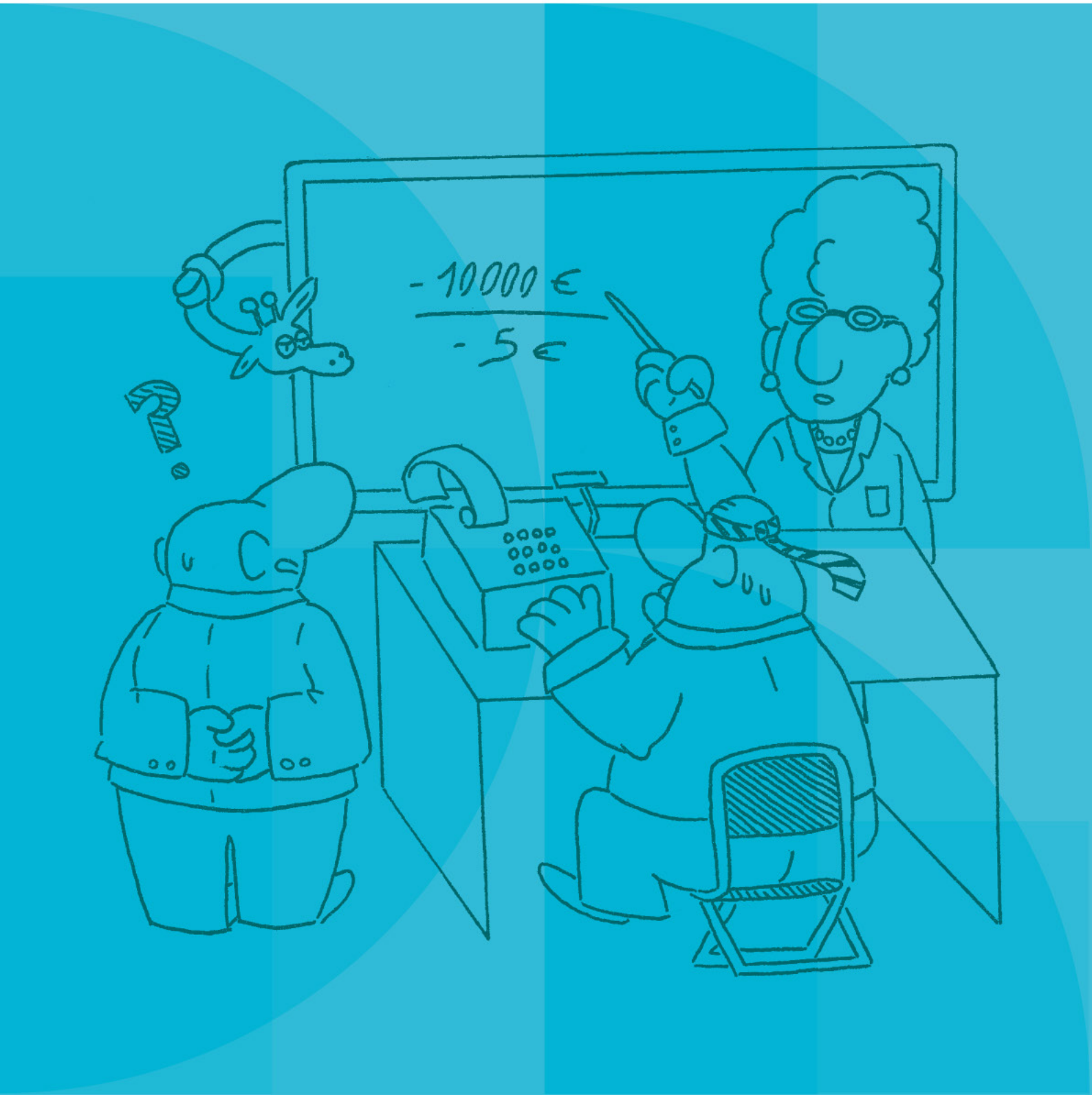
Análisis estadístico de la información financiera

Metodología composicional

Germà Coenders Gallart
Núria Arimany Serrat

Documenta Universitaria



Análisis estadístico de la información financiera

Metodología composicional

Germà Coenders Gallart
Departamento de Economía. Universitat de Girona
https://orcid.org/0000-0002-5204-6882

Núria Arimany Serrat
Departamento de Economía y Empresa. Universitat de Vic-Universitat Central de Catalunya
https://orcid.org/0000-0003-0323-6601

Documenta Universitaria

Datos CIP recomendados por la Biblioteca de la UdG

CIP 519.2 COE

Coenders, Germà, autor
Análisis estadístico de la información financiera : metodología composicional / Germà Coenders Gallart, Núria Arimany Serrat.
– Vic ; Girona : Documenta Universitaria, marzo de 2025. –
1 recurs en línia (135 pàgines) : il·lustracions
Descripció del recurs: 20 maig 2025
ISBN 978-84-9984-701-6

I. Arimany Serrat, Núria, autor 1. Finances – Estadístiques
2. Estadística – Programes d'ordinador 3. Llibres electrònics

CIP 519.2 COE

www.documentauniversitaria.com
info@documentauniversitaria.com
Documenta Universitaria® d'Edicions a Petició, SL

ISBN: 978-84-9984-701-6
DOI: 10.33115/b/9788499847016

Vic y Girona, marzo de 2025

Documenta Universitaria
@DocUniv
documentauniversitaria.com

Índex

1. Introducción

En cualquier investigación académica y profesional, ya sea de organizaciones, instituciones o empresas, se utilizan análisis estadísticos que avalen estudios que van desde las finanzas hasta cualquier área de conocimiento de la empresa, para poder tomar las decisiones económicas oportunas, basadas en datos reales. Los análisis estadísticos permiten identificar tendencias, realizar previsiones y llevar a cabo diagnósticos en diversas áreas del contexto empresarial (finanzas, marketing, producción y recursos humanos).

En el caso de la información financiera, un análisis muy utilizado es el de ratios, proporciones que son una relación matemática que compara dos magnitudes o cantidades, generalmente expresada como una fracción, y que indica las veces que una cantidad contiene a otra, de modo que permite comparar las magnitudes.

En el caso de las ratios financieras o ratios contables clásicas, el análisis permite diagnosticar la situación de las empresas (Altman, 1968; Amat Salas, 2020; Barnes, 1987; Horrigan, 1968; Qin et al., 2022; Ross et al., 2003; Soukal et al., 2024; Staňková y Hampel, 2023; Tascón et al., 2018; Veganzones y Severin, 2021; Willer do Prado et al., 2016). Es decir, permite analizar la información financiera a corto plazo y a largo plazo, junto al análisis de los resultados, mediante diversas ratios que, a pesar de ser un instrumento estático, por referirse a un momento concreto del tiempo, pueden compararse entre distintos períodos de tiempo y entre distintas empresas de un sector.

Este libro trata de análisis estadísticos en los que las ratios financieras se usan precisamente como variables estadísticas. En estos análisis, dichas ratios se deben construir con la metodología de datos composicionales (CoDa, acrónimo del término inglés *compositional data*), que resuelve los principales problemas de las ratios financieras clásicas en el análisis estadístico como la asimetría (Linares-Mustarós et al., 2018), la no normalidad de las distribuciones (Iotti et al., 2024a; 2024b), la no linealidad de las relaciones entre variables (Carreras-Simó y Coenders, 2021), las observaciones atípicas extremas (Deshpande, 2023) y la dependencia de los resultados según qué cifra contable va en el numerador o en el denominador de la ratio (Coenders et al., 2023a; Linares-Mustarós et al., 2022).

A pesar de los mencionados graves problemas de las ratios clásicas, actualmente, muchos análisis de los estados financieros de un sector no se realizan con datos composicionales y, por tanto, el análisis sectorial no es fidedigno en el ámbito contable y estadístico. Lejos de ser un lucimiento metodológico, el análisis CoDa conduce a conclusiones mejores y sustancialmente diferentes a los del análisis clásico siempre que se han comparado sus resultados (Arimany-Serrat et al., 2022; Carreras-Simó y Coenders, 2021; Coenders et al., 2023a; Creixans-Tenas

et al., 2019; Dao et al., 2024; Escaramís y Arbussà, 2025; Jofre-Campuzano y Coenders, 2022; Linares-Mustarós et al., 2018; 2022). Por ello, el objetivo de este libro es generalizar la utilización del análisis financiero sectorial con datos composicionales para conseguir una fiel imagen económica y financiera en el ámbito estadístico y poder tomar las oportunas decisiones empresariales. Cabe destacar que cualquier análisis estadístico, sea financiero o de cualquier área de conocimiento de la empresa, que utilice ratios como variables de entrada, debe contemplar la metodología CoDa objeto de este estudio.

La metodología que hay que utilizar para el análisis estadístico de la información financiera es, pues, la metodología CoDa y en este libro se desarrollan de manera minuciosa los pasos que hay que seguir para usar e interpretar dicha metodología. Además, la hace extensiva a todos los análisis de los estados financieros no solo de un sector, sino de una muestra de empresas seleccionadas con criterios cualesquiera (pymes, empresas del IBEX 35, empresas de una determinada ciudad o provincia, etc.), que utilicen ratios como variables en un estudio financiero o de cualquier área de la empresa. En el libro se detallan las bases y los fundamentos estadísticos de esta metodología CoDa, se simplifica dicha metodología en lo posible, y se adapta a nivel del análisis de los estados financieros, para conseguir una imagen más fiel y una toma de decisiones más precisa que ayude a la supervivencia de las organizaciones, instituciones y empresas en el escenario actual, de cambios sistémicos muy ágiles (Bhimani, 2008; Buchetti et al., 2022; Giacomelli, et al., 2021; Gourinchas, et al., 2020; Minnis y Shroff, 2017; Sunder, 2016; de Vito y Gómez, 2020).

En el análisis de los estados financieros, las aplicaciones composicionales todavía son escasas (Arimany-Serrat y Coenders, 2025; Arimany-Serrat et al., 2022; 2023; Arimany-Serrat y Sgorla, 2024; Carreras-Simó y Coenders, 2020; 2021; Coenders, 2025; Coenders y Arimany-Serrat, 2023; Coenders et al., 2023a; Creixans-Tenas et al., 2019; Dao et al., 2024; Escaramís y Arbussà, 2025; Jofre-Campuzano y Coenders, 2022; Linares-Mustarós et al., 2018; 2022; Molas-Colomer et al., 2024; Mulet-Forteza et al., 2024; Saus-Sala et al., 2021; 2023; 2024), a pesar de la gravedad de los problemas de las ratios clásicas que la metodología CoDa resuelve. En concreto, visualizar todas las empresas individuales del sector con un *biplot* composicional, clasificarlas por análisis clúster, llamado también *análisis de conglomerados composicional*, y utilizar modelos de regresión composicional para relacionar variables financieras y no financieras, da gran solvencia a la interpretación y el diagnóstico del análisis para detectar la salud financiera de un conjunto de empresas.

Este libro puede servir de manual para cursos de análisis de los estados financieros avanzado, para estudiantes que quieran realizar su trabajo de fin de grado o máster o su tesis doctoral en el ámbito del análisis estadístico de la información financiera sobre una muestra de empresas, para investigadores y profesionales que elaboren artículos sobre el mismo tema y para asociaciones profesionales interesadas en unos valores medios de referencia de las ratios. Se presupone que el lector o lectora ha cursado ya temas de introducción al análisis de los estados financieros (o está familiarizado con las ratios financieras más habituales) y temas de introducción a la estadística a nivel de grado.

El libro, después de esta introducción, presenta diferentes apartados: el análisis de los estados financieros utilizando ratios clásicas; los problemas de las ratios clásicas en el análisis estadístico de un conjunto de datos en un sector o de la

categoría de empresas de la que se trate; el análisis de los estados financieros utilizando datos composicionales; las medias sectoriales fidedignas; el *biplot* composicional; el *análisis clúster* composicional; la *regresión* composicional y las conclusiones en forma de un decálogo final. En cada capítulo, en un apartado titulado «Manos a la obra», se describe cómo se realiza el análisis en el sector vitivinícola, con una aplicación a los estados contables de bodegas españolas mediante un tutorial del software abierto composicional CoDaPack (Comas-Cufí y Thió-Henestrosa, 2011; Thió-Henestrosa y Martín-Fernández, 2005). En el libro se ha simplificado la metodología CoDa en todo lo posible y se han reservado las partes complejas en un apartado final de cada capítulo titulado «Para saber más». Por último, la bibliografía es exhaustiva y actualizada, y puede servir a los estudiantes de fin de grado, máster y doctorado para sus tesis.

Este libro ha sido posible gracias a la financiación de los siguientes proyectos: Ministerio de Ciencia, Innovación y Universidades / FEDER-a way of making Europe (proyectos RTI2018-095518-B-C21 y PID2021-123833OB-I00); Departamento de Investigación y Universidades de la Generalitat de Catalunya (proyectos 2021SGR01197 y 2021SGR00403), Ministerio de Sanidad (proyecto CIBERCB06/02/1002) y Departamento de Investigación y Universidades, AGAUR y Departamento de Acción Climática, Alimentación y Agenda Rural de la Generalitat de Catalunya (proyecto 2023-CLIMA-00037). Los últimos detalles sobre las actividades de dichos proyectos se pueden encontrar en https://www.researchgate.net/lab/Lab-on-financial-statement-analysis-as-compositional-data-Germa-Coenders-2. Queremos agradecer a Salvador Linares Mustarós, Elisabet Saus Sala, Elena Rondós Casas, Maria Àngels Farreras Noguer y Mike d'Alessandro Hernández Romero sus comentarios al borrador de este libro.

2. Análisis de los estados financieros utilizando ratios clásicas

Las ratios financieras clásicas, ratios derivadas de la información de los estados financieros, son relaciones matemáticas que comparan dos valores extraídos de los estados financieros. Estas ratios financieras clásicas permiten aproximar la salud financiera, el perfil o el desempeño financiero.

En nuestro libro, las ratios utilizan valores extraídos de dos documentos de los estados financieros, el *balance* y la *cuenta de pérdidas y ganancias,* llamada también *cuenta de resultados*. El documento contable del balance refleja la situación financiera y patrimonial de la empresa en un momento determinado y está compuesto por el *activo*, el *pasivo* y el *patrimonio neto*. Por otra parte, se utiliza el documento contable de la cuenta de pérdidas y ganancias, que refleja la gestión de la empresa, con los ingresos y gastos del ejercicio económico, que engloban los ingresos de explotación y los gastos de explotación.

El análisis de los estados financieros utilizando ratios financieras clásicas estudia la *situación financiera a corto plazo*, la *situación financiera a largo plazo* y la *rentabilidad* de la empresa. Respecto a la situación financiera a corto plazo, se analiza la capacidad de la empresa de atender las obligaciones de pago a corto plazo. El equilibrio financiero a corto plazo se logra cuando la empresa genera el efectivo necesario para pagar las deudas. En cuanto al análisis financiero a largo plazo, se mide la capacidad de satisfacer las deudas a largo plazo, y es recomendable valorar la composición de la estructura financiera, de la estructura del activo y de la estructura económica para valorar el equilibrio a largo plazo. Respecto a la capacidad para aumentar el valor de los capitales invertidos, es decir, de generar *resultados* según las inversiones realizadas, se analizan las rentabilidades, financiera y económica (según la cantidad de beneficios generados por unidad monetaria invertida y según las inversiones en activo al margen de la financiación, respectivamente).

2.1. Ratios financieras clásicas

En este apartado se detallan las ratios financieras clásicas que utilizaremos en este libro para realizar el análisis de los estados financieros. En primer lugar, los valores de los estados financieros utilizados son seis, que se corresponden, los cuatro primeros, a *masas patrimoniales* del balance y, los dos últimos, a valores agregados de la cuenta de pérdidas y ganancias, relacionados directamente con los ingresos y gastos de la actividad de explotación. De

estos seis valores surgen el *activo total* ($x_1 + x_2$), *el pasivo total* ($x_3 + x_4$), el *patrimonio neto* ($x_1 + x_2 - x_3 - x_4$), el *beneficio de explotación* ($x_5 - x_6$) y el *fondo de maniobra* ($x_2 - x_4$).

- x_1: activo no corriente
- x_2: activo corriente
- x_3: pasivo no corriente
- x_4: pasivo corriente
- x_5: ingresos de explotación
- x_6: gastos de explotación

Los valores x_j son siempre positivos (como se detalla en el capítulo 4), y las ratios que se derivan de estos son ratios financieras clásicas que permiten construir un sistema de información para diagnosticar la salud financiera de las empresas, y facilitar la toma de decisiones empresariales (Amat-Salas, 2020). A continuación, se presentan ratios financieras clásicas referentes para el estudio de los resultados y los análisis a corto y largo plazo.

En cuanto al análisis de los resultados, disponemos de estas ratios:

- *Ratio de rotación* (ingresos de explotación sobre activo total):

$$(1)\quad x_5/(x_1 + x_2)$$

- *Ratio de rotación del activo corriente* (ingresos de explotación sobre activo corriente):

$$(2)\quad x_5/x_2$$

- *Ratio de margen* (beneficio de explotación sobre ingresos de explotación):

$$(3)\quad (x_5 - x_6)/x_5$$

- *Ratio de apalancamiento* (activo total sobre patrimonio neto):

$$(4)\quad (x_1 + x_2)/(x_1 + x_2 - x_3 - x_4)$$

- *Ratio de rentabilidad económica* (beneficio de explotación sobre activo total, denominada también ROA, acrónimo inglés de *return on assets*)

$$(5)\quad (x_5 - x_6)/(x_1 + x_2)$$

  que también se obtiene multiplicando el margen por la rotación.
- *Ratio de rentabilidad financiera* (beneficio de explotación sobre patrimonio neto, denominada también ROE, acrónimo inglés de *return on equity*)

$$(6)\quad (x_5 - x_6)/(x_1 + x_2 - x_3 - x_4)$$

  que también se obtiene multiplicando el ROA por el apalancamiento.

