44}$$

Características de los datos

- **Periodicidad**: Datos con frecuencia anual.
- **Cobertura temporal**: Desde 1850 hasta 2023.
- **Componentes**: Viviendas, maquinaria y equipos, equipos de transporte y otras construcciones, expresados como porcentajes que suman 100 % cada año.
- **Formato composicional**: Cada observación es una composición cerrada adecuada para análisis CoDa.
- **Calidad y completitud**: La serie histórica no presenta datos faltantes ni inconsistencias documentadas, lo que garantiza un análisis continuo y fiable.
- **Fuente de los datos**:
  Prados de la Escosura, Leandro (2017). *La economía española en perspectiva histórica*. Fundación Rafael del Pino.
  Archivo Excel: Hoja "Cuadro 8". Disponible en: https://frdelpino.es/investigacion/economia-espanola/economia-espanola-en-perspectiva-historica/

Descarga directa: Hoja “Cuadro 8” https://frdelpino.es/investigacion/wp-content/uploads/2025/01/Contabilidad_Nacional_Historica_de_Espana_1850-2023_frdp.01.2025.xlsx

El objetivo central es obtener pronósticos para una serie de tiempo de naturaleza composicional. Su análisis estadístico requiere técnicas especializadas que respeten su estructura inherente. Para ello, se ha utilizado el enfoque VARIMA (Vector Autoregressive Integrated Moving Average), el cual permite modelar dinámicas multivariantes.

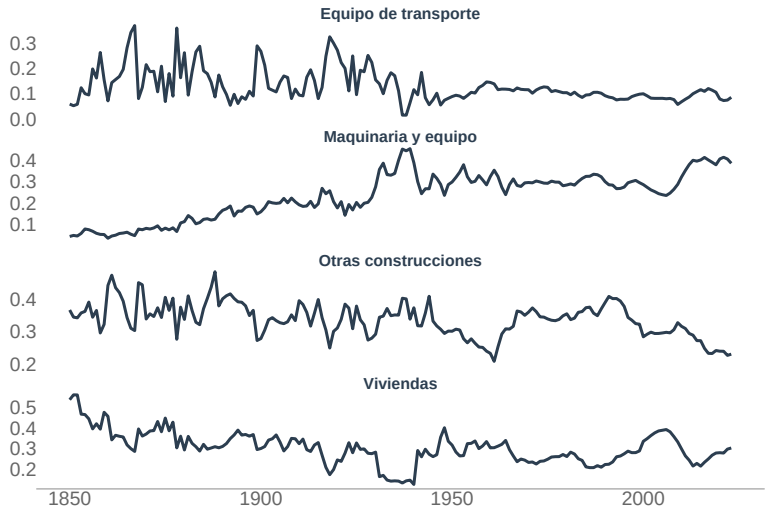

Figura 2: Evolución temporal de viviendas, maquinaria y equipos,equipos de transporte y otras construcciones

El procedimiento de pronóstico se ha realizado separando el conjunto de datos original en dos subconjuntos: uno de entrenamiento y otro de prueba. Esta segmentación es crucial para garantizar la validez del proceso de evaluación del modelo. Una vez definidos los subconjuntos, se aplican las transformaciones necesarias para adecuar los datos composicionales al análisis multivariado. En primer lugar, los datos originales son convertidos en una serie temporal a través de la función `ts()` de R, estableciendo como punto de inicio el año 1850 y una frecuencia anual `(frequency = 1)`. Esta transformación es útil para visualizar la evolución temporal general del conjunto completo de datos, aunque no es adecuada para el análisis estadístico directo debido a la estructura composicional. Por ello, el siguiente paso consiste en aplicar la transformación ilr (isometric log-ratio), que convierte las composiciones en coordenadas euclidianas sin perder información relativa.

La Figura 2 muestra la evolución temporal de la composición de la Formación Bruta de Capital Fijo (FBCF) en España desde 1850 hasta la actualidad, desagregada en cuatro componentes principales: viviendas, otras construcciones, maquinaria y equipo, y equipo de transporte. Se observa que el componente de **viviendas** presenta una tendencia general decreciente en su participación relativa desde mediados del siglo XIX, con algunos repuntes en las últimas décadas.

El componente de **otras construcciones** ha mantenido una participación relativamente estable, aunque con una ligera tendencia descendente en los últimos años. En contraste, la participación de **maquinaria y equipo** muestra un crecimiento sostenido a lo largo del tiempo, reflejando un proceso de industrialización y modernización de la economía. Por su parte, el **equipo de transporte** experimenta una alta volatilidad, especialmente durante las primeras décadas del siglo XX, y se estabiliza posteriormente en niveles intermedios.

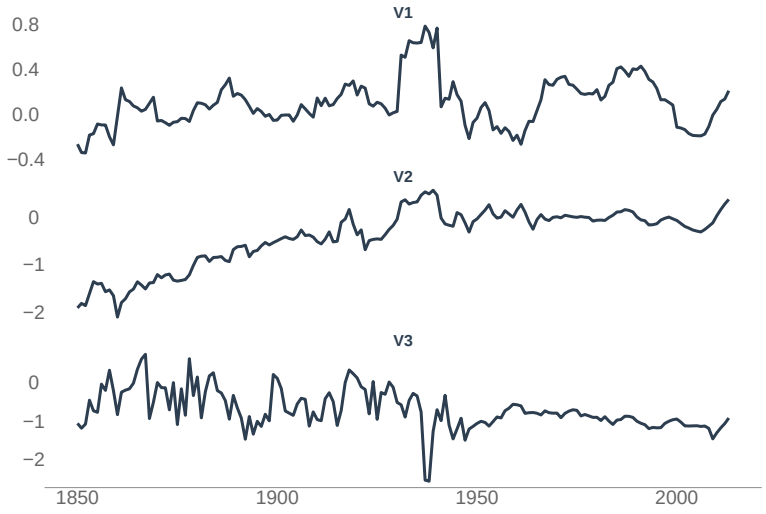

Figura 3: Evolución temporal de la transformación ilr

La Figura 3 presenta la evolución temporal de las coordenadas obtenidas tras aplicar la transformación isométrica log-ratio (ilr) al conjunto de entrenamiento. Cada una de las series representadas (V1, V2 y V3) corresponde a una combinación ortogonal de los componentes composicionales originales, proyectada en un espacio euclídeo. Estas nuevas variables no tienen una interpretación directa en términos económicos individuales, pero permiten modelar las relaciones relativas entre los componentes respetando la estructura composicional de los datos.

Se observan dinámicas diferenciadas en cada coordenada ilr, lo que sugiere patrones subyacentes complejos en la evolución relativa de los componentes de la FBCF. En particular, destacan ciertos cambios bruscos en las coordenadas V1 y V3, especialmente alrededor de los años 1936–1940 y 1950, posiblemente asociados a eventos históricos relevantes que afectaron la estructura de inversión del país. Estas series transformadas constituyen la base para el ajuste del modelo VARIMA, ya que presentan propiedades estadísticas más adecuadas para este tipo de modelado multivariado.