En cuanto al análisis de la situación financiera a corto y largo plazo, las ratios clásicas son:

- *Ratio de endeudamiento* (pasivo total sobre activo total):

$$(7)\quad (x_3 + x_4)/(x_1 + x_2)$$

- *Ratio de endeudamiento a corto plazo* (pasivo corriente sobre activo total):

$$(8)\quad x_4/(x_1 + x_2)$$

- *Ratio de solvencia a largo plazo* (activo total sobre pasivo total, es decir, permutación del numerador y el denominador del endeudamiento):

$$(9)\quad (x_1 + x_2)/(x_3 + x_4)$$

- *Ratio de solvencia a corto plazo* (activo corriente sobre pasivo corriente, denominada también *ratio de liquidez*):

$$(10)\quad x_2/x_4$$

- *Ratio de inmovilización del activo* (activo no corriente sobre activo corriente):

$$(11)\quad x_1/x_2$$

- *Ratio de maduración de la deuda* (pasivo no corriente sobre pasivo corriente, denominada también *ratio de vencimiento de la deuda*):

$$(12)\quad x_3/x_4$$

Muchas de estas ratios tienen variantes que aportan la misma información al estar calculadas a partir de los mismos valores de los estados financieros (Chen y Shimerda, 1981). Por ejemplo, para la solvencia a largo plazo se usa a veces la ratio de patrimonio neto sobre activo total $(x_1 + x_2 - x_3 - x_4)/(x_1 + x_2)$, para el endeudamiento la ratio de pasivo total sobre patrimonio neto $(x_3 + x_4)/(x_1 + x_2 - x_3 - x_4)$, para la inmovilización del activo la ratio de activo no corriente sobre activo total $x_1/(x_1 + x_2)$, y para la maduración de la deuda la ratio de pasivo corriente sobre pasivo total $x_4/(x_3 + x_4)$ llamada *calidad de la deuda*. La prueba de que aportan la misma información es que se pueden obtener unas a partir de las otras. Por ejemplo, si invertimos la calidad de la deuda $x_4/(x_3 + x_4)$, obtenemos $(x_3 + x_4)/x_4$ que es la maduración de la deuda más uno: $(x_3 + x_4)/x_4 = x_3/x_4 + x_4/x_4 = x_3/x_4 + 1$.

Cabe destacar que las ratios financieras clásicas ayudan a evaluar la salud económica y financiera de una empresa individual, o compararla con otras empresas afines. Este es el uso primigenio de las ratios. No obstante, debido a los problemas que detallaremos en el capítulo 3, no son fidedignas para valorar la salud económica y financiera de una muestra estadística (por ejemplo, obtenida de un sector de actividad, de determinadas pymes, de empresas de índices bursátiles de referencia, de empresas de una ciudad o provincia, entre otras). Además, tampoco son fidedignas como variables en otros análisis estadísticos, aisladamente o acompañadas de variables no financieras. Para utilizar ratios en análisis estadísticos debe usarse la metodología CoDa, tal como se detalla a partir del capítulo 4.

2.2. Manos a la obra. Extracción de la muestra de un sector a través de la base de datos SABI

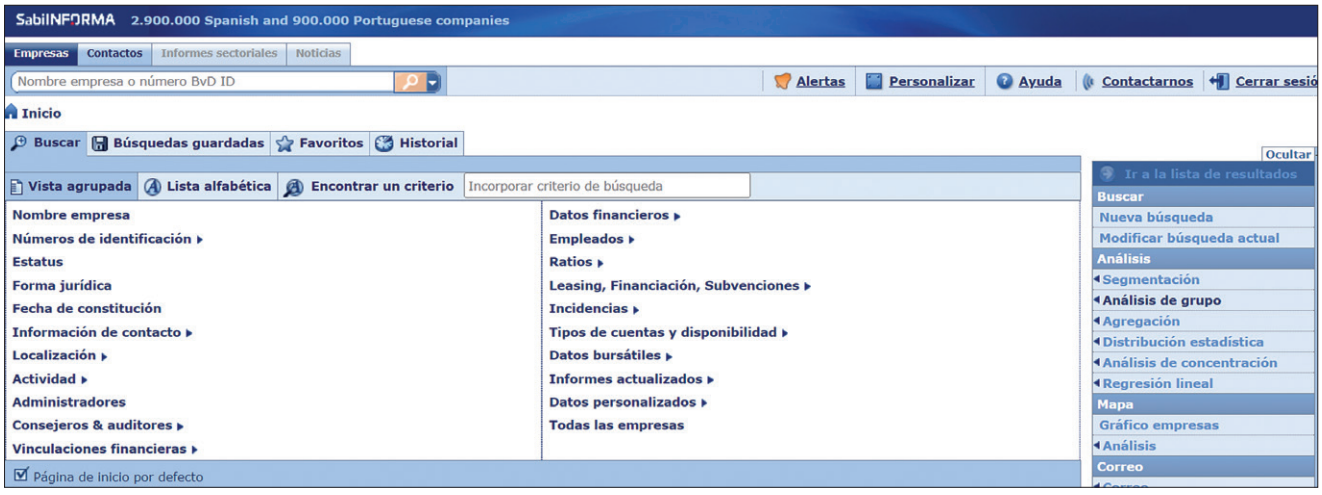
Para el análisis estadístico de la información financiera de un sector precisamos de una muestra objetiva de dicho sector. Y esta muestra, en el entorno hispano-luso, la conseguimos gracias a la base de datos denominada Sistema de Análisis de Balances Ibéricos (SABI), que recoge la información financiera y general de 2,9 millones de empresas de España y 0,9 millones de Portugal. Además, esta base de datos permite realizar análisis precisos de los documentos contables, como el balance, la cuenta de pérdidas y ganancias, el estado de cambios en el patrimonio neto y el estado de flujos de efectivo. Los instrumentos para los análisis utilizados son la información financiera y no financiera, cuantitativa y cualitativa, de las empresas obligadas a presentar cuentas anuales. Esta base de datos es utilizada por investigadores, analistas financieros, inversores, estudiantes y otras partes interesadas para conocer la salud de un ecosistema empresarial objeto de estudio.

Para acceder a SABI, la conexión es posible en la red de la universidad y en remoto, si la institución dispone de dicha base de datos, normalmente desde el catálogo de la biblioteca. Para proceder a extraer la muestra en la base de datos SABI, aplicamos unos primeros filtros:

- que sean empresas mercantiles (*Forma jurídica*: sociedad anónima y sociedad limitada);
- que sean activas (*Estatus*);
- que sean del sector de actividad objeto de estudio (según el CNAE, oportuno en *Actividad* buscando por *Clasificaciones actividades* o buscando por alguna palabra clave, en la búsqueda textual);
- que sean de una determinada zona geográfica (*Localización*) y
- por un período de tiempo.

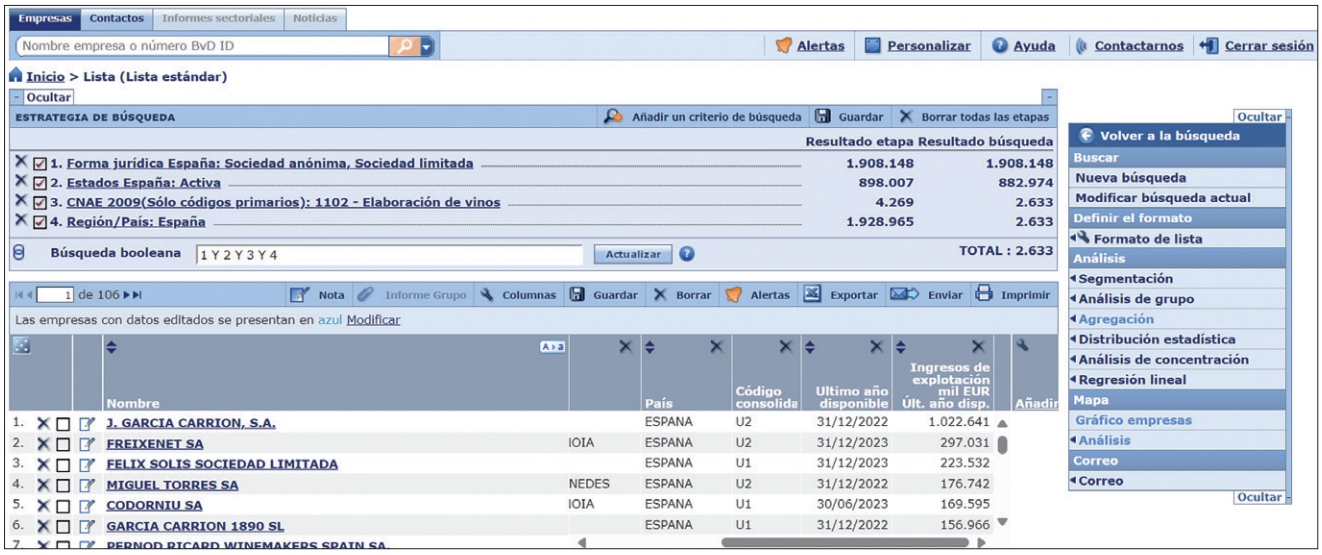
Así pues, la base de datos SABI dispone de diferentes opciones de búsqueda: búsqueda rápida, desde la pantalla de inicio y búsqueda por criterios. También podemos utilizar identificadores como el NIF, o buscar una empresa concreta con el nombre de la empresa. Por defecto, obtendremos un informe estándar de la empresa, que permite exportar la información.

Estos pasos se pueden realizar a través de la pantalla de inicio:

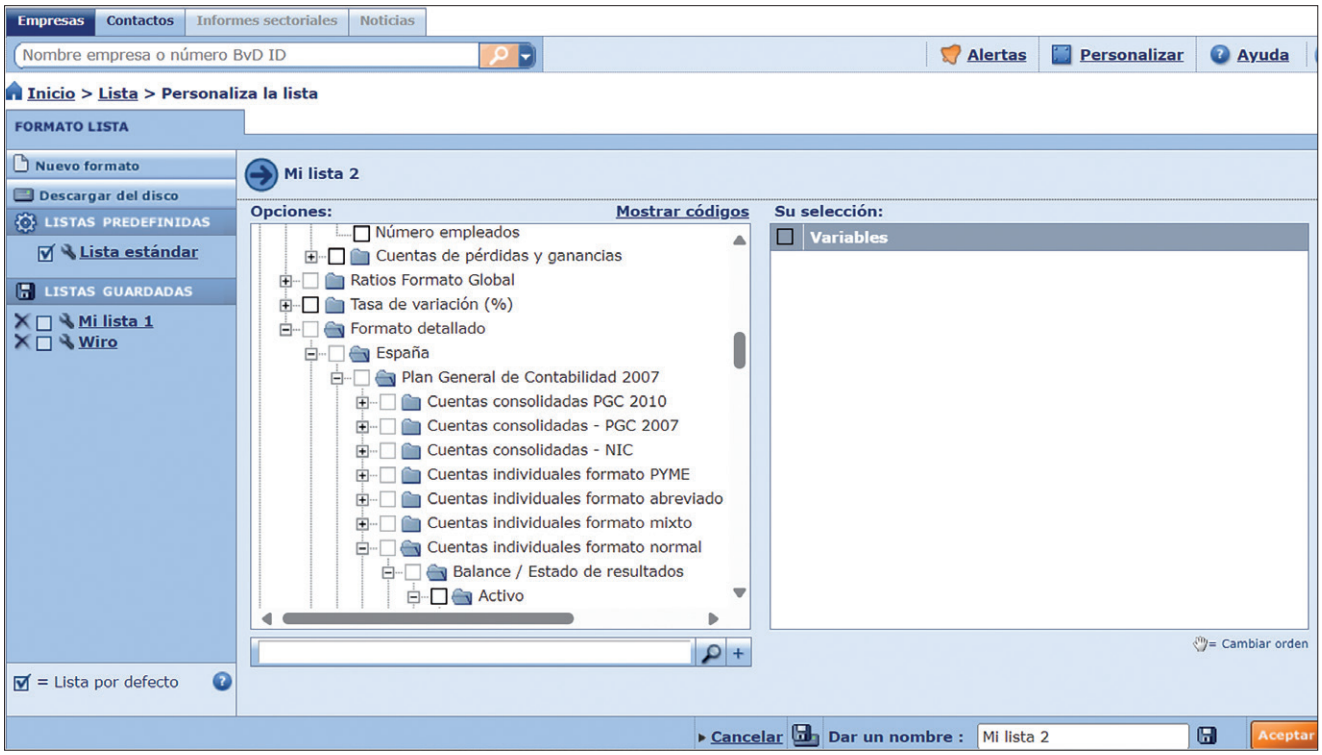


Si escogemos los filtros anteriores y el sector con CNAE 1102 «Elaboración de vinos» en un período concreto, una vez aplicados los filtros, se selecciona la muestra y se visualizan los resultados (*Ir a la lista de resultados*).

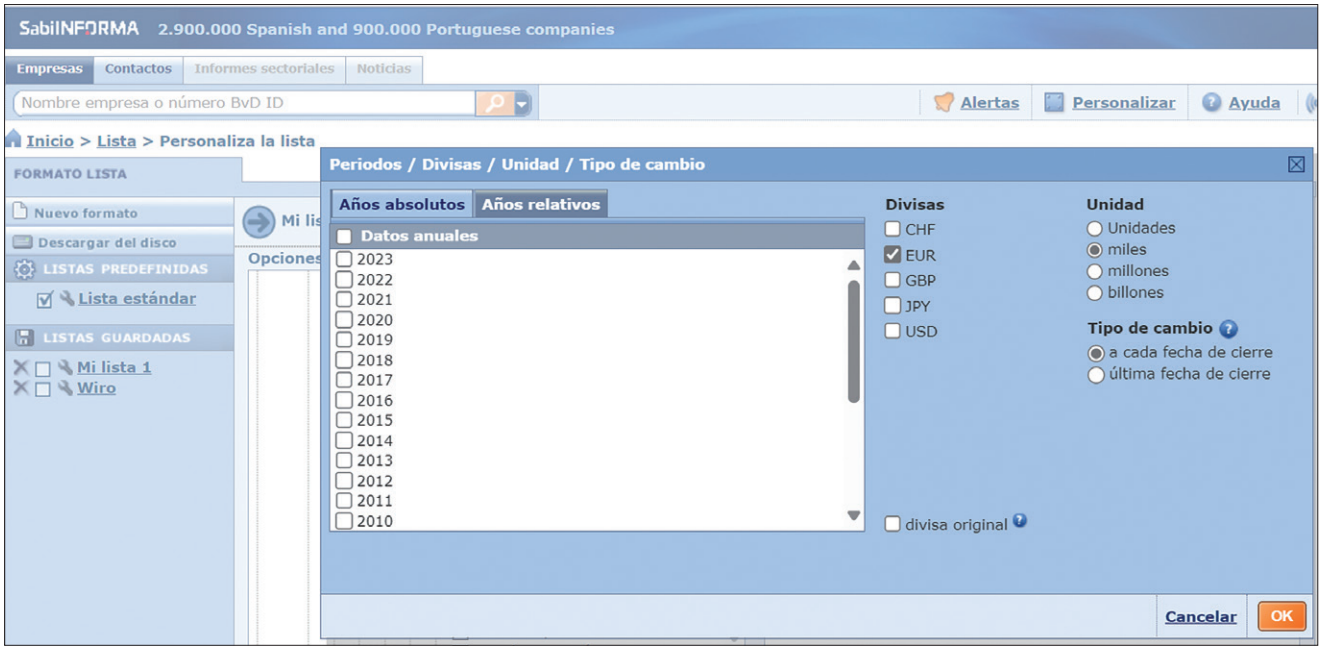
A continuación, en la parte superior derecha se selecciona *Definir el formato* > *Formato de lista* > *Lista estándar*. A la derecha de la tabla de datos, seleccionamos *Añadir*.



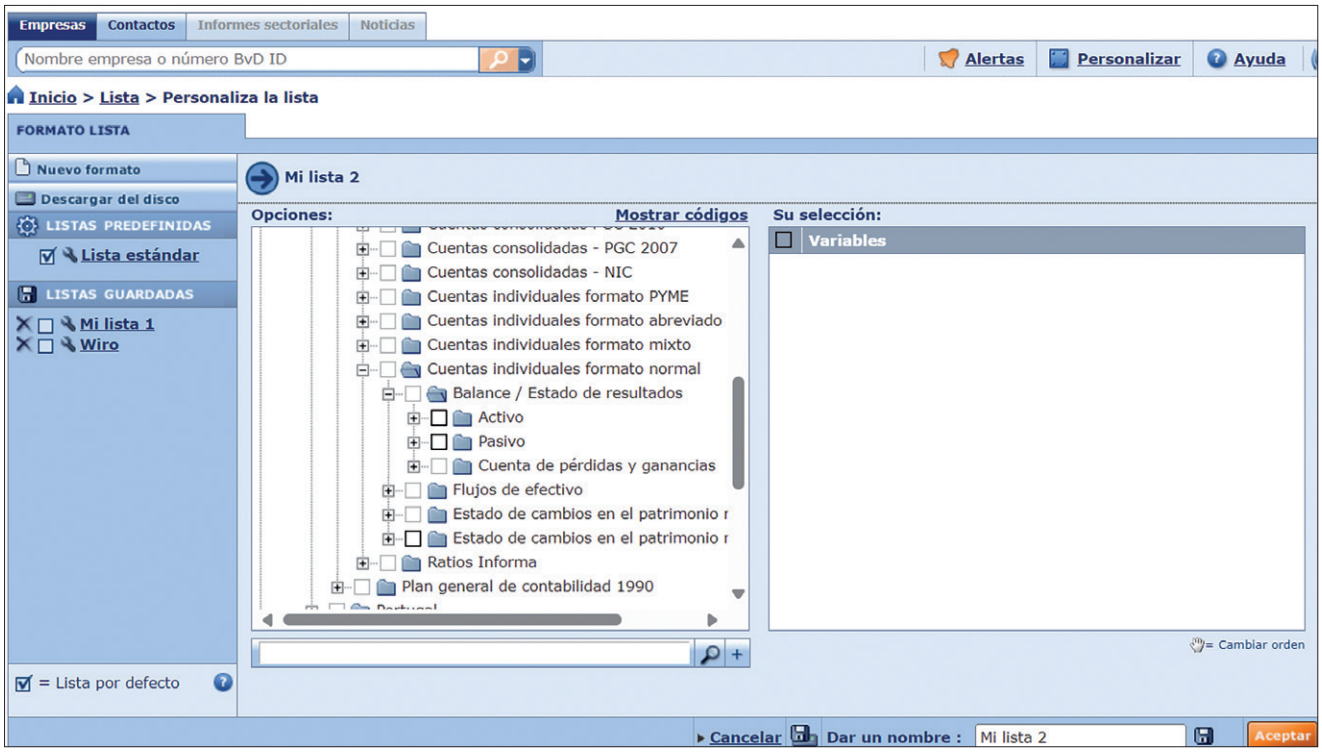
Posteriormente, en la parte izquierda seleccionamos *Datos financieros* > *Formato detallado* > *España* > *Plan General de Contabilidad 2007* > *Cuentas individuales formato normal.*



Tras pedir cada masa patrimonial se debe indicar *Años absolutos* y el año o años deseados. Hay que tener especial cuidado en este punto, ya que la opción por defecto son *Años relativos* por diferencia con el último año disponible, año que puede ser distinto para cada empresa.



Seleccionamos así los valores contables que nos interesan en el estudio y en los ejercicios oportunos, para exportar más adelante el documento a un archivo Excel. Es resaltable que en la base de datos SABI, una vez desplegado el *Balance / Estado de resultados*, x_1 (*Activo no corriente*) y x_2 (*Activo corriente*), se pueden seleccionar directamente; asimismo, la base de datos SABI permite seleccionar en el pasivo x_3 (*Pasivo no corriente*) y x_4 (*Pasivo corriente*). Una vez desplegada la *Cuenta de pérdidas y ganancias / Operaciones continuadas*, se puede seleccionar x_5 (ingresos de explotación), descrito como *Importe neto de la cifra de negocios*. Respecto a x_6 (gastos de explotación), el importe se consigue por diferencia entre los ingresos de explotación y el beneficio o resultado de explotación, ya que la base de datos SABI permite seleccionar el *Resultado de explotación,* pero no los gastos de explotación. Se pueden seleccionar algunas variables no financieras para complementar la información en *Perfil financiero & empleados*, *Información contacto*, *Información legal & cuentas*, *Información grupo & tamaño*, *Marcas*, *Actividades*, *Distribución empleados*, etc.



La hoja Excel descargada debe ser posteriormente editada para ser compatible con el software CoDaPack tal como se detalla en el apartado 4.5.

En función del sector de actividad, o grupo de empresas, para disponer de una muestra robusta podemos recoger empresas que dispongan de información y representen un 75 o 80 % de la facturación de la población total de empresas, como se ha realizado en estudios académicos (e. g., Arimany-Serrat y Coenders, 2025). Si fuera posible disponer de todos los datos de la población, el estudio sería más consistente, pero por diversas causas algunas empresas deben ser descartadas, tal como se indica en los apartados 4.4 (ceros absolutos) y 4.5 (observaciones atípicas).

Finalmente, hay que destacar que el análisis de los estados financieros puede recoger indicadores de otros estados financieros, como los flujos de efectivo de la actividad de explotación, muy útiles en el análisis financiero a corto plazo (*Estado de flujos de efectivo, EFE*) o el *Estado de cambios en el patrimonio neto*, que completa el análisis del patrimonio. Esta información también la facilita la base de datos SABI además de los dos estados financieros usados en este libro (balance y cuenta de resultados).

Por último, es necesario caracterizar el sector de actividad a analizar, buscar información sobre este, y los informes estándar de la base de datos SABI ayudan a este fin; enriquecen la interpretación de los análisis estadísticos realizados. Es recomendable comprobar si se han realizado estudios similares al nuestro o bien estudios que nos ayuden a entender mejor el sector que estamos analizando, utilizando bases de datos bibliográficas como Scopus o Web of Science.

2.3. Para saber más. Otros indicadores útiles

En el análisis de las ratios financieras clásicas a corto y largo plazo, así como en el estudio de los resultados, podemos entrar en más detalle utilizando indicadores adicionales. En primer lugar, se pueden usar más de seis valores positivos de los estados financieros y con ellos definir ratios adicionales (Carreras-Simó y Coenders, 2020). Por ejemplo, se podrían distinguir diversos capítulos de gastos de explotación, las cuentas por pagar a proveedores del resto del pasivo corriente, o las existencias del resto del activo corriente.

En segundo lugar, podemos usar otros estados financieros además del balance y la cuenta de pérdidas y ganancias. Por ejemplo, en el análisis de la solvencia a corto plazo, que mide la capacidad de la empresa para hacer frente a las deudas a corto plazo, para conocer mejor la situación de la tesorería a corto plazo, podemos analizar si el flujo de efectivo de la actividad de explotación, en el ya mencionado EFE, es positivo. En concreto, si los cobros de explotación superan a los pagos de explotación. Ello contribuye a avalar los valores óptimos de la ratio de solvencia a corto plazo utilizada con frecuencia a partir de las cifras del balance. Es decir, ratios calculadas a partir de los datos del EFE corroboran la buena salud de la tesorería. Así pues, el EFE es otro documento clave para analizar la solvencia a corto plazo y evaluar la liquidez, valorando si las entradas de efectivo superan las salidas de efectivo de explotación. Además, también permite determinar la potencial supervivencia de la empresa (Arimany-Serrat y Farreras-Noguer, 2020; Arimany-Serrat et al., 2016; 2022; Bresciani et al., 2016; Rondós-Casas et al., 2018).

En tercer lugar, variables no financieras que tomen valores positivos también pueden formar parte de las ratios. Por ejemplo, el número de empleados puede formar parte de ratios de ventas por empleado, activos por empleado, coste laboral medio por empleado, etc. Cada sector económico puede tener sus variables de interés, por ejemplo, el número de habitaciones en el sector hotelero (Carreras-Simó y Coenders, 2020; Mulet-Forteza et al., 2024).

Por último, incorporar al análisis variables contempladas en las memorias de sostenibilidad, como variables *medioambientales* (agua consumida sobre ingresos de explotación, consumo energético sobre ingresos de explotación, etc.), *sociales* (mujeres empleadas sobre empleo total, tipologías de contratos sobre empleo total, etc.) y de *buen gobierno* (composición del consejo de administración respecto al género, consejeros independientes respecto al total consejeros, etc.), permite corroborar el seguimiento de las Normas Europeas de Información de Sostenibilidad (NEIS), especialmente en caso de grandes empresas, obligadas a divulgar la información de indicadores ESG (*enviromental, social, governance*), ya que la aplicación de estas normas será progresiva. En concreto, el 1 de enero de 2024 para las empresas cotizadas y grandes empresas de más de 500 trabajadores, a partir de 2025 para grandes empresas que superen los 250 trabajadores y sucesivamente para otras empresas de menor tamaño hasta 2027, con voluntad de disponer de una información de sostenibilidad comparable, que se auditará. Con las NEIS, la Comisión Europea espera que se produzca un cambio sustancial en el sistema de información, con voluntad de equiparar la información financiera con la de sostenibilidad (Bastida Vialcanet y Subirats Alcoverro, 2023), y su análisis también pasa por datos composicionales como medidas relativas que son los ejemplos que acabamos de dar y tantos otros indicadores ESG relevantes (Todorov y Simonacci, 2020).

3. Problemas de las ratios clásicas en el análisis estadístico de un conjunto de datos

La tesis principal de este libro es que, en el análisis de los estados financieros de una muestra de empresas, si se utilizan ratios, estas se deben construir mediante la metodología CoDa para salvar los problemas de las ratios financieras clásicas en análisis estadísticos, que son:

- Asimetría.
- No normalidad de las distribuciones.
- No linealidad de las relaciones entre variables.
- Observaciones atípicas extremas.
- Dependencia de los resultados según cuál cifra contable va en el numerador y cuál en el denominador.

El uso de ratios financieras clásicas para diagnosticar la salud financiera de una muestra de empresas requiere de métodos estadísticos fiables. Un primer problema de las ratios financieras es la *asimetría* (Faello, 2015; Frecka y Hopwood, 1983; Linares-Mustarós et al., 2018; Oktaviano et al., 2024; Trejo-Pech et al., 2023), y asociada a ella también la *no normalidad* de las distribuciones (Adcock et al., 2015; Buijink y Jegers, 1986; Deakin, 1976; Iotti et al., 2023; 2024a; 2024b; 2024c; Lueg et al., 2014; Martikainen et al., 1995; McLeay y Omar, 2000; So, 1987; Valaskova et al., 2023).

Asimismo, otros problemas son la *no linealidad* de las relaciones (Balcaen y Ooghe, 2006; Carreras-Simó y Coenders, 2021; Cowen y Hoffer, 1982; Keasey y Watson, 1991) y las *observaciones atípicas extremas* (Deshpande, 2023; Ezzamel y Mar-Molinero, 1990; Frecka y Hopwood, 1983; Kane et al., 1998; Lev y Sunder, 1979; McLeay, 1986; Oktaviano et al., 2024; Watson, 1990), junto con la *dependencia de los resultados de las ratios según la cifra contable del numerador o del denominador* (Coenders et al., 2023a; Frecka y Hopwood, 1983; Linares-Mustarós et al., 2022).

De este modo, los resultados de muchos análisis estadísticos quedan invalidados por estos problemas, y, consecuentemente, también lo son las decisiones de gestión que se toman atendiendo a dichos resultados. Además, estos problemas para las ratios individuales también desencadenan problemas cuando dichas ratios se agrupan en índices compuestos con el análisis factorial y métodos relacionados (Cowen y Hoffer, 1982; Martikainen et al., 1995) y también se encuentran presentes en otras disciplinas científicas que usan ratios en análisis estadísticos (Isles, 2020).

3.1. Consecuencias estadísticas de la asimetría, la no normalidad, la no linealidad, las observaciones atípicas y cuál cifra contable va en el numerador o denominador

Las propiedades de los datos, como la asimetría, la no normalidad, la no linealidad y las observaciones atípicas, así como la forma en que se manejan las ratios (según si una cifra contable va en el numerador o en el denominador), tienen implicaciones importantes en el análisis estadístico y econométrico.

En cuanto a la asimetría, que mide el sesgo de la distribución de los datos, puede ser positiva (cuando la cola de la distribución de los valores altos es más larga) o negativa (en caso contrario), en ambos casos con consecuencias estadísticas. Dicha asimetría va íntimamente asociada a la no normalidad de la distribución. Muchos modelos y contrastes de hipótesis estadísticos asumen que los datos son normales (y, por lo tanto, simétricos) y la no normalidad invalida dichos modelos y contrastes (Adcock et al., 2015; Buijink y Jegers, 1986; Deakin, 1976; Durana et al., 2025; Faello, 2015; Frecka y Hopwood, 1983; Iotti et al., 2023; 2024a; 2024b; 2024c; Linares-Mustarós et al., 2018; Lueg et al., 2014; Martikainen et al., 1995; McLeay y Omar, 2000; So, 1987; Valaskova et al., 2023). Además de los contrastes, si no se tienen en cuenta la asimetría y la no normalidad, las estimaciones, intervalos de confianza y predicciones de dichos modelos no son fiables. La asimetría deriva a menudo del hecho que las ratios tienen un límite inferior igual a cero, es decir, pueden variar de cero a infinito. Esto tiene una consecuencia indeseable adicional. Puede ocurrir que las ratios previstas en modelos estadísticos tengan, a pesar de ello, valores negativos y, por lo tanto, imposibles. En el caso del análisis clúster, la asimetría desemboca en clústeres muy grandes acompañados de otros muy pequeños (Dao et al., 2024; Feranecová y Krigovská, 2016; Jofre-Campuzano y Coenders, 2022; Linares-Mustarós et al., 2018; Lukason y Laitinen, 2019; Sharma et al., 2016).

La no linealidad de las relaciones entre variables afecta de manera especial al análisis de la correlación y la regresión (Balcaen y Ooghe, 2006; Carreras-Simó y Coenders, 2021; Cowen y Hoffer, 1982; Keasey y Watson, 1991) y cuando diversas ratios se agregan formando índices, para lo cual se toman precisamente como punto de partida las correlaciones (Cowen y Hoffer, 1982; Martikainen et al., 1995). Cuando las relaciones no son lineales, las correlaciones son inferiores, y los modelos de regresión adolecen de una bondad de ajuste inferior, sus previsiones son sesgadas y sus contrastes de hipótesis son incorrectos.

Es de sobra conocido que las observaciones atípicas tienen un impacto considerable en cualquier método estadístico clásico, incluso para el cálculo de la simple media sectorial de una ratio clásica o la correlación entre una ratio y una variable no financiera, por no hablar de lo que ocurre cuando se aplican modelos estadísticos algo más complejos (Deshpande, 2023; Ezzamel y Mar-Molinero, 1990; Frecka y Hopwood, 1983; Gupta, 2024; Kane et al., 1998; Lev y Sunder, 1979; Liu et al., 2025; McLeay, 1986; Nyitrai y Virág, 2019; Oktaviano et al., 2024; Watson, 1990). En todos los casos, la presencia o ausencia de una sola observación atípica extrema puede modificar los resultados hasta el punto de afectar las conclusiones que de ellos se extraigan. En el caso del análisis clúster, desembocan en clústeres completamente configurados en torno a las observaciones atípicas (Dao et al., 2024; Feranecová y Krigovská, 2016;

Jofre-Campuzano y Coenders, 2022; Linares-Mustarós et al., 2018; Sharma et al., 2016). A menudo, las observaciones atípicas surgen más por culpa de un pequeño valor en el denominador que por un valor elevado en el numerador. Un ejemplo es la rentabilidad financiera (ROE) que es más sensible si el patrimonio neto está cerca de cero, y pequeñas fluctuaciones en dicho patrimonio pueden provocar grandes cambios en el ROE, hecho inspirador de la imagen de portada del libro, donde el ROE que se calcula en la pizarra sería del 200 000 %.

La dependencia de los resultados según qué cifra contable va en el numerador o en el denominador de la ratio es una auténtica carga de profundidad para el análisis estadístico de ratios clásicas, dada la frecuencia con la que se definen ratios que son una simple permutación de otras, como hemos visto con la solvencia a largo plazo y el endeudamiento de las ecuaciones (7) y (9) (Coenders et al., 2023a; Frecka y Hopwood, 1983; Linares-Mustarós et al., 2022). Realmente, el endeudamiento y la solvencia son dos caras de la misma moneda. Una empresa con endeudamiento igual a ⅓ tendrá necesariamente una solvencia igual a 3, con lo que para el estudio de una empresa individual la permutación no tiene consecuencia alguna. Por ello, los investigadores que usan una versión de la ratio u otra en un análisis estadístico lo hacen a menudo sin sospechar que sus conclusiones estadísticas pueden ser totalmente diferentes. Sin ir más lejos, la correlación entre una variable no financiera y la ratio permutada siempre será diferente de la correlación obtenida con la ratio original, y no solo en su signo. Asimismo, ocurre casi siempre que observaciones que son atípicas en la ratio original aparecen como no atípicas en la ratio permutada, y a la inversa (Coenders et al., 2023a; Linares-Mustarós et al., 2022).

Cada uno de estos factores puede alterar la validez de un análisis estadístico o econométrico basado en ratios financieras clásicas si no se tratan estos problemas utilizando la metodología CoDa. A pesar de ello, es comprensible que al principio no se les prestara atención. Al nacer las ratios financieras a fines del siglo XIX (Brown Sister, 1955; Horrigan, 1968), el análisis estadístico estaba en sus inicios y la *teoría de las escalas de medición*, que dicta los tipos de datos y las operaciones matemáticas y estadísticas válidas sobre las ratios, no había ni siquiera nacido (Stevens, 1946). Esta situación ha cambiado drásticamente, y actualmente se dispone de un vasto cuerpo de investigación estadística y econométrica en el campo de la contabilidad (Gruszczyński, 2022). De hecho, el primer uso del término *econometría* fue hecho por Paweł Ciompa en 1910 en el campo de la contabilidad (Ciompa, 1910). Por lo tanto, es poco comprensible que en pleno siglo XXI siga sin prestarse atención a los problemas mencionados de las ratios clásicas. A pesar de que la literatura existente desde los años 1970 y que hemos ido citando los mostraba claramente, sus voces fueron ignoradas por la gran mayoría de los profesionales y los investigadores aplicados. A lo sumo, se fueron aplicando soluciones parciales de manera *ad hoc*. Las observaciones atípicas se resolvieron en algunos casos simplemente eliminándolas aun a costa de desprenderse de una parte de la información contenida en la muestra, o substituyéndolas por la observación no atípica más cercana (Demiraj et al., 2024; Deshpande, 2023; Ezzamel y Mar-Molinero, 1990; Frecka y Hopwood, 1983; Gupta, 2024; Lev y Sunder, 1979; Liu et al., 2025; Martikainen et al., 1995; Naz et al., 2023; Nyitrai y Virág, 2019; So, 1987; Vu et al., 2023; Watson, 1990). En cuanto a la asimetría, se realizaron transformaciones a menudo difícilmente interpretables como raíces cuadradas y cúbicas o rangos (Deakin, 1976; Ezzamel y Mar-Molinero, 1990; Frecka y Hopwood, 1983; Kane et al., 1998; Lueg et al.,

2014; Martikainen et al., 1995; Mcleay y Omar, 2000; Watson, 1990) y la no normalidad se trató con contrastes estadísticos no paramétricos (Durana et al., 2025; Hazami-Ammar, 2024; Iotti et al., 2023; 2024a; 2024b; 2024c; Latief y Suhendah, 2023; Valaskova et al., 2023) o modelos estadísticos más complejos (Adcock et al., 2015; Trejo-Pech et al., 2023), de modo que se limitó la elección de los métodos estadísticos que había que utilizar.

En este libro se presenta un enfoque simple y unificado para todos los problemas simultáneamente, compatible con cualquier método estadístico y no solo los no paramétricos. Las ratios financieras como cifras contables relativas en lugar de absolutas suponen un campo natural de aplicación del análisis de datos composicionales (CoDa), por ajustarse al análisis de los estados financieros, tratando las cifras contables de manera simétrica, con resultados que no dependen de la permutación del numerador y el denominador, y reduciendo las observaciones atípicas, además de linealizar las relaciones. No se trata de un lucimiento metodológico. Las ratios que utilizan la metodología CoDa siempre han presentado resultados muy diferentes a las ratios financieros clásicas cuando se han comparado unas con otras (Arimany-Serrat et al., 2022; Carreras-Simó y Coenders, 2021; Coenders et al., 2023a; Creixans-Tenas et al., 2019; Dao et al., 2024; Escaramís y Arbussà, 2025; Jofre-Campuzano y Coenders, 2022; Linares-Mustarós et al., 2018; 2022), y en este libro el mismo lector o lectora será capaz de comparar los dos tipos de resultados y evidenciar sus diferencias.