Ajuste del Modelo y Validación Estadística

Una vez transformada la serie composicional mediante log-ratios isométricos (ilr), el siguiente paso en el análisis es la identificación del modelo de series temporales multivariado adecuado para describir la dinámica de las coordenadas transformadas. Para ello, se sigue el procedimiento clásico en el análisis de series de tiempo, adaptado al contexto multivariado, comenzando por el estudio de la estacionariedad de las series transformadas.

En esta etapa, se aplica la prueba de raíz unitaria Dickey-Fuller (ADF) a cada una de las coordenadas resultantes de la transformación ilr. Esta prueba permite determinar si las series presentan una raíz unitaria, es decir, si son no estacionarias en nivel. La estacionariedad es una condición fundamental para aplicar modelos VARIMA, ya que garantiza que las propiedades estadísticas de las series como la media y la varianza sean constantes a lo largo del tiempo. En caso de que alguna coordenada no sea estacionaria, se procederá a su diferenciación hasta alcanzar la estacionariedad.

El Cuadro 21 presenta los resultados de la prueba de raíz unitaria Dickey-Fuller (ADF) aplicada a las tres coordenadas resultantes de la transformación ilr. Esta prueba permite evaluar la estacionariedad

de cada serie mediante el contraste de hipótesis nula de presencia de raíz unitaria (no estacionariedad). Como se puede observar, las dos primeras coordenadas presentan valores-p superiores al umbral del 5 %, por lo que no se rechaza la hipótesis nula y, en consecuencia, se concluye que estas series no son estacionarias. En cambio, la tercera coordenada muestra un valor-p de 0.01, lo que indica evidencia estadística suficiente para rechazar la hipótesis nula y considerar esta serie como estacionaria.

Cuadro 21: Resultados del Test ADF

Statistic	P_value	Stationary
-3.374	0.0613	No
-2.173	0.5043	No
-4.076	0.0100	Sí

Dado que la estacionariedad es una condición fundamental para ajustar modelos VARIMA, se procederá a diferenciar aquellas coordenadas que no cumplen con este requisito. Este proceso de diferenciación transformará las series no estacionarias en series estacionarias en primera diferencia, garantizando así que todas las coordenadas cumplan con los supuestos del modelo. Una vez realizadas las transformaciones necesarias, se continuará con la identificación y estimación del modelo VARIMA sobre las coordenadas estacionarias.

Cuadro 22: Resultados del Test ADF

Statistic	P_value	Stationary
-4.427	0.01	Sí
-6.760	0.01	Sí
-8.024	0.01	Sí

Autocorrelación de las variables transformadas:

La Figura 4 muestra las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de las tres coordenadas $V1$, $V2$ y $V3$ obtenidas tras aplicar la transformación isométrica log-ratio (ilr) a la serie composicional y, posteriormente, la diferenciación con el objetivo de favorecer la estacionariedad.

En la primera coordenada (V1) no se aprecia una señal dominante concentrada exclusivamente en el rezago 1; más bien, los primeros rezagos presentan valores relativamente pequeños alrededor de cero. Sin embargo, destaca un pico negativo aislado alrededor del rezago 10–11, visible en la ACF y acompañado por una señal también apreciable en la PACF. Este patrón puede interpretarse como indicio de dependencia a un horizonte más largo, compatible con un comportamiento recurrente o con algún tipo de periodicidad de media frecuencia. Aun así, al tratarse de una señal puntual, su lectura debe realizarse con cautela, ya que en muestras finitas (y especialmente tras diferenciar) pueden aparecer picos aislados por variabilidad muestral. Por ello, más que fijar directamente un rezago autorregresivo

específico, resulta preferible considerarlo como una guía para incluir en el conjunto de modelos candidatos especificaciones capaces de capturar dependencia a ese horizonte, verificando su pertinencia en etapas posteriores mediante criterios de información y diagnósticos de residuos. Para la segunda coordenada (V2), el comportamiento es relativamente estable en los primeros rezagos, sin una estructura claramente persistente de corto plazo. En general, la ACF oscila alrededor de cero y la PACF no exhibe un corte nítido en rezagos bajos que permita concluir de forma inequívoca una dinámica AR(1) simple. Por su parte, la tercera coordenada (V3) presenta el patrón más distintivo. En la ACF se observan varios rezagos cortos (aproximadamente entre 1 y 3) con autocorrelaciones negativas que sobrepasan las bandas de significancia, y la PACF refuerza esta evidencia al mostrar picos negativos pronunciados también en los primeros rezagos.

En conjunto, estos resultados respaldan la necesidad de considerar una estructura VARIMA que incorpore distintos niveles de memoria temporal para cada componente transformada. La primera componente podría requerir un rezago largo , mientras que la segunda puede ser representada por una estructura más parsimoniosa , y la tercera justificaría un rezago intermedio. Así, una especificación inicial con orden autorregresivo máximo igual a 10 permite capturar toda la dinámica relevante detectada en las funciones ACF y PACF, siendo recomendable posteriormente contrastar diferentes configuraciones mediante criterios de información como el AIC o BIC.

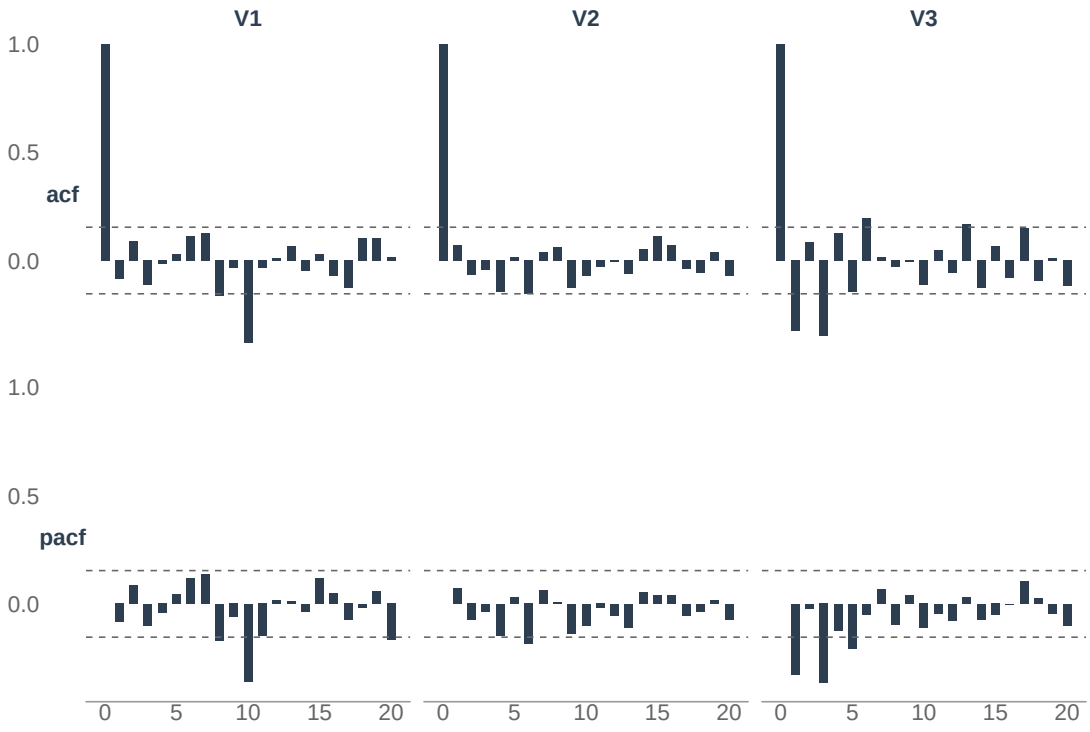

Figura 4: ACF y PACF de la variables transformadas

Para determinar la estructura óptima del modelo, se aplicó a la serie transformada la función `VARMA()` del paquete **MTS** en R, la cual implementa un procedimiento de búsqueda automática basado en el criterio de información de Akaike (AIC). Esta función evalúa diferentes combinaciones posibles de órdenes (p, d, q), donde p representa el orden autorregresivo, d el orden de diferenciación y q el orden

del promedio móvil, seleccionando aquella configuración que minimiza el AIC y, por tanto, garantiza un equilibrio adecuado entre calidad del ajuste y complejidad del modelo. Como resultado, se identificó inicialmente un modelo VARIMA(1,1,0) como óptimo bajo este criterio.

No obstante, al evaluar los residuos del modelo mediante las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF), se evidenció la presencia de autocorrelación significativa, lo cual viola uno de los supuestos fundamentales del modelo: la independencia de los errores. Ante esta limitación, se decidió ampliar el análisis considerando modelos con mayor número de retardos autorregresivos. Se evaluaron modelos con hasta 10 rezagos, seleccionando finalmente el modelo VARIMA(10,1,0), el cual no solo logró reducir el AIC, sino que también cumplió con los supuestos de independencia en los residuos, ofreciendo un ajuste más robusto y consistente con la estructura temporal de las componentes transformadas.

Estimaciones del Modelo

El modelo VARIMA(10,1,0) fue ajustado sobre las coordenadas ilr diferenciadas. La expresión final del modelo estimado es:

$$\Delta \mathbf{z}_t = \mu + \sum_{i=1}^{10} \Phi_i \Delta \mathbf{z}_{t-i} + \varepsilon_t \tag{45}$$

con:

$$\mu = \begin{bmatrix} 0{,}00097 \\ 0{,}02281 \\ -0{,}00043 \end{bmatrix} \quad \Phi_1 = \begin{bmatrix} -0{,}1418 & 0{,}0897 & 0{,}0138 \\ 0{,}1247 & 0{,}0125 & 0{,}0079 \\ -0{,}2235 & -0{,}0233 & -0{,}3641 \end{bmatrix} \quad \Phi_2 = \begin{bmatrix} 0{,}0773 & -0{,}0005 & -0{,}0051 \\ 0{,}0809 & -0{,}0959 & -0{,}0586 \\ 0{,}3958 & -0{,}2488 & -0{,}2332 \end{bmatrix}$$

$$\Phi_3 = \begin{bmatrix} 0{,}0352 & -0{,}0508 & 0{,}0085 \\ 0{,}1772 & -0{,}1742 & 0{,}0107 \\ 0{,}2949 & 0{,}0981 & -0{,}4962 \end{bmatrix} \quad \Phi_4 = \begin{bmatrix} 0{,}0653 & -0{,}0051 & 0{,}0576 \\ 0{,}0687 & -0{,}1951 & 0{,}0108 \\ -0{,}8617 & 0{,}3390 & -0{,}1808 \end{bmatrix}$$

$$\Phi_5 = \begin{bmatrix} 0{,}1123 & -0{,}0679 & 0{,}0439 \\ 0{,}1362 & 0{,}0096 & -0{,}0402 \\ 0{,}1402 & -0{,}1209 & -0{,}2591 \end{bmatrix} \quad \Phi_6 = \begin{bmatrix} 0{,}0349 & 0{,}1438 & -0{,}0215 \\ 0{,}1223 & -0{,}2257 & 0{,}0248 \\ -0{,}1418 & -0{,}1304 & -0{,}0466 \end{bmatrix}$$

$$\Phi_7 = \begin{bmatrix} -0{,}0377 & 0{,}0404 & -0{,}0022 \\ -0{,}0532 & 0{,}0073 & -0{,}0009 \\ -0{,}2853 & -0{,}7584 & -0{,}0351 \end{bmatrix} \quad \Phi_8 = \begin{bmatrix} -0{,}1214 & 0{,}0527 & 0{,}0008 \\ -0{,}1828 & 0{,}0743 & -0{,}0737 \\ 0{,}0170 & 0{,}2965 & -0{,}0717 \end{bmatrix}$$

$$\Phi_9 = \begin{bmatrix} -0{,}0587 & -0{,}0900 & -0{,}0075 \\ -0{,}0365 & -0{,}1871 & -0{,}0233 \\ 0{,}3979 & -0{,}5850 & 0{,}0065 \end{bmatrix} \quad \Phi_{10} = \begin{bmatrix} -0{,}3869 & 0{,}0913 & 0{,}0170 \\ -0{,}1452 & -0{,}0360 & -0{,}0145 \\ -0{,}0092 & -0{,}0039 & -0{,}1066 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 0{,}0067 & 0{,}0033 & -0{,}0012 \\ 0{,}0033 & 0{,}0112 & -0{,}0001 \\ -0{,}0012 & -0{,}0001 & 0{,}1026 \end{bmatrix}$$

Varios coeficientes del modelo resultaron estadísticamente significativos, destacándose algunos términos autorregresivos diagonales como $\Phi_{11}^{(10)} = -0{,}3869$ y efectos cruzados como $\Phi_{31}^{(4)} = -0{,}8617$, lo que evidencia una fuerte interdependencia entre las componentes de la serie transformada.

El ajuste global del modelo se respalda en los siguientes valores de los criterios de información:

$$\text{AIC} = -10{,}788, \quad \text{BIC} = -9{,}023$$

Estos valores indican un adecuado balance entre calidad de ajuste y parquedad del modelo.

La estimación del modelo VARIMA(10,1,0) ha capturado de forma eficaz la estructura temporal subyacente en las coordenadas composicionales ilr diferenciadas. En análisis posteriores, se procederá a la transformación inversa ilr^{-1} para interpretar las predicciones en el espacio composicional original.