Desde sus inicios en la década de 1980, CoDa se aplicó a diferentes campos de las ciencias naturales y sociales, y hoy en día ha alcanzado un elevado grado de madurez, evidenciado por su presencia en libros de texto (Van den Boogaart y Tolosana-Delgado, 2013; Filzmoser et al., 2018; Greenacre, 2018; Pawlowsky-Glahn et al., 2015) y software (Calle et al., 2023; Comas-Cufí y Thió-Henestrosa, 2011; Van den Boogaart y Tolosana-Delgado, 2013; Filzmoser et al., 2018; Greenacre, 2018; Palarea-Albaladejo y Martín-Fernández, 2015; Thió-Henestrosa y Martín-Fernández, 2005). Tras cuarenta años, sigue siendo de rabiosa actualidad y sujeta a nuevos desarrollos (Coenders et al., 2023b; Greenacre et al., 2023).

En este libro se muestra cómo aplicar la metodología CoDa en el caso concreto del análisis de los estados financieros, simplificando y adaptando los métodos CoDa para explotar sus ventajas en la comunidad contable hoy en día.

3.2. Para saber más. Lecturas complementarias

Existe diverso material publicado sobre la metodología CoDa con diversos grados de dificultad. Para un primer contacto con la metodología, recomendamos el capítulo de Ferrer-Rosell et al. (2022) en el campo de la comunicación, el capítulo de Ferrer-Rosell et al. (2021) en el campo del marketing, o el artículo de Coenders y Ferrer-Rosell (2020) en el campo del turismo. Existe poco material en español; en esta lengua recomendamos el artículo de Egozcue y Pawlowsky-Glahn (2016), aunque en algunos aspectos el nivel es superior al que presentamos en este libro.

Aparte de estas breves introducciones, existen diversos manuales que presuponen al lector o lectora un alto nivel de estadística (Van den Boogaart y Tolosana-Delgado, 2013; Filzmoser et al., 2018; Greenacre, 2018; Pawlowsky-Glahn et al., 2015). Recomendamos en especial los dos últimos de la lista como algo más accesibles.

Otra vía de entrada al tema son los artículos de investigación aplicada. Existen ya numerosas aplicaciones de la metodología composicional en sectores diversos como el farmacéutico (Linares-Mustarós et al., 2018), el de corte y confección (Linares-Mustarós et al., 2018), el hospitalario (Creixans-Tenas et al., 2019), el de distribución alimentaria (Carreras-Simó y Coenders, 2020), el de comercio al por menor (Carreras-Simó y Coenders, 2021), el vinícola (Arimany-Serrat et al., 2022; 2023; Coenders, 2025; Linares-Mustarós et al., 2022), el cervecero (Arimany-Serrat y Sgorla, 2024; Coenders et al., 2023a), el turístico (Mulet-Forteza et al., 2024; Saus-Sala et al., 2021; 2023; 2024), el apícola (Arimany-Serrat y Coenders, 2025), el pesquero (Dao et al., 2024), el conservero (Dao et al., 2024) y el gasolinero (Jofre-Campuzano y Coenders, 2022), y tiene potencial para aplicarse a cualquier otro sector.

Por último, las referencias de cabecera sobre los problemas de las ratios financieras clásicas cuando se analizan estadísticamente son las de Linares et al. (2018; 2022). Un resumen en español se encuentra en Coenders et al. (2023a).

4. Análisis de los estados financieros utilizando datos composicionales

4.1. Los estados financieros como composición

El análisis composicional surgió en los campos de la química y la geología para estudiar la composición química de rocas o muestras de tipo diverso (Aitchison, 1982; 1986; Buccianti et al., 2006), aunque tras más de cuarenta años de desarrollo (Coenders et al., 2023b; Greenacre et al., 2023) se encuentre presente en casi todas las ramas de las ciencias, incluidas las ciencias económicas y sociales (Coenders y Ferrer-Rosell, 2020; Fry, 2011; Martínez-García et al., 2023). Así, se ha venido aplicando desde hace tiempo en distintas ramas de las finanzas, como los seguros (Belles-Sampera et al., 2016; Boonen et al., 2019; Gan y Valdez, 2021; Verbelen et al., 2018), la microfinanciación (Davis et al., 2017), el riesgo sistémico (Fiori y Coenders, 2025; Fiori y Porro, 2023; Porro, 2022), las finanzas familiares (Fry et al., 1996; 2000; 2001; Gokhale et al., 2024; Mclaren et al., 1995; Tian et al., 2024), los tipos de cambio (Gámez-Velázquez y Coenders, 2020; Maldonado et al., 2021a; 2021b), los portafolios (Glassman y Riddick, 1996; Joueid y Coenders, 2018; Vega-Gámez y Alonso-González, 2024), la estructura de propiedad de los fondos propios (Ahmed et al., 2023), los mercados financieros (Kokoszka et al., 2019; Li et al., 2019; Ortells et al., 2016; Vega-Baquero y Santolino, 2022a; Wang et al., 2019), la banca (Vega-Baquero y Santolino, 2022b), la calificación de bonos (Tallapally, 2009) y los presupuestos municipales (Voltes-Dorta et al., 2014).

Paralelamente a la expansión a ámbitos de conocimiento diversos, el énfasis inicial de la metodología CoDa sobre el análisis de las partes de un todo en análisis químicos se ha trasladado al análisis de ratios (Egozcue y Pawlowsky-Glahn, 2019), con lo cual su aplicabilidad al análisis de los estados financieros es inmediata.

Una composición en D partes se define como un conjunto de D números estrictamente positivos, llamados *partes*, cuya magnitud relativa es de interés para el investigador o investigadora (Aitchison, 1986):

$$x_1, x_2, \ldots, x_D \text{ con } x_j > 0, j = 1,2,\ldots,D \tag{13}$$

La mencionada importancia relativa de las partes se encuentra recogida en sus ratios (Egozcue y Pawlowsky-Glahn, 2019). La composición es la materia prima con la que trabaja la metodología CoDa. Aunque en otros campos científicos

que usan composiciones, las D partes a veces tienen suma constante (en un análisis químico las partes suman el 100 % del peso, del volumen o de las moléculas), esto no es de ninguna manera un requisito. Igualmente, con los estados financieros no suele suceder que la suma sea constante.

Solo deben seguirse dos reglas para introducir valores contables extraídos de los estados financieros en una composición de D partes, que consisten en evitar los valores negativos y su superposición (Coenders y Arimany-Serrat, 2023; Creixans-Tenas et al., 2019):

- Aunque a veces las ratios financieras implican valores contables que pueden ser negativos, la literatura financiera desaconseja su uso, ya que pueden provocar una discontinuidad, observaciones atípicas o incluso una inversión de la interpretación cuando el valor contable que puede ser negativo está en el denominador (Lev y Sunder, 1979; Linares-Mustarós et al., 2022). Por ejemplo, según la ecuación (6), una empresa con beneficio de explotación y patrimonio neto ambos negativos tendría un ROE aparentemente positivo. Esta paradoja también ha inspirado la imagen de portada del libro.
  Los valores contables negativos también se desaconsejan desde el punto de vista de la teoría de las escalas de medición. El cálculo de una ratio es una operación que tiene sentido solo para las variables medidas en una *escala de razón*, que deben tener un cero absoluto (Stevens, 1946), cosa que excluye los valores negativos.
  En general, los valores contables son negativos porque llevan implícita la resta de otros valores contables positivos, que son los que hay que utilizar. Esto significa, por ejemplo, que, al construir ratios, se deben utilizar directamente los ingresos y los gastos en lugar del beneficio, los activos y pasivos totales en lugar del patrimonio neto, o los activos y pasivos corrientes en lugar del fondo de maniobra. Esta limitación no implica pérdida de información de ningún tipo. Por ejemplo, una ratio que transmita la misma información que la ratio de *margen* clásica (beneficio de explotación sobre ingresos de explotación) puede construirse únicamente a partir de los valores no negativos de ingresos y gastos. Sean x_5= ingresos de explotación, x_6 = gastos de explotación y $x_7 = x_5 - x_6$ = beneficio de explotación. La ratio siempre positiva de ingresos sobre gastos (x_5/x_6) es una simple transformación de la ratio problemática de beneficios sobre ingresos (x_7/x_5) de la ecuación (3). Cuando aumenta x_7/x_5, también lo hace x_5/x_6:

$$\frac{x_5}{x_6} = \frac{x_5}{x_5 - x_7} = \frac{1}{\frac{x_5 - x_7}{x_5}} = \frac{1}{1 - \frac{x_7}{x_5}} \tag{14}$$

- También hay que tener en cuenta que las partes no pueden superponerse. Por ejemplo, no se podría usar x_8 = activos totales y x_1 = activos no corrientes porque x_1 está incluido en x_8. En la terminología composicional, x_8 = activos totales es una *amalgama* de x_1 = activos no corrientes y x_2 = activos corrientes. Usar a la vez amalgamas y sus partes constituyentes es extremadamente problemático (Pawlowsky-Glahn et al., 2015). Más bien, la elección entre usar solo la amalgama o solo las partes individuales debe hacerse en la etapa de definición del problema y no se puede cambiar posteriormente (Van den Boogaart y Tolosana-Delgado, 2013). No es esencial utilizar todas las partes constituyentes, lo que se conoce como *subcomposición* en la terminología de

datos composicionales. En consecuencia, las opciones factibles para manejar x_1, x_2 y x_8 son: a) usar solo x_8; b) usar x_1 y x_2, c) usar solo x_1; y d) usar solo x_2.

Los siguientes $D = 6$ valores extraídos de los estados financieros en el capítulo 2 se usan como partes en casi todo el libro y cumplen los requisitos mencionados de no negatividad y no superposición:

- x_1: activo no corriente
- x_2: activo corriente
- x_3: pasivo no corriente
- x_4: pasivo corriente
- x_5: ingresos de explotación
- x_6: gastos de explotación

Estas seis partes son las más usadas hasta el momento en el análisis de los estados financieros con la metodología CoDa (Arimany-Serrat y Coenders, 2025; Arimany-Serrat et al., 2023; Coenders, 2025; Creixans-Tenas et al., 2019; Dao et al., 2024; Jofre-Campuzano y Coenders, 2022; Saus-Sala et al., 2024).

4.2. Log-ratios por pares

El enfoque habitual para el análisis CoDa es utilizar métodos estadísticos ya existentes sobre datos transformados. Los *logaritmos de las ratios*, abreviadamente *log-ratios*, son la transformación estándar en CoDa (Aitchison, 1986). El caso más simple de una log-ratio es el que se toma solo dos valores contables (*log-ratios por pares*, e. g., Creixans-Tenas et al., 2019; Greenacre, 2018; 2019; Mulet-Forteza et al., 2024; Saus-Sala et al., 2021) y también se puede entender como la diferencia logarítmica entre los dos valores contables de partida:

$$\log\left(\frac{x_1}{x_2}\right) = \log\left(x_1\right) - \log\left(x_2\right) \tag{15}$$

A diferencia de una ratio clásica, que está delimitada entre cero e infinito, una log-ratio es simétrica en el sentido de que su rango o recorrido es de menos infinito a más infinito, toda la *recta real*, lo que convierte la log-ratio en una *variable real*. Esto tiene dos ventajas clave:

- Por un lado, coincide con el rango de la distribución de probabilidad normal, como sabemos muy usada en estadística. Nada garantiza que una variable real vaya a estar distribuida normalmente, pero sí está garantizado que una variable delimitada entre cero e infinito no puede ser nunca normal.
- Por otro lado, las previsiones en regresión lineal (véase el capítulo 8) también van de menos infinito a más infinito, como hemos indicado en el apartado 3.1. Al ajustar una ratio clásica como variable dependiente en un modelo de regresión lineal, algunos valores previstos podrían ser valores imposibles para una ratio clásica, es decir, por debajo de cero. Por el contrario, la previsión de una log-ratio nunca será un valor imposible

de obtener para una log-ratio. Algunas ratios financieras clásicas son, en realidad, fracciones de un total y también tienen un límite superior igual a 1, lo que agrava el problema, al ser también imposibles las previsiones superiores a 1.

Además, una log-ratio es simétrica en el sentido de que la permutación de las partes del numerador y del denominador conduce a la misma distancia de cero y no afecta a ninguna otra propiedad de la log-ratio que su signo (Linares-Mustarós et al., 2022):

$$\log\left(\frac{x_1}{x_2}\right) = \log\left(x_1\right) - \log\left(x_2\right) = -\left(\log\left(x_2\right) - \log\left(x_1\right)\right) = -\log\left(\frac{x_2}{x_1}\right) \quad (16)$$

Por ejemplo, la correlación de un indicador externo no financiero con una log-ratio permutada es igual a la correlación con la log-ratio original con el signo invertido como hemos indicado en el apartado 3.1. Esta propiedad no es válida para las ratios financieros clásicas (Frecka y Hopwood, 1983). Correlacionar x_1/x_2 con un indicador no financiero puede dar resultados contradictorios comparado con correlacionarlo con x_2/x_1 (Coenders et al., 2023a; Linares-Mustarós et al., 2022). No hay otra razón que el acuerdo para usar una ratio o su permutación. Para una sola empresa, el hecho de que $x_1/x_2 = 0{,}5$ proporciona la misma información que el hecho de que $x_2/x_1 = 2$. Sin embargo, en los análisis estadísticos en el ámbito sectorial, los resultados de una y otra ratio pueden ser contradictorios. Tal como hemos visto en el capítulo 2, estas permutaciones son corrientes en la práctica; el ejemplo más habitual son las ratios de solvencia a largo plazo y endeudamiento.

Por último, si uno de los valores contables que se comparan en la ratio es cercano a cero, puede dar lugar a una ratio clásica atípica cuando se coloca en el denominador y a una ratio típica cuando se coloca en el numerador (Ezzamel y Mar-Molinero, 1990; Frecka y Hopwood, 1983; Kane et al., 1998; Lev y Sunder, 1979; Martikainen et al., 1995). Para las log-ratios el resultado será el mismo (Coenders et al., 2023a; Linares-Mustarós et al., 2018; 2022; Molas-Colomer et al., 2024) y, en general, se dan muchas menos observaciones atípicas extremas (Coenders et al., 2023a; Dao et al., 2024; Linares-Mustarós et al., 2018; 2022) que con las ratios clásicas (Ezzamel y Mar-Molinero, 1990; Frecka y Hopwood, 1983; Kane et al., 1998; Lev y Sunder, 1979; McLeay, 1986; Watson, 1990).

La tabla 1 muestra un ejemplo sencillo de siete empresas ficticias y dos valores contables, x_1 y x_2. Para facilitar el cálculo, mostramos los *logaritmos en base 10* representados como $\log_{10}(x)$, que solo dicen cuántas veces 10 tiene que multiplicarse por sí mismo para obtener el valor deseado. Por ejemplo, $\log_{10}(1\,000\,000) = 6$ porque $10^6 = 1\,000\,000$:

$$\log_{10}\left(10^x\right) = x \quad (17)$$

Dado que $10^0 = 1$, $\log_{10}(1) = 0$. Su simetría en torno a 1 y 0 se constata con ejemplos como este: $0{,}000001 = 1/1\,000\,000 = 1/10^6 = 10^{-6}$, con lo que $\log_{10}(0{,}000001) = -6$.

La tabla 1 muestra, en primer lugar, cómo interpretar las log-ratios. En la empresa 4 se cumple que $x_1 = x_2$. En consecuencia, la ratio clásica x_1/x_2 es igual

a 1 y la log-ratio correspondiente $\log_{10}(x_1/x_2)$ es igual a 0. En las empresas 1 a 3 $x_1 > x_2$. En consecuencia, la ratio clásica x_1/x_2 es mayor que 1 y la log-ratio $\log_{10}(x_1/x_2)$ es positiva. En las empresas 4 a 7 ocurre lo contrario: $x_1 < x_2$. En consecuencia, la ratio clásica x_1/x_2 es menor que 1 y la log-ratio $\log_{10}(x_1/x_2)$ es negativa. Cuando la ratio clásica aumenta, también lo hace la log-ratio.

Al igual que las ratios, los logaritmos se centran en las diferencias relativas entre las empresas. Es por ello por lo que las ratios y los logaritmos son mutuamente compatibles (Stevens, 1946) y deben usarse juntos de manera rutinaria para datos en una escala de razón, cuando la diferencia que importa entre dos valores es la diferencia relativa, lo que significa que radica en su cociente (ratio) y no en su resta.

Por ejemplo, si tomamos las empresas 3, 4 y 5 en la tabla 1 (valores de x_2 100, 1 000 y 10 000), en términos relativos, la diferencia relativa entre 1 000 y 100, que es 1 000/100 = 10, es la misma que la diferencia relativa entre 10 000 y 1 000, que es 10 000/1 000 = 10. En consecuencia, sus diferencias logarítmicas 3 – 2 = 1 y 4 – 3 = 1 son las mismas. Solo una vez aplicado el logaritmo, la resta vuelve a tener sentido. La resta es una operación esencial en estadística. Por ejemplo, el residuo de un modelo es el valor real menos el valor previsto, la varianza de una variable se basa en la resta de la media a cada uno de los valores, etc.

Nótese que los valores x_1 y x_2 en la tabla 1 son perfectamente simétricos, mientras que las ratios clásicas x_1/x_2 y x_2/x_1 no son simétricas en absoluto. En la ratio x_1/x_2 las empresas 4 a 7 están concentradas en el reducido intervalo entre 0 y 1, mientras que las empresas 1 a 3 alcanzan valores elevadísimos. En la ratio x_1/x_2 las empresas 1 y 2 aparecen como atípicas con valores de 10 000 y 1 000 000, debido a la pequeñez del denominador, y en la ratio x_2/x_1 les ocurre lo mismo a las empresas 6 y 7.

Por el contrario, los logaritmos de las ratios $\log_{10}(x_2/x_1)$ y $\log_{10}(x_1/x_2)$ son completamente simétricos, no tienen observaciones atípicas (observaciones que se alejen más del resto de lo que el resto se alejan entre ellas), y la permutación del numerador y el denominador solo conduce a una inversión del signo.