En el Cuadro 23 se presentan los resultados de la prueba de normalidad de Shapiro–Wilk aplicada a las tres variables del modelo. Para las variables 1 y 3 se obtienen valores p inferiores a 0.05, por lo que se rechaza la hipótesis nula de normalidad en esas componentes. En cambio, para la variable 2 el valor $p = 0.1180$ es mayor que 0.05, de modo que no se rechaza la normalidad según esta prueba.

No obstante, es importante destacar que en modelos multivariantes como el C-VARIMA, el supuesto de normalidad no es estrictamente necesario para que el modelo sea válido. Aunque la normalidad puede facilitar algunas inferencias, muchos procedimientos de estimación y diagnóstico son robustos frente a desviaciones de este supuesto.

Cuadro 23: Resultados de la prueba de normalidad Shapiro-Wilk

Variable	Estadistico	p_value
1	0.9197432	0.0000002
2	0.9857572	0.1180248
3	0.9415195	0.0000057

Autocorrelación de los residuos:

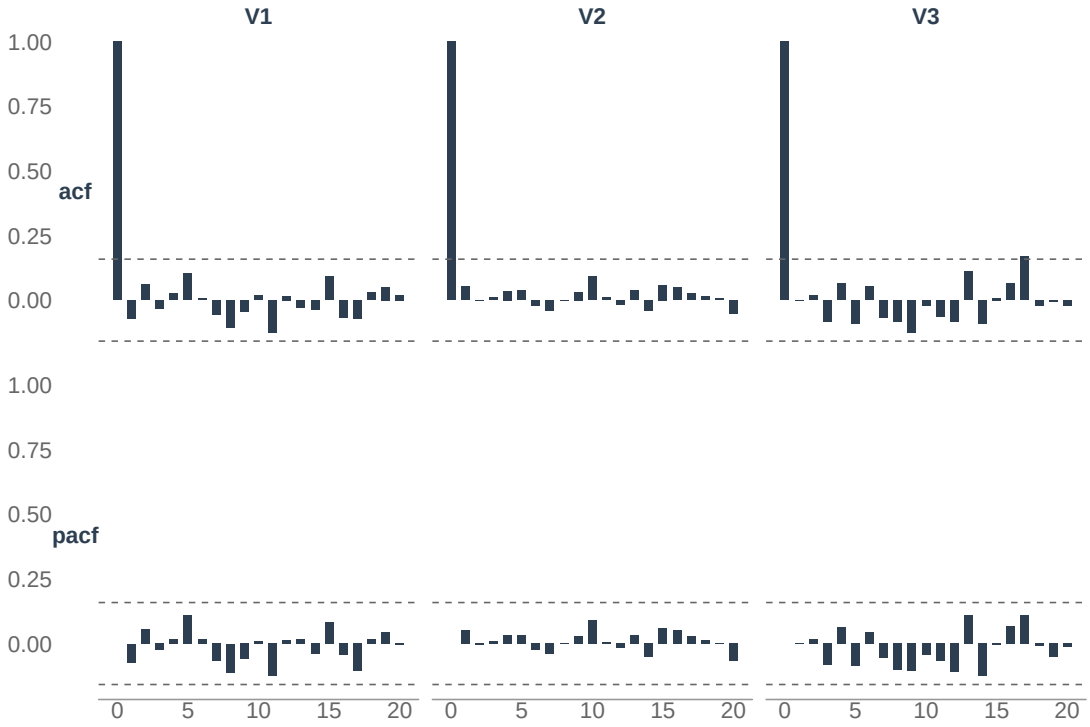

Figura 5: ACF y PACF de los residuos

Una vez estimado el modelo VARIMA(10,1,0) sobre las componentes transformadas mediante la transformación isométrica log-ratio (ilr), se procede a evaluar la validez del ajuste a través del análisis de los residuos. La Figura 5 presenta las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de los residuos correspondientes a las tres componentes del modelo.

En términos generales, se observa que, para las tres series residuales (V1, V2 y V3), las autocorrelaciones estimadas se encuentran dentro de los límites de confianza del 95 % para la mayoría de los rezagos, lo cual es indicativo de que los residuos se comportan de manera aproximada a ruido blanco. Específicamente, en la componente V1, la ACF muestra una caída abrupta después del primer rezago, sin picos significativos posteriores, mientras que la PACF se mantiene dentro de los límites, sin evidencia de autocorrelación directa. Este comportamiento sugiere que no persiste estructura serial relevante no capturada por el modelo para esta componente.

En la componente V2 se aprecia un patrón similar: las funciones ACF y PACF no presentan valores significativos más allá del rezago cero, lo que indica que la dinámica temporal de esta variable fue adecuadamente modelada y no se detectan residuos autocorrelados sistemáticos. Finalmente, en la componente V3, si bien se identifica un pequeño incremento en los residuos a partir del rezago 15 en la ACF, dicho valor no excede el umbral de significancia estadística, por lo que no compromete la validez global del modelo. La PACF correspondiente a esta componente tampoco revela valores significativos, lo cual refuerza la hipótesis de que los residuos son esencialmente aleatorios.

En conjunto, estos resultados respaldan la idoneidad del modelo VARIMA(10,1,0) propuesto, en tanto que los residuos no presentan evidencia de autocorrelación serial remanente. Este comportamiento

es fundamental para asumir la validez de las inferencias obtenidas del modelo, dado que la ausencia de estructura temporal en los residuos garantiza que la información dinámica de las series ha sido capturada de forma adecuada mediante los términos autorregresivos y la diferenciación aplicada.

Predicciones

En esta sección se presentan las predicciones generadas por el modelo C-VARIMA(10,1,0), estimado previamente sobre la serie composicional transformada. A partir de las estimaciones obtenidas, se realizan pronósticos multivariados que permiten evaluar la capacidad del modelo para replicar la dinámica observada en los datos.

En la visualización se muestran las trayectorias de predicción junto con los valores reales de la serie, lo que permite apreciar de manera general el ajuste del modelo a lo largo del tiempo.

Figura 6: Evolución temporal y predicciones de la variables: Equipo de transporte, Maquinaria y equipo, Otras construcciones y Viviendas

Para una mejor evaluación visual del desempeño reciente del modelo, se presenta un segundo gráfico centrado únicamente en los últimos 20 valores observados, donde se puede distinguir con mayor claridad la proximidad entre las predicciones y los datos reales. Esta visualización detallada permite constatar que el modelo logra capturar adecuadamente la estructura temporal de las componentes composicionales, manteniendo un comportamiento coherente en el corto plazo.

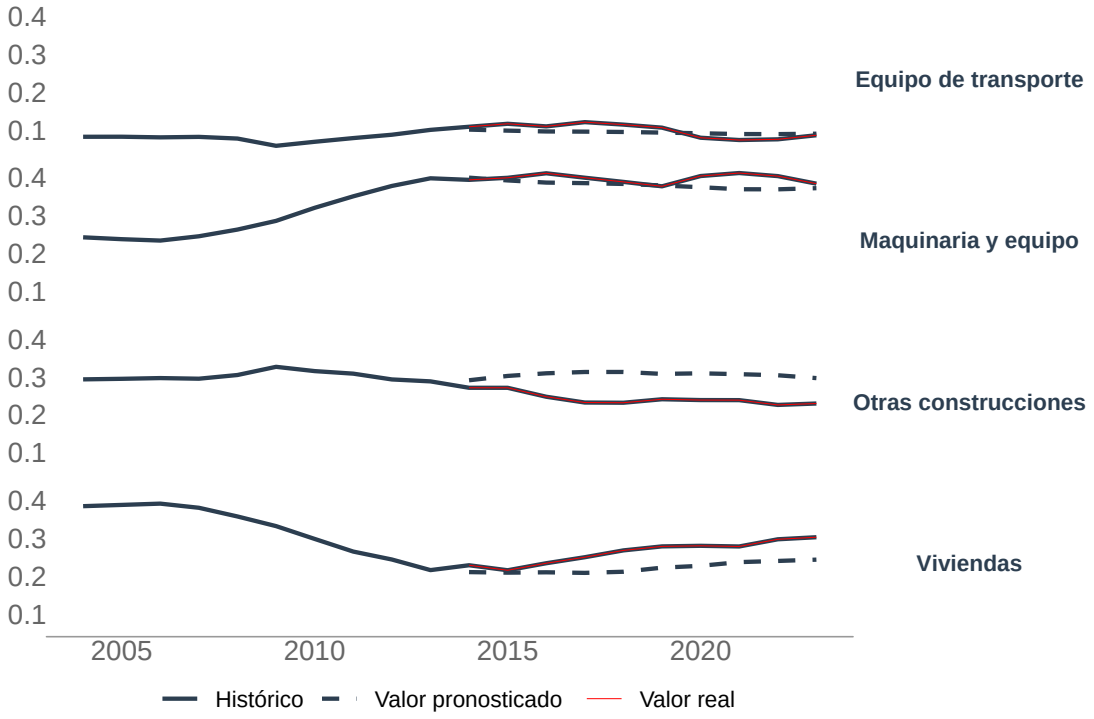

Figura 7: Evolución temporal y predicciones de la variables: Equipo de transporte, Maquinaria y equipo, Otras construcciones y Viviendas (Las últimas 20 observaciones)

El modelo C-VARIMA(10,1,0) muestra un desempeño razonable en la predicción. Tanto el MAE (Error Absoluto Medio) como el RMSE (Raíz del Error Cuadrático Medio) son bajos, lo que indica precisión en las predicciones absolutas. El MAE mide el promedio de las diferencias absolutas entre valores observados y predichos, mientras que el RMSE penaliza de manera más severa los errores grandes, reflejando así la presencia o ausencia de errores extremos en las predicciones. Además, el MAPE (Error Porcentual Absoluto Medio), que permite interpretar el error en términos relativos o porcentuales, también se mantiene en niveles aceptables. En general, un MAPE menor al 10 % se considera excelente, entre 10 % y 20 % bueno, y entre 20 % y 50 % aceptable. Estos valores bajos en las métricas de error respaldan la elección del modelo como una herramienta adecuada para representar y predecir la dinámica temporal de los datos composicionales analizados.

Cuadro 24: Resultados de métricas para evaluar la precisión de modelo

MAE	RMSE	MAPE
0.034	0.0371	15.16

Conclusión

Los resultados obtenidos a partir del estudio de simulación y la aplicación empírica permiten extraer conclusiones significativas en cuanto al comportamiento y desempeño de los modelos composicionales de series temporales, particularmente el modelo C-VARIMA bajo transformaciones log-ratio.

En el estudio de simulación, se evaluaron nueve escenarios que combinaban distintas configuraciones de autocorrelación y niveles de concentración de la distribución Dirichlet. A lo largo de todos los escenarios, se observó que tanto la transformación log-ratio isométrica (ilr) como la aditiva (alr) arrojaron resultados muy similares en términos de error cuadrático medio (RMSE), lo que indica una capacidad comparable para predecir la dinámica de los datos composicionales. No obstante, el análisis del criterio de información de Akaike (AIC) reveló una ventaja sistemática a favor de la transformación ilr, al presentar valores consistentemente más bajos, lo cual implica una mejor parsimonia en el ajuste del modelo. Esto, sumado a sus propiedades geométricas como la ortogonalidad de las coordenadas y la preservación de distancias euclídeas refuerza la idoneidad de la transformación ilr para contextos multivariantes composicionales.

En el caso práctico, aplicado a la serie histórica de la Formación Bruta de Capital Fijo (FBCF) en España (1850–2023), el modelo C-VARIMA(10,1,0) ajustado sobre coordenadas ilr permitió capturar adecuadamente la evolución temporal conjunta de los componentes composicionales. La estructura de autocorrelación observada en la ACF y PACF justificó el uso de un modelo autorregresivo de primer orden, mientras que las pruebas ADF confirmaron la necesidad de diferenciación para alcanzar la estacionariedad. A pesar de que los residuos no cumplieron el supuesto estricto de normalidad, el modelo mostró un desempeño predictivo razonable, confirmado visualmente mediante la cercanía entre las trayectorias observadas y las predichas, y cuantitativamente a través de métricas como RMSE, MAE y MAPE.

En conjunto, tanto la simulación como la aplicación empírica validan la capacidad del modelo C-VARIMA transformado con coordenadas ilr para modelar y predecir con eficacia datos composicionales temporales, respetando su estructura inherente.