Empresa	x_1	x_2	x_2/x_1	x_1/x_2	$\log_{10}(x_1)$	$\log_{10}(x_2)$	$\log_{10}(x_2/x_1)$	$\log_{10}(x_1/x_2)$
1	1 000 000	1	0,000001	1 000 000	6	0	–6	6
2	100 000	10	0,0001	10 000	5	1	–4	4
3	10 000	100	0,01	100	4	2	–2	2
4	1 000	1 000	1	1	3	3	0	0
5	100	10 000	100	0,01	2	4	2	–2
6	10	100 000	10 000	0,0001	1	5	4	–4
7	1	1 000 000	1 000 000	0,000001	0	6	6	–6

Tabla 1. Ejemplo simplificado con siete empresas y dos valores contables

Los *logaritmos naturales*, llamados también *logaritmos neperianos* (es decir, los *logaritmos en base e* = 2,718281828…), representados en adelante como $\log(x)$, son más comunes en economía y finanzas. Son los que se utilizan en la mayoría de los softwares de CoDa y se utilizarán en el libro a partir de ahora, pero se podría utilizar cualquier otra base sin afectar a las propiedades del

análisis CoDa. Las propiedades ya vistas de los logaritmos y alguna adicional se presentan aquí, ya particularizadas para los logaritmos en base *e*:

$$\log\left(e^{x}\right) = x$$
$$e^{\log(x)} = x$$
$$\log\left(\frac{1}{x}\right) = -\log(x)$$
$$\log\left(\frac{x_1}{x_2}\right) = \log\left(x_1\right) - \log\left(x_2\right)$$
$$\log\left(x_1 x_2\right) = \log\left(x_1\right) + \log\left(x_2\right)$$
$$\log\left(\frac{x_1}{x_2}\right) = -\log\left(\frac{x_2}{x_1}\right)$$
$$\log\left(x_1^{x_2}\right) = x_2 \log\left(x_1\right) \quad (18)$$

Algunas log-ratios por pares basadas en las partes x_1 a x_6 se relacionan con las ratios clásicas presentadas en el capítulo 2 (Creixans-tenas et al., 2019). La rotación del activo corriente compara los ingresos con el activo corriente y sustituye la ecuación (2):

$$y_1 = \log\left(\frac{x_5}{x_2}\right) \quad (19)$$

La comparación de ingresos y gastos proporciona una noción de margen, tal como se ha visto en la ecuación (14), y sustituye la ecuación (3):

$$y_2 = \log\left(\frac{x_5}{x_6}\right) \quad (20)$$

La comparación del activo y el pasivo corrientes indica solvencia a corto plazo, como lo hace la ecuación (10):

$$y_3 = \log\left(\frac{x_2}{x_4}\right) \quad (21)$$

La comparación de los activos no corriente y corriente indica la inmovilización del activo, igual que la ecuación (11):

$$y_4 = \log\left(\frac{x_1}{x_2}\right) \quad (22)$$

La comparación de los pasivos no corriente y corriente indica la maduración o vencimiento de la deuda, como la ecuación (12):

$$y_5 = \log\left(\frac{x_3}{x_4}\right) \tag{23}$$

Potencialmente, se pueden calcular $D(D-1)/2$ diferentes log-ratios por pares, aunque algunas de ellas pueden no tener ninguna interpretación financiera o interés teórico, con lo que la elección de las log-ratios se convierte en un asunto potencialmente problemático. También se debe tener mucho cuidado para evitar que las log-ratios sean mutuamente redundantes, lo que significa que la información de algunas log-ratios ya esté contenida en otras (Barnes, 1987; Chen y Shimerda, 1981; Pohlman y Hollinger, 1981). Como ejemplo de elección de log-ratio redundante, se podría considerar agregar la log-ratio $y_6 = \log\left(x_3 / x_1\right)$ para indicar en qué medida el activo no corriente está siendo financiado por pasivos no corrientes. Lo que ocurre es que y_6 ya está contenida en las otras log-ratios. En particular, y_6 se puede obtener como $y_5 - y_4 - y_3$:

$$\begin{aligned} y_5 - y_4 - y_3 &= \log\left(\frac{x_3}{x_4}\right) - \log\left(\frac{x_1}{x_2}\right) - \log\left(\frac{x_2}{x_4}\right) = \\ &\log\left(x_3\right) - \log\left(x_4\right) - \left(\log\left(x_1\right) - \log\left(x_2\right)\right) - \left(\log\left(x_2\right) - \log\left(x_4\right)\right) = \\ &\log\left(x_3\right) - \log\left(x_1\right) = \log\left(\frac{x_3}{x_1}\right) = y_6 \end{aligned} \tag{24}$$

Algunas pautas para evitar la redundancia en las log-ratios por pares se encuentran en Greenacre (2019) y fueron aplicadas por Creixans-Tenas et al. (2019) y Coenders y Arimany-Serrat (2023) en el contexto de los estados financieros. Greenacre (2019) recomienda dibujar un *grafo* en el que los valores de los estados financieros son los *vértices* (*nodos*), y las log-ratios son las *aristas* (*arcos*). El grafo debe ser necesariamente *conexo* y *acíclico*. Esto significa que:

- Es posible unir entre sí dos valores contables cualesquiera siguiendo los arcos (es decir, las log-ratios).
- No puede haber circuitos cerrados. Es decir, al seguir los arcos del grafo desde un valor contable hasta cualquier otro, no se puede visitar ningún valor dos veces. En otras palabras, solo hay un camino posible para unir dos valores contables cualesquiera.

Los arcos se pueden dibujar como flechas sin afectar las propiedades del grafo. Las flechas apuntan al numerador de la log-ratio solo con fines ilustrativos. En otras palabras, los valores contables se consideran unidos incluso cuando para unirlos hay que ir en contra del sentido de las flechas.

Se puede demostrar por contradicción que tal grafo tiene exactamente $D-1$ arcos (log-ratios). Si tiene menos arcos, no puede conectar todos los valores contables, y si tiene más arcos, tiene que haber un ciclo (Greenacre, 2019). Se puede, asimismo, demostrar que las $D-1$ log-ratios por pares así elegidas contienen

toda la información sobre las D partes composicionales iniciales, es decir, toda la información sobre la importancia relativa de los D valores contables.

Si bien cualquier grafo que cumpla con estas condiciones funcionará, estadísticamente hablando, es una buena práctica trazar un grafo con interpretación financiera, basado en el conocimiento experto o que aporte luz al objetivo de la investigación. Las $D - 1 = 6 - 1 = 5$ log-ratios y_1 a y_5 (ecuaciones (19) a (23)), que, como hemos visto aquí y en el apartado 2.1, son indicadores interpretables de resultados y situación financiera, cumplen las condiciones de acuerdo con el grafo acíclico conexo en el panel superior de la figura 1.

Como ejemplo de elección inapropiada de las log-ratios, al agregar y_6 (ecuación (24)) al grafo, se crea un ciclo (figura 1, panel central). Hay dos formas de unir x_4 con x_1: a través de x_2 y a través de x_3 (recuérdese que no es necesario seguir los sentidos de las flechas). Además, x_1, x_2, x_4, y x_3 definen un ciclo cerrado de vuelta a x_1. El panel inferior de la figura 1 muestra un ejemplo de grafo no conexo, incluso si el número de log-ratios es correcto en $D - 1 = 5$. No hay forma de unir, por ejemplo, x_1 y x_6. El ciclo del panel central persiste.

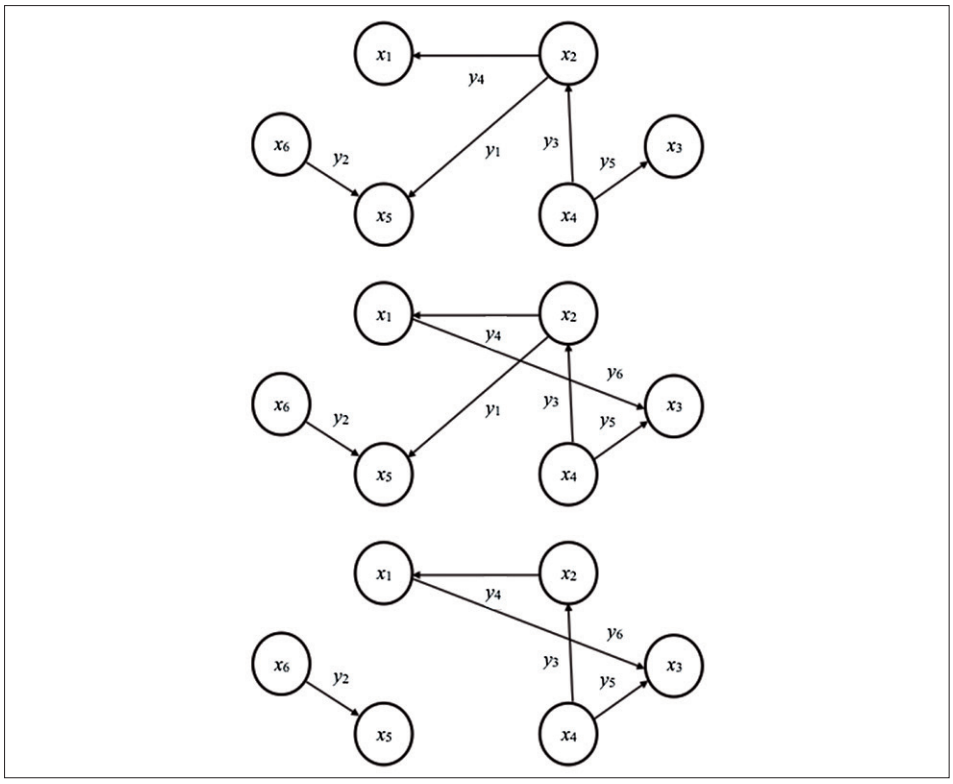


Figura 1. Grafo acíclico y conexo (superior), grafo cíclico conexo (centro), grafo cíclico inconexo (inferior)

Se debe advertir a los usuarios que puede haber más de una forma de elegir un conjunto sensato de $D - 1$ log-ratios por pares interpretables y no redundantes, y los resultados de algunos análisis estadísticos que se basan en distancias (por ejemplo, *biplots*, *análisis en componentes principales* y *análisis clúster*, como se utilizan en los capítulos 6 y 7) dependen de esta elección (Hron et al., 2021). Estos métodos estadísticos basados en la distancia requieren log-ratios alternativas, como se muestra en el apartado 4.3. Por el contrario, las log-ratios por pares son apropiadas como variables de entrada para los modelos estadísticos como la regresión (capítulo 8).

4.3. Log-ratios centradas

Las log-ratios por pares no son la única posibilidad en la metodología CoDa. Esta metodología puede prescindir por completo de la elección manual de las log-ratios al garantizar que las D llamadas *log-ratios centradas* o clr (Aitchison, 1983) también contengan toda la información sobre la importancia relativa de los D valores contables. Cualquier log-ratio que pueda interesar al investigador o investigadora es una función de estas D log-ratios centradas. Cada log-ratio centrada compara una parte, en el numerador, con la *media geométrica* de todas las partes para cada empresa individual, en el denominador. Las clr no tienen interpretación financiera en sí mismas, pero se utilizan como datos de entrada en métodos de análisis descriptivo multivariante, como el análisis clúster, el análisis en componentes principales y los *biplots*, como se muestra en los capítulos 6 y 7:

$$clr_j = \log\left(\frac{x_j}{\sqrt[D]{x_1 x_2 \ldots x_D}}\right) \quad \text{con } j = 1,2,\ldots,D \tag{25}$$

La media geométrica de n valores cualesquiera se define así como la raíz enésima de su producto:

$$g(x) = \sqrt[n]{x_1 x_2 \cdots x_n} \tag{26}$$

En nuestro caso, con $D = 6$ tenemos seis log-ratios centradas:

$$\begin{aligned}
clr_1 &= \log\left(\frac{x_1}{\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}}\right) \\
clr_2 &= \log\left(\frac{x_2}{\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}}\right) \\
clr_3 &= \log\left(\frac{x_3}{\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}}\right) \\
clr_4 &= \log\left(\frac{x_4}{\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}}\right) \\
clr_5 &= \log\left(\frac{x_5}{\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}}\right) \\
clr_6 &= \log\left(\frac{x_6}{\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}}\right)
\end{aligned} \tag{27}$$

Todas las log-ratios por pares posibles están contenidas en las log-ratios centradas. Nótese, por ejemplo, cómo se puede obtener y_1 a partir de clr_5 y clr_2:

$$clr_5 - clr_2 = \log\left(\frac{x_5}{\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}}\right) - \log\left(\frac{x_2}{\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}}\right) =$$

$$\log(x_5) - \log\left(\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}\right) - \left(\log(x_2) - \log\left(\sqrt[6]{x_1 x_2 x_3 x_4 x_5 x_6}\right)\right) =$$

$$\log(x_5) - \log(x_2) = \log\left(\frac{x_5}{x_2}\right) = y_1 \quad (28)$$

Precisamente, otro problema de las ratios clásicas que la metodología CoDa resuelve es que se suele tomar un excesivo número de ratios clásicas y son casi siempre redundantes (Barnes, 1987; Chen y Shimerda, 1981; Pohlman y Hollinger, 1981). Nunca se necesitarán más de D ratios para recoger la información relativa de D valores de los estados financieros. En el apartado 2.1, definimos hasta doce ratios clásicas a partir de las mismas seis partes, y aun con algunas variantes adicionales.

Como mostraremos en los capítulos 6 y 7, incluso si se utilizan log-ratios centradas como datos de entrada de los análisis estadísticos, la interpretación de los resultados puede volver a las log-ratios por pares más fácilmente interpretables (capítulo 6) y, en algunos casos, incluso a las ratios financieras clásicas de las ecuaciones (1) a (12), tal como se explica en el capítulo 7. Véase Saus-Sala et al. (2021; 2023; 2024).

Por contraposición a las ratios financieras clásicas, es legítimo llamar conjuntamente a las log-ratios por pares y las log-ratios centradas *ratios financieras composicionales*.

4.4. Reemplazamiento de ceros

Una limitación comúnmente mencionada de CoDa es que las partes contables de interés no pueden contener valores cero para que se calculen las log-ratios (Martín-Fernández et al., 2011). Sin embargo, un hecho que a menudo se pasa por alto es que ocurre exactamente lo mismo con las ratios financieras clásicas: un valor contable cero no es relativo a nada y, por lo tanto, la ratio no es una operación válida según la teoría de las escalas de medición (Stevens, 1946). La ratio se utiliza para medir cuántas veces una magnitud contiene otra, y esto no tiene respuesta cuando una de las magnitudes es cero. Si el valor cero es el denominador, la ratio clásica ni siquiera se puede calcular.

A diferencia del caso del análisis de ratios financieras clásicas, la metodología CoDa incluye una caja de herramientas avanzada para la *imputación de ceros* (también conocida como *reemplazamiento de ceros*) antes del cálculo de las log-ratios y bajo los supuestos estadísticos más comunes (Martín-Fernández et al., 2012). Esto proporciona a la metodología CoDa una ventaja de salida en comparación con el análisis de ratios financieras clásicas en presencia de ceros

y, en última instancia, hace posible el análisis de los estados financieros incluso cuando algunos valores contables de interés son iguales a cero. Resumiéndolo en pocas palabras, los ceros se reemplazan por valores pequeños que tengan sentido y que posean ciertas propiedades estadísticas.

La necesidad de algún tipo de tratamiento de los ceros se reconoció desde el principio del desarrollo de la metodología CoDa (Aitchison, 1982). Véanse Mariadassou y Coenders (2025) y Coenders et al. (2023b) para reseñas recientes sobre el tema. Desde los primeros métodos simples (Aitchison, 1986; Fry et al., 2000; Martín-Fernández et al., 2003), los desarrollos pasaron a métodos avanzados, con Palarea-Albaladejo y Martín-Fernández (2008; 2015) y Martín-Fernández et al. (2011; 2012) como referencias clave.

En la literatura de análisis de los estados financieros composicionales, el método de imputación más popular es, con mucho, el método esperanza-maximización (EM) por log-ratios (Palarea-Albaladejo y Martín-Fernández, 2008). Este método es similar al método EM estándar para imputar los *datos faltantes*, ya que el valor imputado se predice a partir de los valores disponibles con un modelo estadístico. Los datos faltantes constituyen simplemente valores ausentes o desconocidos para una variable y una empresa determinadas. Nada indica *a priori* que un valor faltante vaya a ser grande o pequeño. Sin embargo, en el caso del reemplazamiento de ceros en composiciones, hay una diferencia importante: el método EM añade la restricción de que los valores imputados deben ser pequeños. En particular, se constriñen a estar por debajo del valor mínimo no cero observado de cada parte o por debajo de cualquier otro umbral especificado por el usuario o usuaria, llamado *límite de detección*.

Los límites de detección definidos por el usuario o usuaria son especialmente útiles en el siguiente caso. Si el valor mínimo distinto de cero corresponde a una empresa con un valor muy bajo, el reemplazamiento por debajo de este límite podría hacer que los ceros reemplazados se conviertan en observaciones atípicas. En este caso, recomendamos establecer el límite de detección un poco más alto. Según nuestra experiencia, los límites de detección en torno a la media del 5 % de los valores más bajos distintos de cero tienden a funcionar bien.

Las metodologías de imputación de ceros requieren que el número de ceros sea pequeño, idealmente inferior al 20 %, para cualquiera de las partes (Palarea-Albaladejo y Martín-Fernández, 2008). Por lo tanto, antes de la imputación, se deben examinar los porcentajes de ceros. Esto puede impedir la división de los activos y pasivos en cuentas muy detalladas, como los inmuebles, las marcas comerciales, los inventarios, los clientes, la tesorería, los valores, los proveedores, los préstamos a corto plazo, los bonos, los préstamos a largo plazo, etc., algunas de las cuales son cero para una gran parte de las empresas, por lo menos si la muestra de datos incluye pymes.

En otras palabras, la elección del número y el detalle de los D valores contables debe estar sujeta a la presencia de ceros. Si algunas partes contienen más del 20 % de ceros, es posible que el usuario o usuaria desee sumarlas con otras partes conceptualmente similares con menos ceros y así reducir D. Por ejemplo, si los préstamos a corto plazo tienen un 30 % de ceros y las cuentas a pagar a proveedores tienen un 5 % de ceros, la suma de ambos en una categoría de pasivo corriente dará como resultado como máximo un 5 % de ceros (o incluso menos, si los ceros no coinciden en las mismas empresas). Por lo tanto, antes de decidir

qué valores contables hay que agregar, es útil examinar no solo sus porcentajes de ceros y su similitud conceptual, sino también la coocurrencia de ceros mediante el llamado *gráfico de patrones de ceros*. Dicho gráfico muestra cuántas empresas presentan cada una de las combinaciones posibles de ceros. En el ejemplo anterior, si ninguna empresa tiene simultáneamente ceros tanto para los préstamos a corto plazo como para los proveedores, la parte agregada estará completamente libre de ceros. Estas agregaciones se denominan *amalgamas* en la literatura de CoDa, concepto que ya había aparecido en el apartado 4.1.