Código de R

```
################################################################
## 0. LIBRERÍAS
################################################################
library(MCMCpack)
library(compositions)
library(MTS)
library(dplyr)
library(tidyr)
library(tibble)

################################################################
## 1. FUNCIÓN DE SIMULACIÓN:
##    C-VARMA(p,q) con innovaciones W_t ~ Dirichlet(1,1,1)
################################################################

varima_dirichlet_sim <- function(model, n,
                                 alpha_w = c(1, 1, 1),  # RUIDO BLANCO W_t (FIJO)
                                 n.start = 100,
                                 seed = NULL) {
  if (!is.null(seed)) set.seed(seed)
  if (!is.list(model)) stop("'model' must be a list")

  # Determinar k, p y q a partir de ar/ma
  if (!is.null(model$ar)) {
    k <- nrow(model$ar[, , 1])
    p <- dim(model$ar)[3]
  } else {
    k <- nrow(model$ma[, , 1])
    p <- 0
  }
  if (!is.null(model$ma)) {
    q <- dim(model$ma)[3]
  } else {
    q <- 0
  }
  d <- if (!is.null(model$order)) model$order[2] else 0

  total_n <- n + n.start
```

```
  # W_t ~ Dirichlet(alpha_w) => ruido blanco composicional
  innov <- t(MCMCpack::rdirichlet(total_n, alpha_w))  # matriz k x total_n

  # Serie X_t
  X <- matrix(0, nrow = k, ncol = total_n)
  epsilon <- innov  # para parte MA

  # Estado inicial: partimos de la primera innovación
  X[, 1] <- innov[, 1]

  for (t in (max(p, q) + 1):total_n) {
    AR_part <- rep(0, k)
    MA_part <- rep(0, k)

    if (p > 0) {
      for (lag in 1:p) {
        AR_part <- AR_part + model$ar[, , lag] %*% X[, t - lag]
      }
    }
    if (q > 0) {
      for (lag in 1:q) {
        MA_part <- MA_part + model$ma[, , lag] %*% epsilon[, t - lag]
      }
    }

    X_temp <- AR_part + MA_part + innov[, t]
    X[, t] <- X_temp / sum(X_temp)  # cierre composicional
  }

  # Diferenciación inversa si d > 0 (aquí d = 0)
  if (d > 0) {
    for (i in 1:k) {
      dif <- diffinv(X[i, ], differences = d)
      X[i, ] <- dif[(d + 1):length(dif)]
    }
  }

  # Serie final: filas = tiempo, columnas = componentes
  ts(t(X[, (n.start + 1):(n.start + n)]))
}

############################################################
## 2. POTENCIA COMPOSICIONAL PARA ESCENARIOS "alpha^X"
##    x -> C(x^r): cambia la concentración de X_t
############################################################

ajustar_concentracion <- function(X_mat, r) {
  # X_mat: matriz T x D con composiciones (filas suman 1)
  X_pow <- X_mat^r
  X_pow / rowSums(X_pow)
}
```

```
############################################################
## 3. AJUSTE VARMA(p,q) EN ILR Y ALR + MÉTRICAS
##    (RMSE, MAE, ME, AIC, BIC)
############################################################

ajustar_y_metricas <- function(series_list, n_obs = 10,
                               p_order, q_order) {
  resultados <- list()
  n_series   <- length(series_list)

  for (i in seq_len(n_series)) {
    serie_i <- series_list[[i]]
    n_total <- nrow(serie_i)

    X_train <- serie_i[1:(n_total - n_obs), ]
    X_test  <- serie_i[(n_total - n_obs + 1):n_total, ]

    ## ---------- ILR ----------
    ilr_train <- ilr(X_train)
    model_ilr <- tryCatch(
      VARMA(ilr_train, p = p_order, q = q_order),
      error = function(e) NULL
    )

    if (!is.null(model_ilr)) {
      pred_ilr         <- VARMApred(model_ilr, h = n_obs)$pred
      pred_ilr_símplex <- ilrInv(pred_ilr)
      err_ilr <- X_test - pred_ilr_símplex

      RMSE_ILR <- apply(err_ilr^2, 2, function(x) sqrt(mean(x)))
      MAE_ILR  <- apply(abs(err_ilr), 2, mean)
      ME_ILR   <- apply(err_ilr, 2, mean)

      AIC_ILR  <- model_ilr$aic
      BIC_ILR  <- model_ilr$bic
    } else {
      RMSE_ILR <- rep(NA, 3)
      MAE_ILR  <- rep(NA, 3)
      ME_ILR   <- rep(NA, 3)
      AIC_ILR  <- NA
      BIC_ILR  <- NA
    }

    ## ---------- ALR ----------
    alr_train <- alr(X_train)
    model_alr <- tryCatch(
      VARMA(alr_train, p = p_order, q = q_order),
      error = function(e) NULL
    )

    if (!is.null(model_alr)) {
      pred_alr         <- VARMApred(model_alr, h = n_obs)$pred
```

```
      pred_alr_símplex <- alrInv(pred_alr)
      err_alr <- X_test - pred_alr_símplex

      RMSE_ALR <- apply(err_alr^2, 2, function(x) sqrt(mean(x)))
      MAE_ALR  <- apply(abs(err_alr), 2, mean)
      ME_ALR   <- apply(err_alr, 2, mean)

      AIC_ALR  <- model_alr$aic
      BIC_ALR  <- model_alr$bic
    } else {
      RMSE_ALR <- rep(NA, 3)
      MAE_ALR  <- rep(NA, 3)
      ME_ALR   <- rep(NA, 3)
      AIC_ALR  <- NA
      BIC_ALR  <- NA
    }

    componentes <- colnames(serie_i)

    for (j in seq_along(componentes)) {
      resultados[[length(resultados) + 1]] <- tibble(
        serie_id   = i,
        componente = componentes[j],
        RMSE_ILR   = RMSE_ILR[j],
        MAE_ILR    = MAE_ILR[j],
        ME_ILR     = ME_ILR[j],
        AIC_ILR    = AIC_ILR,
        BIC_ILR    = BIC_ILR,
        RMSE_ALR   = RMSE_ALR[j],
        MAE_ALR    = MAE_ALR[j],
        ME_ALR     = ME_ALR[j],
        AIC_ALR    = AIC_ALR,
        BIC_ALR    = BIC_ALR
      )
    }
  }

  bind_rows(resultados)
}

############################################################
## 4. DEFINICIÓN DE LOS 9 ESCENARIOS
##    - alpha^X (0.5, 1, 5) ~ (0.5,0.5,0.5), (1,1,1), (5,5,5)
##    - Φ y Θ diagonales
##    - Tipo de modelo: C-VAR(1), C-VMA(1), C-VARMA(1,1)
############################################################

escenarios <- tibble(
  escenario    = 1:9,
  alphaX_label = rep(c("(0.5,0.5,0.5)", "(1,1,1)", "(5,5,5)"), each = 3),
  alphaX_r     = rep(c(0.5, 1, 5), each = 3),
```

```
  # Φ solo para documentar
  phi = list(
    c(0.8, 0.5, 0.2),  # Esc 1
    c(0.0, 0.0, 0.0),  # Esc 2
    c(0.8, 0.5, 0.2),  # Esc 3
    c(0.8, 0.5, 0.2),  # Esc 4
    c(0.0, 0.0, 0.0),  # Esc 5
    c(0.8, 0.5, 0.2),  # Esc 6
    c(0.8, 0.5, 0.2),  # Esc 7
    c(0.0, 0.0, 0.0),  # Esc 8
    c(0.8, 0.5, 0.2)   # Esc 9
  ),

  # Θ (diag) por escenario
  theta = list(
    c(0.0, 0.0, 0.0),  # Esc 1: C-VAR(1)
    c(0.8, 0.5, 0.2),  # Esc 2: C-VMA(1)
    c(0.2, 0.5, 0.8),  # Esc 3: C-VARMA(1,1)
    c(0.0, 0.0, 0.0),  # Esc 4: C-VAR(1)
    c(0.8, 0.5, 0.2),  # Esc 5: C-VMA(1)
    c(0.2, 0.5, 0.8),  # Esc 6: C-VARMA(1,1)
    c(0.0, 0.0, 0.0),  # Esc 7: C-VAR(1)
    c(0.8, 0.5, 0.2),  # Esc 8: C-VMA(1)
    c(0.2, 0.5, 0.8)   # Esc 9: C-VARMA(1,1)
  )
) %>%
  mutate(
    tipo_modelo = case_when(
      escenario %in% c(1, 4, 7) ~ "C-VAR(1)",
      escenario %in% c(2, 5, 8) ~ "C-VMA(1)",
      escenario %in% c(3, 6, 9) ~ "C-VARMA(1,1)"
    ),
    phi_str   = sapply(phi,   function(v) paste0("(", paste(v, collapse = ", "), ")")),
    theta_str = sapply(theta, function(v) paste0("(", paste(v, collapse = ", "), ")"))
  )

############################################################
## 5. SIMULACIÓN COMPLETA PARA LOS 9 ESCENARIOS
##    - En C-VAR(1):   = coef_grid,   = 0
##    - En C-VMA(1):   = coef_grid,   = 0
##    - En C-VARMA(1,1):
##        ( ,  ) = (0.2, 0.8), (0.5, 0.5), (0.8, 0.2)
############################################################

# Parámetros globales
n_sim   <- 100   # número de series por escenario
T       <- 100  # longitud de cada serie
k       <- 3    # número de componentes
n_burn  <- 100   # burn-in
n_obs   <- 10   # horizonte de predicción
alpha_w <- c(1, 1, 1)  # RUIDO BLANCO W_t FIJO: Dirichlet(1,1,1)
```

```
# Rejilla de coeficientes a evaluar
coef_grid <- c(0.8, 0.5, 0.2)

resultados_todos <- list()

for (row in seq_len(nrow(escenarios))) {

  esc        <- escenarios$escenario[row]
  tipo_mod   <- escenarios$tipo_modelo[row]
  r_val      <- escenarios$alphaX_r[row]

  # Parte MA: solo existe en C-VMA(1) y C-VARMA(1,1)
  tiene_MA <- tipo_mod %in% c("C-VMA(1)", "C-VARMA(1,1)")
  q_order  <- if (tiene_MA) 1 else 0

  ## ---- bucle sobre los valores del coeficiente (  o   / combinaciones) ----
  for (coef in coef_grid) {

    # Definir   y   que realmente se usan en este escenario
    if (tipo_mod == "C-VAR(1)") {
      #   varía,   = 0
      phi_used   <- coef
      theta_used <- 0
    } else if (tipo_mod == "C-VMA(1)") {
      #   varía,   = 0
      phi_used   <- 0
      theta_used <- coef
    } else if (tipo_mod == "C-VARMA(1,1)") {
      # Aquí imponemos las 3 combinaciones:
      # θ = 0.8 - Φ = 0.2
      # θ = 0.5 - Φ = 0.5
      # θ = 0.2 - Φ = 0.8
      if (coef == 0.8) {
        phi_used   <- 0.2
        theta_used <- 0.8
      } else if (coef == 0.5) {
        phi_used   <- 0.5
        theta_used <- 0.5
      } else if (coef == 0.2) {
        phi_used   <- 0.8
        theta_used <- 0.2
      } else {
        stop("coef_grid debe ser 0.8, 0.5 o 0.2 para C-VARMA(1,1).")
      }
    }

    ## ---- parte AR ----
    if (tipo_mod %in% c("C-VAR(1)", "C-VARMA(1,1)")) {
      ar_matrix <- array(0, dim = c(k, k, 1))
      diag(ar_matrix[, , 1]) <- phi_used   # mismo   para X, Y, Z
      p_order <- 1
    } else {
```

```
  ar_matrix <- NULL
  p_order   <- 0
}

## ---- parte MA ----
if (tiene_MA) {
  ma_matrix <- array(0, dim = c(k, k, 1))
  diag(ma_matrix[, , 1]) <- theta_used # mismo   para X, Y, Z
} else {
  ma_matrix <- NULL
}

## ---- definir lista 'model' para la simulación ----
model <- list(
  ar    = ar_matrix,
  ma    = ma_matrix,
  order = c(p_order, 0, q_order)  # (p, d, q) con d = 0
)

## 1) Simular series base con ruido fijo
series_base <- lapply(seq_len(n_sim), function(i) {
  X_i <- varima_dirichlet_sim(
    model   = model,
    n       = T,
    alpha_w = alpha_w,
    n.start = n_burn,
    seed    = 1000 * esc + 10 * i + round(100 * coef)
  )
  colnames(X_i) <- c("X", "Y", "Z")
  X_i
})

## 2) Ajustar concentración de X_t según alpha^X (potencia r)
series_r <- lapply(series_base, function(X_mat) {
  ajustar_concentracion(X_mat, r = r_val)
})

## 3) Ajustar modelos ILR/ALR con VARMA(p,q) coherente
tabla_metricas <- ajustar_y_metricas(series_r,
                                     n_obs   = n_obs,
                                     p_order = p_order,
                                     q_order = q_order) %>%
  mutate(
    escenario    = esc,
    tipo_modelo  = tipo_mod,
    alphaX_label = escenarios$alphaX_label[row],
    alphaX_r     = r_val,
    phi          = phi_used,
    theta        = theta_used,
    # Etiquetas de texto simples con los valores usados
    phi_str      = paste0("(", phi_used, ")"),
    theta_str    = paste0("(", theta_used, ")")
```

```
      )

    resultados_todos[[length(resultados_todos) + 1]] <- tabla_metricas
  }
}

# Tabla detallada con TODAS las simulaciones y métricas
tabla_resultados_raw <- bind_rows(resultados_todos)

################################################################
## 6. TABLA RESUMEN: PROMEDIOS POR ESCENARIO,  ,   Y COMPONENTE
################################################################

tabla_resumen <- tabla_resultados_raw %>%
  group_by(
    escenario, tipo_modelo,
    alphaX_label, alphaX_r,
    theta, phi,              # <-   y   ya vienen numéricos
    phi_str, theta_str,
    componente
  ) %>%
  summarise(
    RMSE_ILR = mean(RMSE_ILR, na.rm = TRUE),
    MAE_ILR  = mean(MAE_ILR,  na.rm = TRUE),
    ME_ILR   = mean(ME_ILR,   na.rm = TRUE),
    AIC_ILR  = mean(AIC_ILR,  na.rm = TRUE),
    BIC_ILR  = mean(BIC_ILR,  na.rm = TRUE),

    RMSE_ALR = mean(RMSE_ALR, na.rm = TRUE),
    MAE_ALR  = mean(MAE_ALR,  na.rm = TRUE),
    ME_ALR   = mean(ME_ALR,   na.rm = TRUE),
    AIC_ALR  = mean(AIC_ALR,  na.rm = TRUE),
    BIC_ALR  = mean(BIC_ALR,  na.rm = TRUE),
    .groups = "drop"
  )

# Versión "limpia" para exportar a Excel
tabla_resumen_excel <- tabla_resumen %>%
  select(
    escenario, tipo_modelo,
    alphaX_label, alphaX_r,
    theta, phi,
    phi_str, theta_str,
    componente,
    RMSE_ILR, RMSE_ALR, AIC_ILR, AIC_ALR
  )
```