En relación con el problema de los ceros, las empresas inactivas, como se revela por tener valores cero en los ingresos de explotación o en los activos totales, deben eliminarse completamente del archivo de datos. Si las empresas están inactivas, simplemente no pertenecen a la población del estudio, y no tiene sentido reemplazar la información contable que les falta con algún tipo de pequeño valor que tenga sentido (imagínese cómo se verían el margen o la rotación con los ingresos de explotación iguales a cero reemplazados por valores muy pequeños). Recomendamos a los investigadores que siempre descarten estas empresas, tanto desde el punto de vista conceptual como práctico. Esta situación se denomina indistintamente *ceros absolutos*, *ceros esenciales*, *ceros estructurales* o *ceros verdaderos* en la literatura de CoDa, y el consenso es que no son aptos para su reemplazamiento (Martín-Fernández et al., 2011). Cuando hay ceros absolutos, se trata de una simple imposibilidad del funcionamiento correcto de una empresa. Una empresa no puede funcionar con poco o nada de ingresos de explotación. En cambio, sí puede funcionar con poco o nada de pasivo no corriente (financiándose totalmente con deudas a corto plazo) o con poco o nada de activo no corriente (recurriendo a fórmulas de alquiler). Los ceros de estas dos partes no serían absolutos.

4.5. Manos a la obra con CoDaPack. Preparamos los datos para el análisis

A partir de este capítulo, cada uno tiene un apartado titulado «Manos a la obra con CoDaPack» donde el contenido teórico se ilustra con datos reales. Se da respuesta a preguntas de investigación concretas con ayuda de un programa informático especializado en datos composicionales llamado CoDaPack. Animamos a los lectores a reproducir por sí mismos cada uno de los resultados con los archivos que se facilitan en formato Excel.

La muestra que usamos para la mayoría de los ejemplos de este libro proviene del estudio del sector vitivinícola de Arimany-Serrat et al. (2023). Se trata de un sector clave en España, que consta de setenta denominaciones de origen y 42 indicaciones geográficas protegidas, y representa el 13 % de la superficie de viñedo a escala mundial (Vizcaíno et al., 2020). Antes de la pandemia de COVID-19, el sector generaba el 2,2 % del valor añadido y el 2,4 % del empleo en España, y este país era el primer exportador mundial de vino en volumen y el tercero en valor (Vizcaíno et al., 2020). El sector es altamente diverso, incluye empresas solo elaboradoras y también cultivadoras, empresas exportadoras y orientadas al mercado nacional, pymes y grandes empresas, y empresas que disponen de sus propias marcas comerciales y otras que producen para otras marcas (Arimany-Serrat et al., 2023;

Coenders, 2025), con lo que las variables que explican la salud financiera del sector son muy diversas (Castillo Valero y García Cortijo, 2013).

Se parte de los estados financieros de una muestra de 370 empresas obtenida de la base de datos SABI, extraída con los siguientes filtros: empresas mercantiles (sociedades anónimas y sociedades limitadas) con diez o más empleados, en territorio español, activas durante el período 2019-2020 y del sector vitivinícola (CNAE 1102 «Elaboración de vinos»). Se eliminaron empresas sin ingresos de explotación, al considerarse ceros absolutos. En Arimany-Serrat et al. (2023) se usaban los datos de 2019 y 2020; en los ejemplos de este libro, los de 2019. Los resultados no corresponden exactamente con los del artículo por un tratamiento distinto de los ceros y las observaciones atípicas y una codificación más agregada de las comunidades autónomas. El archivo Excel se llama *vinicolas.xls* y se halla disponible en ResearchGate (https://doi.org/10.13140/RG.2.2.22798.57925).

Las variables disponibles son las siguientes. Un primer grupo de variables son características no financieras:

- *Id*: código de identificación de cada empresa, de 1 a 370.
- *CA*: comunidad autónoma agrupada (variable categórica codificada textualmente como *AND*, *PV*, *CAT*, etc.).
  - *AND*: variable binaria que indica (codificadas con el código numérico «1») las empresas sitas en Andalucía (5,7 %); el «0» es para las otras comunidades.
  - *PV*: variable binaria que indica las empresas sitas en el País Vasco (7,3 %).
  - *CAT*: variable binaria que indica las empresas sitas en Cataluña (16,5 %).
  - *CL*: variable binaria que indica las empresas sitas en Castilla y León (20,3 %).
  - *CM*: variable binaria que indica las empresas sitas en Castilla-La Mancha (10,3 %).
  - *GAL*: variable binaria que indica las empresas sitas en Galicia (7,6 %).
  - *MUR*: variable binaria que indica las empresas sitas en Murcia (4,1 %).
  - *NAV*: variable binaria que indica las empresas sitas en Navarra (4,3 %).
  - *RIO*: variable binaria que indica las empresas sitas en La Rioja (7,3 %).
  - *VAL*: variable binaria que indica las empresas sitas en la Comunidad Valenciana (4,1 %).
  - *OTRAS*: variable binaria que indica las empresas sitas en otras comunidades autónomas con poca implantación del sector vitivinícola (12,7 %).
- *Edad*: número de años que la empresa ha estado activa (variable numérica).
- *SA*: variable binaria que indica (codificadas como «1») las sociedades anónimas (33,0 %); el «0» es para las sociedades limitadas.
- *Exporta*: variable binaria que indica las empresas que exportan al menos una parte de su producción (28,6 %); el «0» es para las que no lo hacen.
- *Subvenciones*: variable binaria que indica las empresas que reciben alguna subvención (70,8 %); el «0» es para las que no la reciben.
- *log_empleados*: logaritmo del número de empleados (variable numérica).
- *Genero*: distribución de género de los empleados (variable categórica codificada textualmente como *Alto*, *Bajo*, *ND*).
  - *Genero_alto*: variable binaria que indica las empresas con porcentaje de mujeres entre el total de empleados igual o superior a la mediana del sector (27 %).
  - *Genero_bajo*: variable binaria que indica las empresas con porcentaje de mujeres inferior a la mediana del sector (27 %).

  - *Genero_no_divulgado*: variable binaria que indica las empresas que no divulgan el porcentaje de mujeres (45,9 %).

Un segundo grupo de variables son los $D = 6$ valores contables x_1 a x_6, por supuesto, numéricos:

- *x1_ANC*: x_1 activo no corriente.
- *x2_AC*: x_2 activo corriente.
- *x3_PNC*: x_3 pasivo no corriente.
- *x4_PC*: x_4 pasivo corriente.
- *x5_IE*: x_5 ingresos de explotación.
- *x6_GE*: x_6 gastos de explotación.

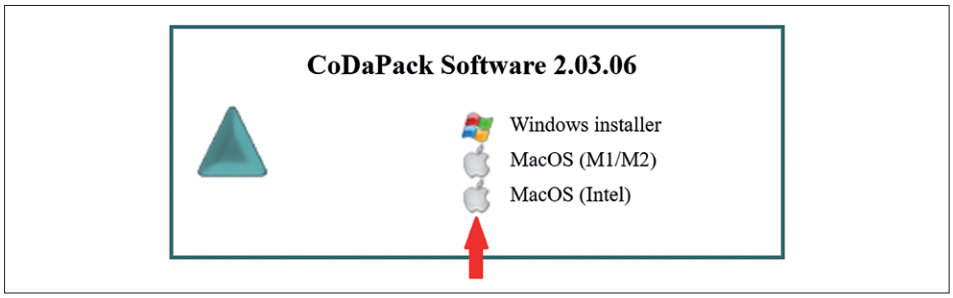
Un tercer grupo de variables (por supuesto numéricas) son las ratios financieras clásicas, que no son necesarias para el análisis composicional, pero sirven para ilustrar sus problemas con la asimetría, las observaciones atípicas, las medias aritméticas, la permutación de numerador y denominador, etc. al final de este mismo capítulo y también en el capítulo 8. Existen cinco empresas con patrimonio neto $x_1 + x_2 - x_3 - x_4$ negativo para las cuales el cálculo del apalancamiento y el ROE no tienen sentido, con lo que estas dos ratios se han omitido en el archivo de datos. Quedan las siguientes:

- *Rotación* (rotación del activo total): ingresos de explotación sobre activo total (ecuación (1)).
- *Rotacion_AC* (rotación del activo corriente): ingresos de explotación sobre activo corriente (ecuación (2)).
- *Margen*: beneficio de explotación sobre ingresos de explotación (ecuación (3)).
- *ROA* (rentabilidad económica): beneficio de explotación sobre activo total (ecuación (5)).
- *Endeudamiento*: pasivo total sobre activo total (ecuación (7)).
- *Endeudamiento_CP* (endeudamiento a corto plazo): pasivo corriente sobre activo total (ecuación (8)).
- *Solvencia_LP* (solvencia a largo plazo): activo total sobre pasivo total, o permutación del numerador y el denominador del endeudamiento (ecuación (9)).
- *Solvencia_CP* (solvencia a corto plazo): activo corriente sobre pasivo corriente (ecuación (10)).
- *Inmovilizacion* (inmovilización del activo): activo no corriente sobre activo corriente (ecuación (11)).
- *Maduracion_deuda*: pasivo no corriente sobre pasivo corriente (ecuación (12)).

Nótese que las variables que indican características no numéricas de las empresas (es decir, variables cualitativas) admiten una doble codificación. La codificación original suele ser como variables categóricas en las que cada valor posible (categoría) recibe un código distinto que puede ser alfanumérico. Otra es como variable binaria que se refiere a una sola de las categorías y codificada numéricamente como «0» (ausencia de la categoría) y «1» (presencia de la categoría).

En todos los ejemplos usamos el programa CoDaPack2.03.06. Se trata de un programa informático especializado en análisis CoDa de libre distribución (disponible en https://ima.udg.edu/codapack) y de funcionamiento por menús (Comas-Cufí y Thió-Henestrosa, 2011; Thió-Henestrosa y Martín-Fernández, 2005). Para una introducción véase Ferrer-Rosell et al. (2022) y Coenders y Arimany-Serrat (2023). Para instalar el programa,

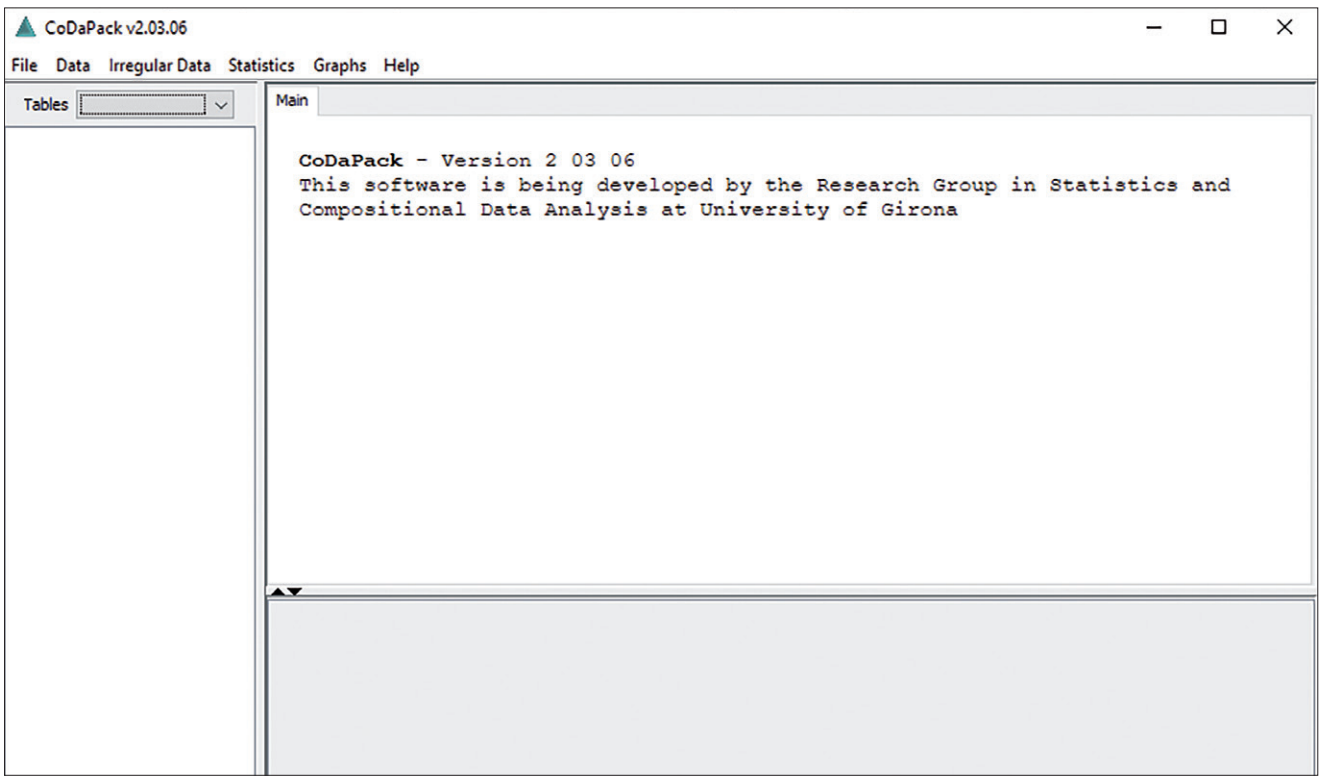
accedemos a https://ima.udg.edu/codapack y simplemente clicamos sobre el icono de la versión Windows o Mac que necesitemos.



Una vez descargado el archivo de instalación, lo ejecutamos y seguimos sus instrucciones.

Después de instalarlo, entramos en el programa CoDaPack. Su interfaz de usuario se compone de los siguientes elementos:

- Una barra de menús.
- Un desplegable llamado *Tables* que permitirá pasar de una tabla de datos a otra cuando tengamos más de una.
- Debajo de este aparecerá la lista de variables la tabla activa en cada momento.
- La ventana *Main* contendrá los resultados numéricos cuando se vayan obteniendo. Solo los numéricos, pues los resultados gráficos aparecen en ventanas emergentes.
- Debajo de esta, aparecerá una visualización de la tabla de datos activa en cada momento.



Para ser utilizables con CoDaPack, los archivos de Excel deben cumplir una serie de requisitos, con lo que los archivos descargados de SABI deberán editarse para garantizar que los cumplan:

- Solo deben contener una hoja.
- Las empresas se disponen por filas y las variables, por columnas.
- Los nombres de las variables aparecen en la primera fila y los datos, a partir de la segunda fila.
- Los datos pueden ser texto o números, no fórmulas.
- Los nombres de las variables solo pueden contener letras del alfabeto inglés, números, puntos y guiones bajos «_», y no pueden incluir espacios ni tildes.
- Los ceros en los datos contables deben introducirse como tales ceros. Hay que tener en cuenta que en SABI suelen estar introducidos como «n. d.», lo cual hay que modificar.
- Los datos faltantes en las variables no contables (variables desconocidas en alguna de las empresas) deben introducirse como «NA».

El archivo *vinicolas.xls* ya cumple todas estas indicaciones.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	CA	AND	PV	CAT	CL	CM	GAL	MUR	NAV	RIO	VAL	OTRAS	Edad	SA	Exporta	Subvenciones	log_empleados	Genero	Genero_alto	Genero_bajo	Genero_no_divulgado
2	PV	0	1	0	0	0	0	0	0	0	0	0	45	1	0	1	2,77	Bajo	0	1	0
3	PV	0	1	0	0	0	0	0	0	0	0	0	36	1	1	1	2,89	Bajo	0	1	0
4	PV	0	1	0	0	0	0	0	0	0	0	0	34	1	0	1	3,18	ND	0	0	1
5	PV	0	1	0	0	0	0	0	0	0	0	0	23	1	1	1	3,97	Alto	1	0	0
6	PV	0	1	0	0	0	0	0	0	0	0	0	22	1	1	1	3,04	ND	0	0	1
7	CM	0	0	0	0	1	0	0	0	0	0	0	34	1	1	1	4,52	Alto	1	0	0
8	CM	0	0	0	0	1	0	0	0	0	0	0	31	1	0	1	2,56	Bajo	0	1	0
9	CM	0	0	0	0	1	0	0	0	0	0	0	25	1	0	1	2,48	Bajo	0	1	0
10	CM	0	0	0	0	1	0	0	0	0	0	0	24	1	1	1	3,30	ND	0	0	1
11	CM	0	0	0	0	1	0	0	0	0	0	0	19	1	0	1	2,83	ND	0	0	1

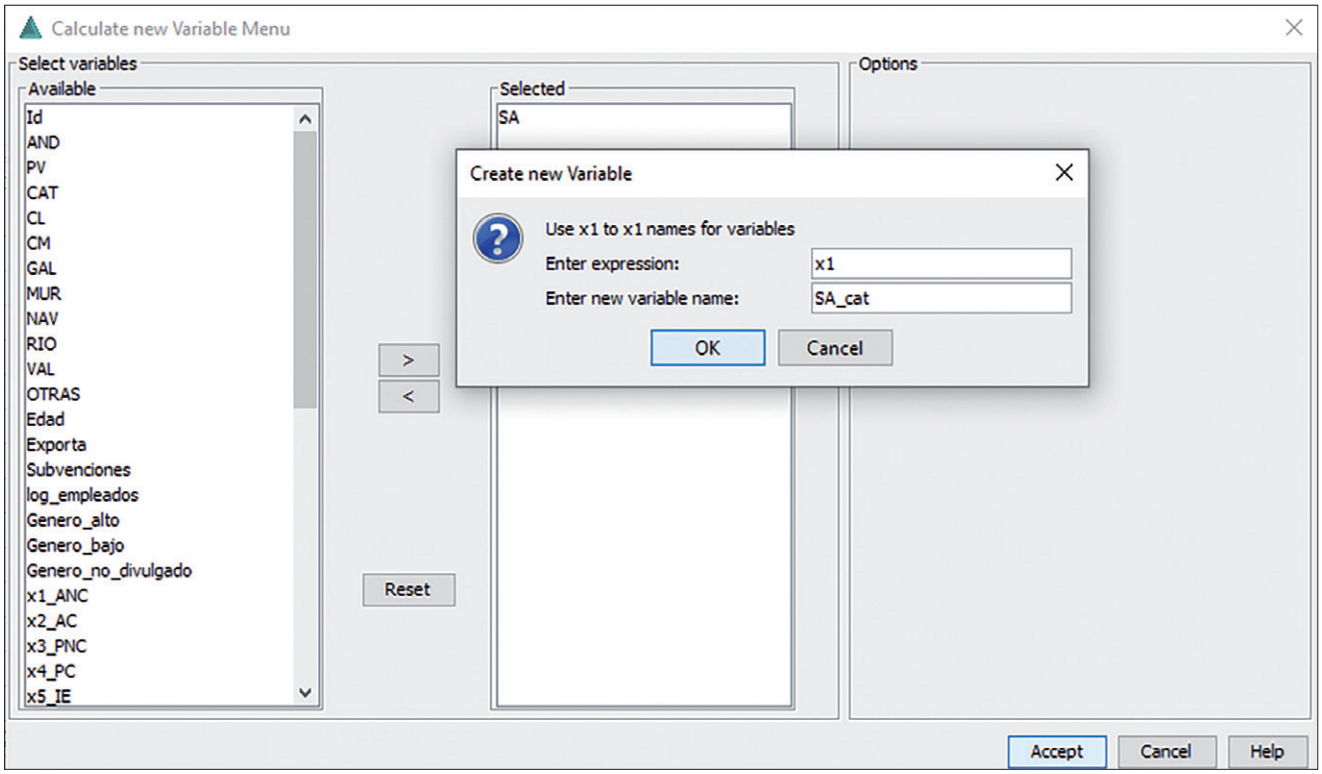
Para leer el archivo de datos Excel (por ahora, CoDaPack es compatible solo con archivos en formato *.xls*, no con archivos *.xlsx*), seleccionamos el menú *File > Import > Import XLS Data* con las opciones por defecto. Las variables *CA* y *Genero* están resaltadas en color naranja, para indicar su carácter categórico. Como contienen texto, solo pueden ser categóricas. Las variables cualitativas codificadas como binarias («1» y «0») pueden definirse en CoDaPack como categóricas o como numéricas según convenga.

Las variables *Genero* y *CA* ya están disponibles en su doble versión categórica (aparecen en color naranja en la base de datos) y numérica binaria (en blanco). Es interesante tener esta doble versión de cada variable, pues cada una se adapta a distintos análisis. Queremos, por lo tanto, disponer de la versión categórica para las variables cualitativas del archivo que ahora son solo numéricas binarias: *SA*, *Exporta* y *Subvenciones*.

	CA	AND	PV	CAT	CL	CM	GAL	MUR	NAV	RIO	VAL	OTRAS	Edad	SA	Exporta	Subvenciones	log_empleados	Genero	Genero_alto	Genero_bajo	Genero_no_divulgado
1	PV	0	1.00...	0	0	0	0	0	0	0	0	0	45.000000	1.00...	0	1.000000	2.772589	Bajo	0	1.000000	0
2	PV	0	1.00...	0	0	0	0	0	0	0	0	0	36.000000	1.00...	1.000...	1.000000	2.890372	Bajo	0	1.000000	0
3	PV	0	1.00...	0	0	0	0	0	0	0	0	0	34.000000	1.00...	0	1.000000	3.178054	ND	0	0	1.000000
4	PV	0	1.00...	0	0	0	0	0	0	0	0	0	23.000000	1.00...	1.000...	1.000000	3.970292	Alto	1.000000	0	0
5	PV	0	1.00...	0	0	0	0	0	0	0	0	0	22.000000	1.00...	1.000...	1.000000	3.044522	ND	0	0	1.000000
6	CM	0	0	0	0	1.00...	0	0	0	0	0	0	34.000000	1.00...	1.000...	1.000000	4.521789	Alto	1.000000	0	0
7	CM	0	0	0	0	1.00...	0	0	0	0	0	0	31.000000	1.00...	0	1.000000	2.564949	Bajo	0	1.000000	0
8	CM	0	0	0	0	1.00...	0	0	0	0	0	0	25.000000	1.00...	0	1.000000	2.484907	Bajo	0	1.000000	0
9	CM	0	0	0	0	1.00...	0	0	0	0	0	0	24.000000	1.00...	1.000...	1.000000	3.295837	ND	0	0	1.000000
10	CM	0	0	0	0	1.00...	0	0	0	0	0	0	19.000000	1.00...	0	1.000000	2.833213	ND	0	0	1.000000

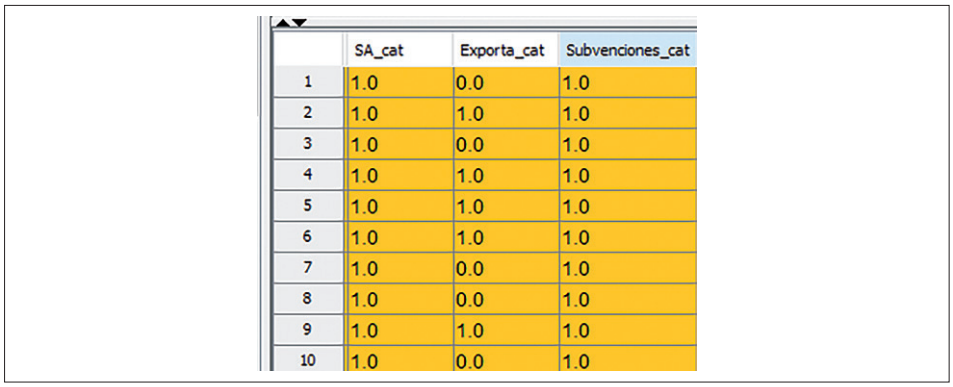
Empezamos por realizar una copia de cada una de esas variables. Entramos en el menú *Data* > *Manipulate* > *Calculate New Variable*. Este menú sirve para crear nuevas variables a partir de variables existentes. Internamente, CoDaPack llama las variables existentes que vamos a usar *x1*, *x2*, etc. Por ejemplo, si quisiéramos sumar dos variables, la operación a realizar sería *x1* + *x2*. Obsérvese que la letra *x* está en minúscula siempre.

Introducimos la variable numérica binaria *SA* sobre la cual queremos operar en el cuadro *Selected*. En este caso se trata de una única variable. Tras clicar *Accept*, aparece un nuevo cuadro de diálogo donde hay que entrar la operación *x1* en la casilla superior (pues solo queremos hacer una copia de la variable, sin modificarla) y el nombre que queramos dar a la nueva variable (solo letras del alfabeto inglés, números, puntos y guiones bajos «_», sin espacios ni tildes) en la casilla inferior. Clicamos *OK*.



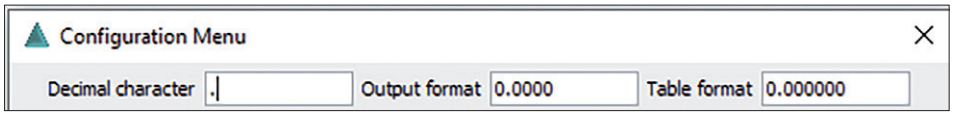
Una vez creada la copia de la variable binaria numérica *SA* llamada *SA_cat*, para convertirla en categórica vamos al menú *Data* > *Manipulate* > *Numeric to Categoric* e introducimos *SA_cat* en *Selected*.

Repetimos la doble operación para *Exporta* y *Subvenciones*. Tras hacerlo, obtenemos la versión categórica de las tres variables al final de la tabla de datos:



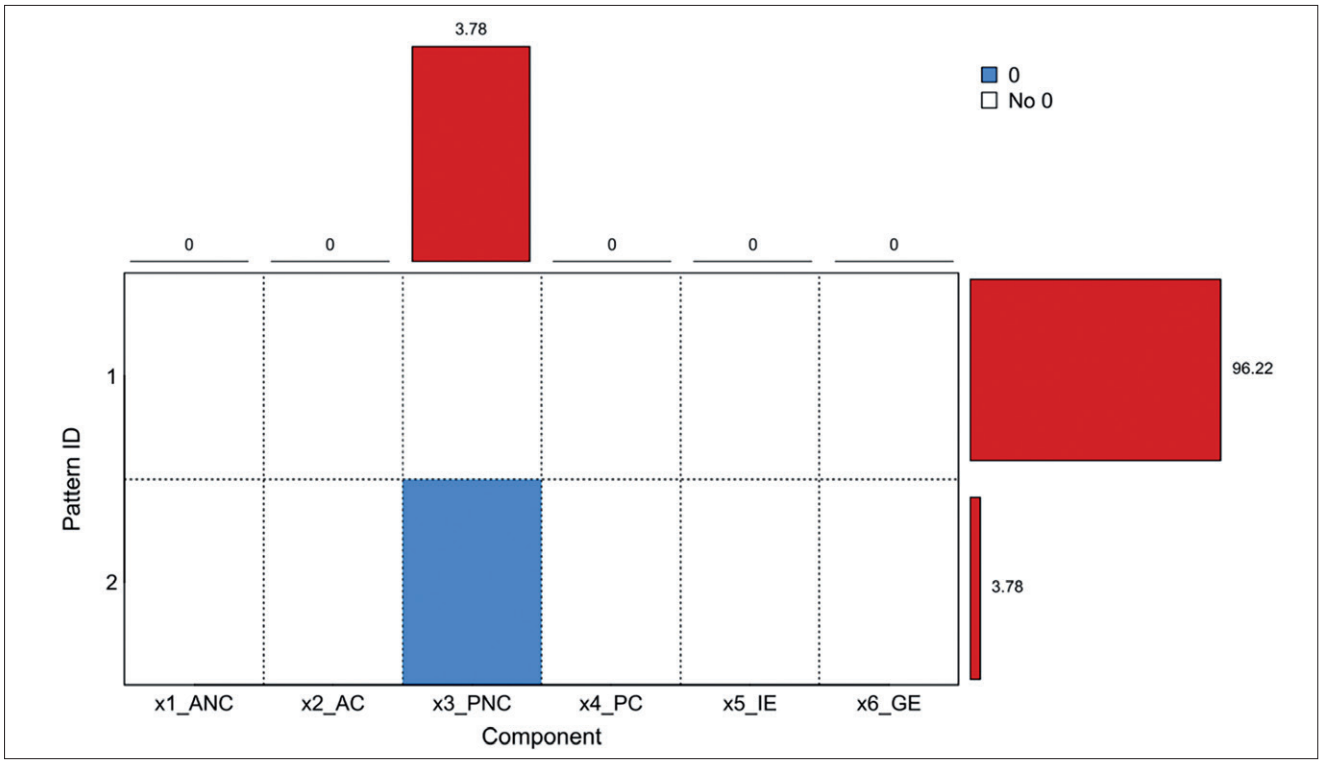
	SA_cat	Exporta_cat	Subvenciones_cat
1	1.0	0.0	1.0
2	1.0	1.0	1.0
3	1.0	0.0	1.0
4	1.0	1.0	1.0
5	1.0	1.0	1.0
6	1.0	1.0	1.0
7	1.0	0.0	1.0
8	1.0	0.0	1.0
9	1.0	1.0	1.0
10	1.0	0.0	1.0

También es posible transformar una variable categórica codificada numéricamente a numérica con el menú *Data* > *Manipulate* > *Categoric to Numeric*. Queda una última recomendación previa. En el menú *File* > *Configuration*, seleccionamos el número de decimales que queremos visualizar en los resultados y en la tabla de datos (cuatro y seis, respectivamente). La tabla de datos se refrescará con los decimales adicionales al clicar sobre cualquier casilla de esta tabla.



Empecemos los análisis propiamente dichos. En primer lugar, hay que mirar si los datos contables contienen ceros para alguno de los *D* valores seleccionados por medio del *gráfico de patrones de ceros*. El menú *Irregular Data* > *Zero Patterns* calcula los porcentajes de ceros por parte y globalmente, y los porcentajes de coocurrencia de ceros, después de introducir simultáneamente las partes *x1_ANC, x2_AC, x3_PNC, x4_PC, x5_IE* y *x6_GE* en el cuadro *Selected* con las opciones *Show percentages* y *Plot Pattern*. En el gráfico obtenido, las barras verticales sobre la tabla indican que la parte *PNC* tiene un 3,78 % de ceros (14 de las 370 observaciones) y el resto de las partes están completas. Las barras horizontales a la derecha de la tabla indican que el 96,22 % de las empresas están completas y hay un único patrón de ceros con ceros en *x3_PNC* (casilla azul). Un patrón de ceros es cada combinación posible de partes iguales a cero presentes en el archivo de datos. Si hubiera más de una parte con ceros, la tabla contendría más patrones de ceros. Por ejemplo, si dos partes tuvieran empresas con ceros, aparecerían una fila con los ceros solo en la primera parte, una fila con los ceros solo en la segunda parte y una fila con los ceros en ambas partes simultáneamente, reconocible por tener dos casillas azules.

Los gráficos de CoDaPack se generan en una ventana separada. Para conservar el gráfico, recomendamos maximizar dicha ventana y realizar una captura de pantalla. Así se han hecho todos los gráficos de este libro.



Obtenemos la misma información en formato texto en la ventana *Main*. Los ceros están identificados en los patrones como «+». Primero salen los porcentajes por cada patrón de ceros y, finalmente, los porcentajes de ceros por cada parte.

```
Zero Patterns
Patterns ('+' means 0, '-' means No 0)

 Patt.ID x1_ANC x2_AC x3_PNC x4_PC x5_IE x6_GE  No.Unobs Patt.Perc
       1      -     -      -     -     -      -         0      96.2
       2      -     -      +     -     -      -         1       3.8

Percentage cells by component
x1_ANC  x2_AC x3_PNC  x4_PC  x5_IE x6_GE
   0.0    0.0    3.8    0.0    0.0    0.0
```

A continuación, fijamos el límite de detección en el valor más bajo observado distinto de cero. Entramos en el menú *Irregular Data* > *Set Detection Limit*, introducimos *x3_PNC* en el cuatro *Selected* y marcamos la opción *Column Minimum*. Tras hacerlo, vemos que en la tabla de datos aparece el límite de detección entre corchetes al lado de los ceros. Por ejemplo, para las observaciones 318 y 329 del archivo vemos que dicho límite es 5,25:

	x1_ANC	x2_AC	x3_PNC	x4_PC	x5_IE	x6_GE
318	4577.122...	2036.887...	0 [5.25]	3771.200...	1285.899...	1228.201...
319	15728.05...	3955.807...	70.543570	4031.829...	1653.441...	2012.100...
320	2051.132...	711.7868...	668.7270...	264.4771...	1520.182...	1322.217...
321	2861.906...	2196.725...	229.6850...	1477.971...	5276.452...	4214.031...
322	2520.659...	2934.786...	2936.752...	1012.017...	1618.468...	1593.391...
323	1167.250...	1601.264...	247.7226...	388.1080...	1833.334...	1465.320...
324	7156.916...	4364.927...	2918.417...	2409.002...	4051.487...	3936.391...
325	876.4228...	2076.862...	72.172330	946.7309...	1901.962...	1884.578...
326	5042.785...	3352.542...	1645.904...	772.6502...	3277.713...	3359.780...
327	1094.887...	1316.777...	778.4570...	1150.841...	2411.160...	2324.842...
328	162.8329...	713.2903...	8.853580	689.2541...	4248.100...	4130.252...
329	950.7724...	988.1659...	0 [5.25]	111.6936...	888.9982...	863.3540...

El porcentaje de ceros es suficientemente bajo para proceder a su reemplazamiento. El menú *Irregular Data* > *Log-Ratio EM Zero Replacement* reemplaza los ceros por el método EM por log-ratios de Palarea-Albaladejo y Martín-Fernández (2008). Introducimos simultáneamente las partes *x1_ANC*, *x2_AC*, *x3_PNC*, *x4_PC*, *x5_IE* y *x6_GE* en el cuadro *Selected*. Es importante la inclusión de todas las partes disponibles, no solo las que tienen ceros. No modificamos nada de las opciones por defecto que nos propone el programa. Seis nuevas variables sin ceros aparecen al final del archivo. Sus nombres son *z.x1_ANC*, *z.x2_AC*, *z.x3_PNC*, *z.x4_PC*, *z.x5_IE* y *z.x6_GE*. Para las mismas observaciones 318 y 329 del archivo vemos los valores reemplazados 3,1911 y 2,1476, por supuesto, inferiores a 5,25. Los valores distintos de cero no se ven modificados:

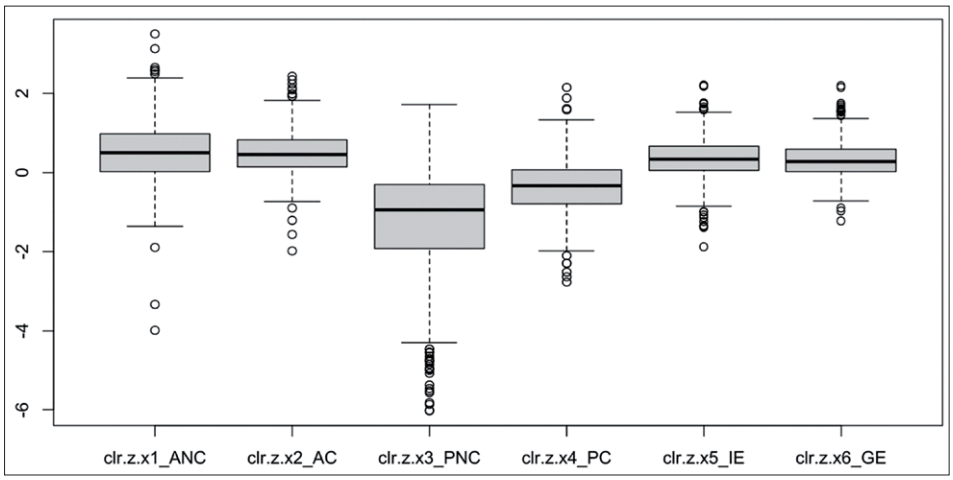
	z.x1_ANC	z.x2_AC	z.x3_PNC	z.x4_PC	z.x5_IE	z.x6_GE
318	4577.122...	2036.887...	3.191122	3771.200...	1285.899...	1228.201...
319	15728.05...	3955.807...	70.543570	4031.829...	1653.441...	2012.100...
320	2051.132...	711.7868...	668.7270...	264.4771...	1520.182...	1322.217...
321	2861.906...	2196.725...	229.6850...	1477.971...	5276.452...	4214.031...
322	2520.659...	2934.786...	2936.752...	1012.017...	1618.468...	1593.391...
323	1167.250...	1601.264...	247.7226...	388.1080...	1833.334...	1465.320...
324	7156.916...	4364.927...	2918.417...	2409.002...	4051.487...	3936.391...
325	876.4228...	2076.862...	72.172330	946.7309...	1901.962...	1884.578...
326	5042.785...	3352.542...	1645.904...	772.6502...	3277.713...	3359.780...
327	1094.887...	1316.777...	778.4570...	1150.841...	2411.160...	2324.842...
328	162.8329...	713.2903...	8.853580	689.2541...	4248.100...	4130.252...
329	950.7724...	988.1659...	2.147584	111.6936...	888.9982...	863.3540...