Bibliografía

[1] Abramowitz, M., y Stegun, I. A. (Eds.). (1964). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (Applied Mathematics Series 55). National Bureau of Standards. (10th printing, 1972)

[2] Aitchison J and Brown JAC (1969) The Lognormal Distribution with Special Reference to its Uses in Econometrics. Department of Applied Economics Monograph: 5. Cambridge University Press, Cambridge (UK). 176 p.

[3] Aitchison J and Shen SM 1980 Logistic-normal distributions. Some properties and uses. Biometrika 67(2), 261-272.

[4] Aitchison J 1982 The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44(2), 139-177

[5] Aitchison J 1985 A general class of distributions on the símplex. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 47(1), 136-146.

[6] Aitchison J 1986 The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd (reprinted 2003 with additional material by The Blackburn Press), London (UK). 416 p.

[7] Azzalini A 2005 The skew normal distribution and related multivariate families. Scandinavian Journal of Statistics 32(2), 159-188.

[8] Barcelo-Vidal C, Aguilar L and Martín-Fernandez J 2007 Time series of compositional data: A first approach. In Proceedings of the 22nd International Workshop of Statistical Modelling (IWSM 2007), (ed. del Castillo J, Espinal A and Puig P). Institut d'Estadística de Catalunya (IDESCAT), Barcelona (Spain). pp. 81-86.

[9] Bell W, Bozik J, McKenzie S and Shulman H 1986 Time series analysis of household headship pro portions: 1959-1985. Technical Report 86/01, Statistical Research Division, Bureau of the Census, Washington (USA).

[10] Bergman J 2008 Compositional time series: An application. In Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop (ed. Daunis-i Estadella J and Martin-Fernandez J), p. http://hdl.handle.net/10256/723. University of Girona, Girona (Spain). CD-ROM

[11] Bhaumik A, Dey DK and Ravishanker N 2003 A dynamic linear model approach for compositional time series analysis. Technical Report, University of Connecticut (USA).

[12] Billheimer, D., Guttorp, P., Fagan, W.F. 2001. Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456), 1205–1214.

[13] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M. 2016. *Time Series Analysis: Forecasting and Control* (5th ed.). Hoboken, NJ: Wiley.

[14] Brunsdon TM 1987 Time series of compositional data. PhD thesis, University of Southampton, Southampton (UK)

[15] Brunsdon TM and Smith TMF 1998 The time series analysis of compositional data. Journal of Official Statistics 14(3), 237-253.

[16] Buccianti, A., Pawlowsky-Glahn, V. 2005. New perspectives on compositional data analysis: a response to the comments on 'compositional data and their analysis: some comments'. *Mathematical Geology*, 37(7), 829–848.

[17] Chayes F (1960) On correlation between variables of constant sum. Journal of Geophysical Research 65(12), 4185-4193.

[18] Connor, R.J., Mosimann, J.E. 1969 Concepts of independence for proportions with a generalization of the Dirichlet distribution. J. Am. Stat. Assoc. 64(325), 194-206

[19] Cox D and Snell E 1989 Analysis of Binary Data, 2nd edition. Chapman and Hall/CRC, London (UK). p. 236.

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

[21] G.K. Grunwald, A.E. Raftery, P. Guttorp, Time series of continuous proportions, J. R. Stat. Soc. Ser. B Stat. Methodol. 55 (1993) 103-116.

[22] Khusna, H., Ahsan, M., and Prastyo, D. D. (2017). Number of foreign tourist arrival forecasting using percentile error bootstrap based on varima model. IPTEK Journal of Proceedings Series, 3(2):79-83.

[23] Kotz S, Balakrishnan N and Johnson NL 2000 Continuous Multivariate Distributions. Volume I, Models and Applications. Wiley Series in Probability and Statistics. Wiley-Interscience, New York, NY (USA). 730 p.

[24] Lütkepohl, H. 2005. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

[25] Martin-Fernandez JA, Barcel ´ o-Vidal C and Pawlowsky-Glahn V 2003 Dealing with zeros and missing ´ values in compositional data sets using nonparametric imputation Mathematical Geology 35(3), 253-278

[26] Mateu-Figueras G, Pawlowsky-Glahn V and Barcel´o-Vidal C 2005 The additive logistic skew-normal distribution on the símplex. Stochastic Environmental Research and Risk Assessment (SERRA) 19(3),205-214.

[27] McAlister D 1879 The law of the geometric mean. Proceedings of the Royal Society of London 29, 367-376

[28] Mills T 2009 Forecasting obesity trends in England. Journal of the Royal Statistics Society, Series A 172(1), 107-17

[29] Mills T 2010 Forecasting compositional time series. Journal Quality and Quantity 44, 673-690

[30] Nolan T and Smith G 1995 Time series analysis of the prevalence of endoparasitic infections in cats and dogs presented to a veterinary teaching hospital. Veterinary Parasitology 59(2), 87-96

[31] Pawlowsky-Glahn V and Egozcue JJ 2001 Geometric approach to statistical analysis on the símplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15(5), 384-398.

[32] Mateu-Figueras, G. y Pawlowsky-Glahn, V. (2008). A Critical Approach to Probability Laws in Geochemistry. En G. Bonham-Carter y Q. Cheng (eds.), Progress in Geomathematics (pp. 39–52).

[33] 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 LX,489-502.

[34] Pitombeira-Neto, A. R., Loureiro, C. F. G. (2016). A dynamic linear model for the estimation of time-varying origin-destination matrices from link counts. Journal of Advanced Transportation, 50(8), 2116-2129.

[35] Quintana JM and West M 1988 Time series analysis of compositional data. In Bayesian Statistics 3 (ed. Bernardo JM, DeGroot MH, Lindley DV and Smith AFM). Oxford University Press, New York, NY (USA). pp. 747-756

[36] Ravishanker N, Dey D and Iyengar M 2001 Compositional time series analysis of mortality proportions. Communications in Statistics - Theory and Methods 30(11), 2281-2291

[37] Silva D 1996 Modelling compositional time series from repeated surveys. PhD thesis University of Southampton, Southampton (UK)

[38] Silva D and Smith T 2001 Modelling compositional time series from repeated surveys. Survey Method ology 27, 205-215.

[39] Smith T and Brunsdon T 1989 The time series analysis of compositional data. In Proceedings of the Survey Research Methods Section. American Statistical Association, Alexandria, VA (USA). pp. 26-32.

[40] Tanner J (1949) Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. Journal of Applied Physiology 2(1), 1-15.

[41] Tjøstheim, D., Paulsen, J. 2017. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. Oxford: Oxford University Press.

[42] Wang H, Liu Q, Mok H, Fu L and Tse W 2007 A hyperspherical transformation forecasting model for compositional data. European Journal of Operational Research 179(2), 459-468.