Para poder usar este procedimiento de CoDaPack, al menos una de las partes tiene que estar completa (sin ceros) y todas las empresas deben tener por lo menos dos partes distintas de cero. En análisis de los estados financieros esto suele ser fácil de conseguir. Ya hemos indicado que las empresas con ceros en los ingresos de explotación se suelen eliminar al considerarse ceros absolutos, o sea que por lo menos esta parte estará siempre completa.

Construimos ahora las log-ratios centradas o clr. El menú *Data* > *Transformation* > *CLR* almacena las log-ratios centradas como variables adicionales al final del archivo de datos. Hay que introducir simultáneamente z.*x1_ANC*, *z.x2_AC*,

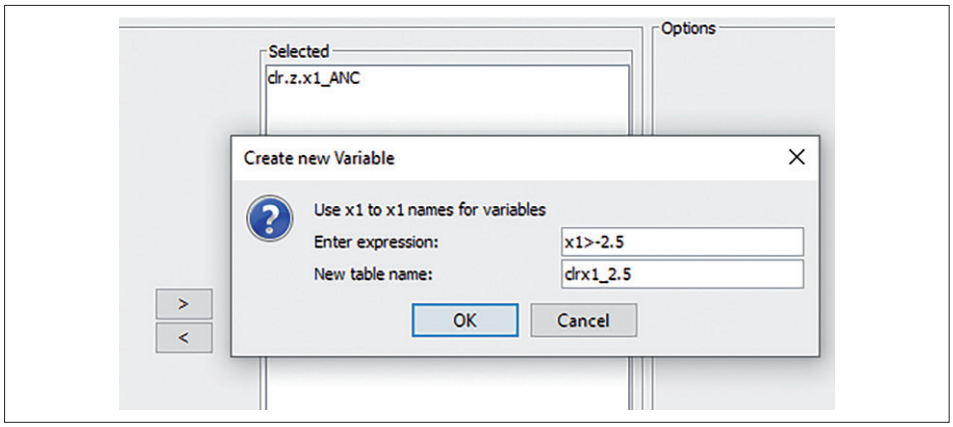
z.x3_PNC, *z.x4_PC*, *z.x5_IE* y *z.x6_GE* en el cuadro *Selected*, con la opción *Raw-CLR*. Las variables clr creadas son *clr.z.x1_ANC*, *clr.z.x2_AC*, *clr.z.x3_PNC*, *clr.z.x4_PC, clr.z.x5_IE* y *clr.z.x6_GE*.

El menú *Graphs* > *Boxplot* representa los llamados *gráficos de caja* o *diagramas de caja* después de introducir simultáneamente las seis log-ratios centradas *clr.z.x1_ANC*, *clr.z.x2_AC*, *clr.z.x3_PNC*, *clr.z.x4_PC*, *clr.z.x5_IE* y *clr.z.x6_GE* en el cuadro *Selected*. Más adelante presentaremos una explicación completa de este tipo de gráficos. Por el momento, nos basta con identificar las observaciones atípicas en el extremo superior o inferior, pero solo nos van a preocupar las extremas, que se encuentren separadas de manera destacada y visible del resto de las observaciones.

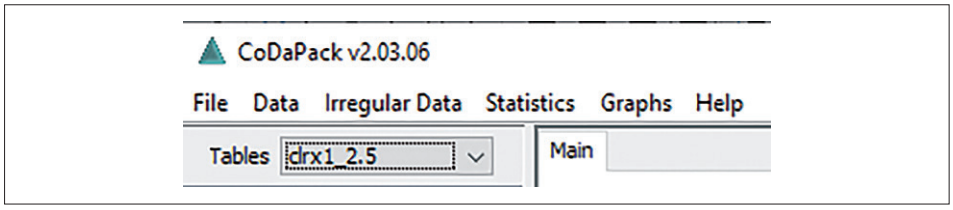


Se identifican dos observaciones atípicas extremas en la parte inferior de la log-ratio que tiene x_1 como numerador (*clr.z.x1_ANC*). En cambio, no aparecen observaciones atípicas extremas en la log-ratio que tiene x_3 como numerador (*clr.z.x3_PNC*), que es la que ha sido objeto de reemplazamiento de ceros. Por lo tanto, en este caso el reemplazamiento de ceros no ha provocado observaciones atípicas extremas, pero sí vamos a crear una nueva tabla de datos eliminando los dos valores inferiores a –2,5 en *clr.z.x1_ANC*.

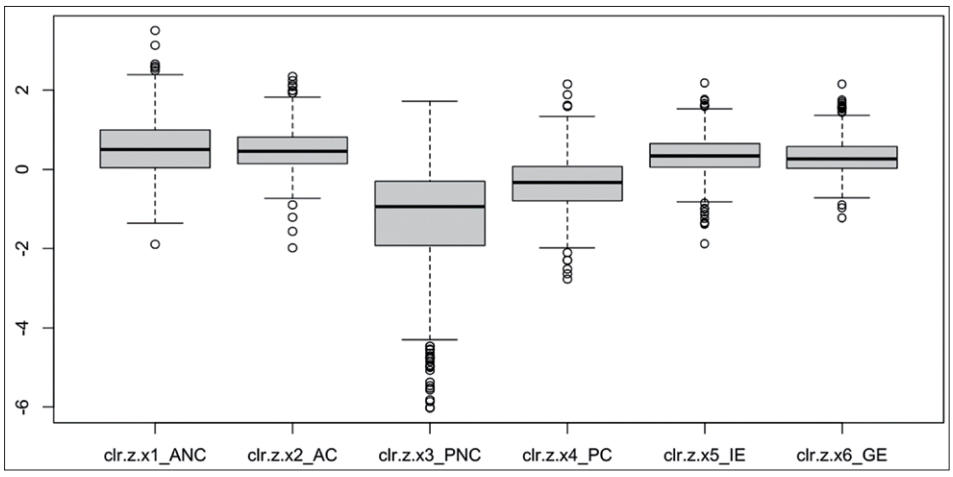
Para eliminar dichas observaciones atípicas, vamos al menú *Data* > *Filters* > *Advanced Filter*. Seleccionamos la variable *clr.z.x1_ANC* en el cuadro *Selected*. Esta variable queda ahora representada por *x1* y queremos seleccionar las observaciones no atípicas, que según el gráfico son las superiores a –2,5 (teniendo en cuenta que para CoDaPack el separador decimal es el punto y no la coma). Damos nombre a la nueva tabla (con las mismas restricciones que para los nombres de las variables). Decidimos llamarla *clrx1_2.5*.



En principio, la tabla *clrx1_2.5* será ahora la tabla activa, aunque podemos volver a la tabla anterior con totas las observaciones siempre que queramos, accediendo a la lista desplegable *Tables*. Constatamos que la nueva tabla tiene 368 observaciones.



Constatamos también que, si repetimos los gráficos de caja sobre la nueva tabla de datos, las dos observaciones atípicas extremas ya no están.



Construimos ahora las log-ratios por pares habituales. El menú *Data* > *Transformation* > *ALR* almacena las log-ratios por pares como variables adicionales al final del archivo de datos, después de introducir las dos partes involucradas en el cuadro *Selected,* la parte del numerador primero, la parte del denominador después, y con la opción *Raw-ALR*. Por ejemplo, para la ratio y_1 de

rotación del activo corriente (ecuación (19)), introducimos en este orden *z.x5_IE* y *z.x2_AC* y la variable creada se llama *alr.z.x5_IE_z.x2_AC*. Hacemos lo propio con las otras cuatro log-ratios. Por ejemplo, para la ratio y_2 (ecuación (20)) de margen seleccionamos en este orden *z.x5_IE* y *z.x6_GE*, y la variable creada se llama *alr.z.x5_IE_z.x6_GE*. Clicando dos veces sobre los nombres de las variables en el encabezamiento de la tabla de datos se pueden cambiar por otros más intuitivos a gusto del usuario o usuaria (siempre que contengan solo números, puntos, guiones bajos y letras del alfabeto inglés sin tildes). Hemos optado por acortarlos y dejar solo los nombres de las partes del numerador y denominador:

	x5_IE_x2_AC	x5_IE_x6_GE	x2_AC_x4_PC	x1_ANC_x2_AC	x3_PNC_x4_PC
1	-1.294372	0.054809	2.971900	0.505125	-3.871676
2	-0.585103	-0.089715	2.211848	-1.364384	1.061223
3	-0.622068	0.016909	0.685162	0.002902	-0.347644
4	-0.483557	0.196189	1.367095	-0.404833	-0.257101
5	-0.534650	0.019419	0.534568	0.464866	-0.281287
6	0.047397	0.088651	0.305631	-0.673737	-1.865234
7	0.125116	0.319065	1.711677	0.455939	-1.542482
8	-1.276212	-1.221000	1.130680	1.716863	2.239427
9	0.563421	0.023621	0.247827	-0.791556	-1.255305
10	-0.502014	0.036074	0.566276	0.774234	0.983305

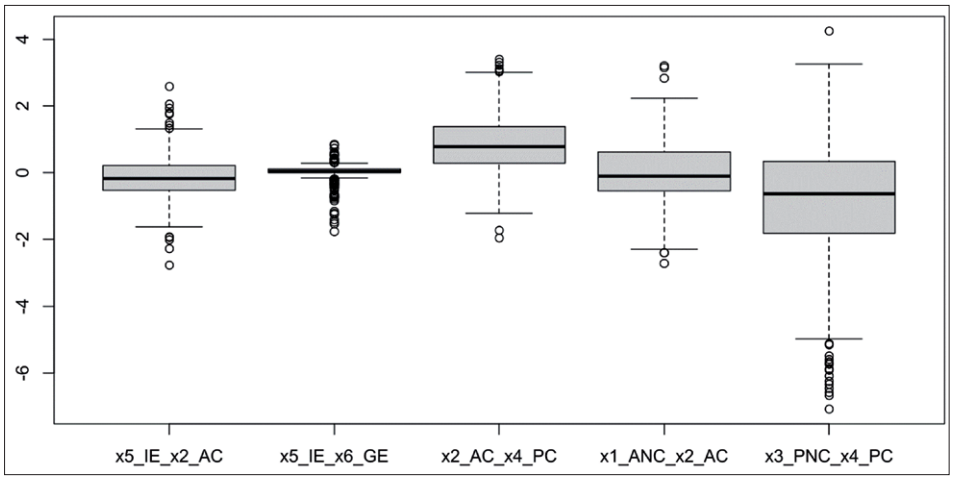
Reconocemos en *x2_AC_x4_PC* la log-ratio y_3 (ecuación (21)), que indica solvencia a corto plazo; en *x1_ANC_x2_AC* la log-ratio y_4 (ecuación (22)), que indica inmovilización del activo, y en *x3_PNC_x4_PC* la log-ratio y_5 (ecuación (23)), que indica maduración de la deuda.

Es un buen momento para guardar todas las variables que hemos ido creando y, a la vez, ambas tablas (con y sin observaciones atípicas) por medio del menú *File* > *Save as*. Ambas tablas se guardan de una sola vez en un único archivo en el formato nativo de CoDaPack con la terminación *.cdp*. Si antes de guardar el archivo queremos eliminar alguna variable que ya no vayamos a necesitar, entramos en el menú *Data* > *Delete Variables*, teniendo en cuenta que las variables seleccionadas solo se eliminarán de la tabla activa.

En sesiones futuras podrán abrirse todas las tablas ejecutando una sola vez el menú *File* > *Open Workspace*.

También es posible volver a exportar los datos a Excel con el menú *File* > *Export* > *Export Data to XLS*.

Con el menú *Graphs* > *Boxplot* podemos representar los diagramas de caja después de introducir simultáneamente las cinco log-ratios por pares en el cuadro *Selected*.



Recordemos que el gráfico de caja es una representación gráfica de los *cuartiles* de una variable llamados *Q1*, *Q2* y *Q3*, o a veces *percentil 25*, *percentil 50* y *percentil 75*. Q2 y el percentil 50 se conocen como *mediana*, término a no confundir con media, ni aritmética ni geométrica. Dichos cuartiles pueden obtenerse numéricamente en el menú *Statistics* > *Classical Statistics Summary*, si marcamos la opción *Percentile*.

Main

Clasical statistics summary:
NA's:
0

Sample size:
368

Statistics

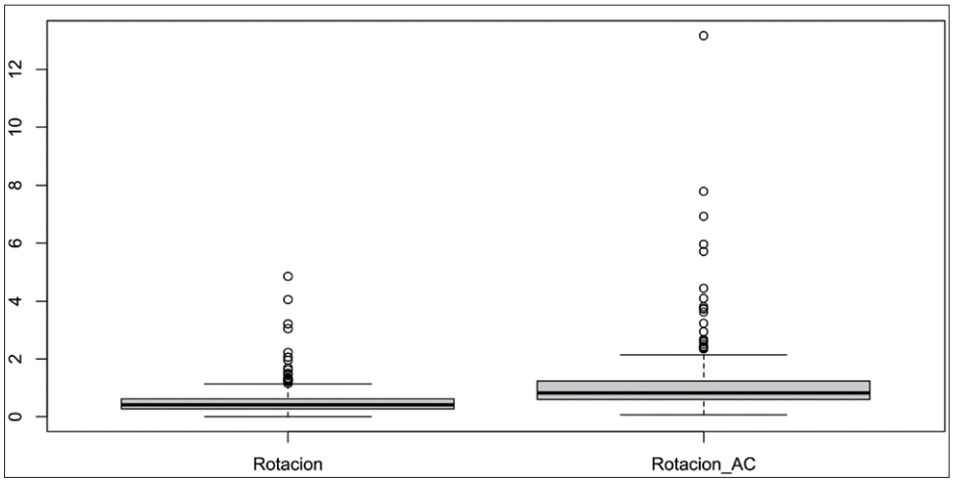
	0	25	50	75	100
x5_IE_x2_AC	-2.7746	-0.5214	-0.1786	0.2174	2.5774
x5_IE_x6_GE	-1.7703	0.0014	0.0327	0.1137	0.8470
x2_AC_x4_PC	-1.9660	0.2823	0.7842	1.3860	3.3915
x1_ANC_x2_AC	-2.7084	-0.5396	-0.1028	0.6331	3.2031
x3_PNC_x4_PC	-7.0748	-1.7995	-0.6343	0.3444	4.2328

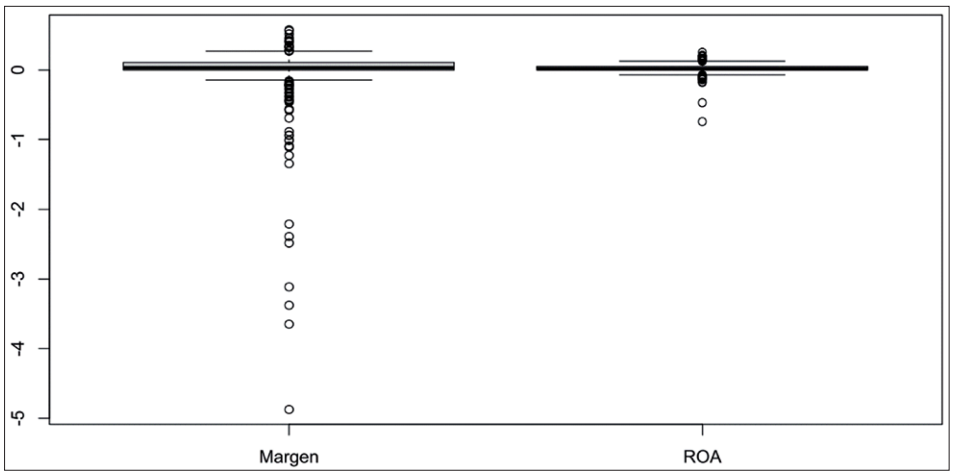
Por ejemplo, para la log-ratio por pares y_5 de maduración de la deuda (*x3_PNC_x4_PC*: pasivo no corriente sobre pasivo corriente), la caja abarca entre los límites Q1 = –1,80 y Q3 = 0,34, que contienen el 50 % central de los valores, es decir, de las empresas. El 25 % de los valores más bajos se halla entre el mínimo (–7,07) y el primer cuartil Q1 o percentil 25 (–1,80); el siguiente 25 % de los valores está entre –1,80 y la mediana (–0,63); el siguiente 25 %, entre la mediana y el tercer cuartil Q3 = 0,34, y el 25 % más alto, entre 0,34 y el máximo (4,23). Tenemos el 50 % de las observaciones por arriba y por debajo de la mediana, que es la línea horizontal en trazo grueso dentro de la caja en el gráfico. También tenemos el 50 % de los valores dentro de la caja, que está delimitada por los cuartiles primero (inferior) y tercero (superior). Las líneas por encima y por debajo de la caja abarcan el resto de las observaciones no atípicas, y las atípicas se representan aparte y de manera individual.

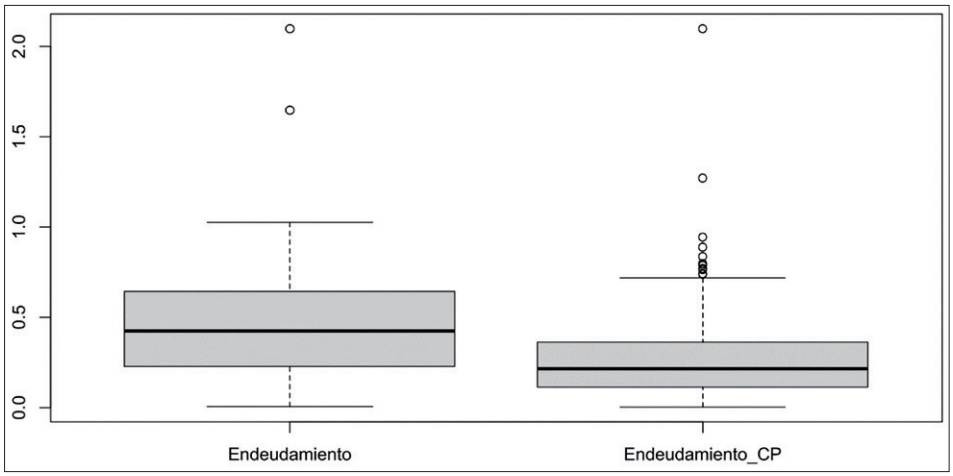
El gráfico de caja permite ver:

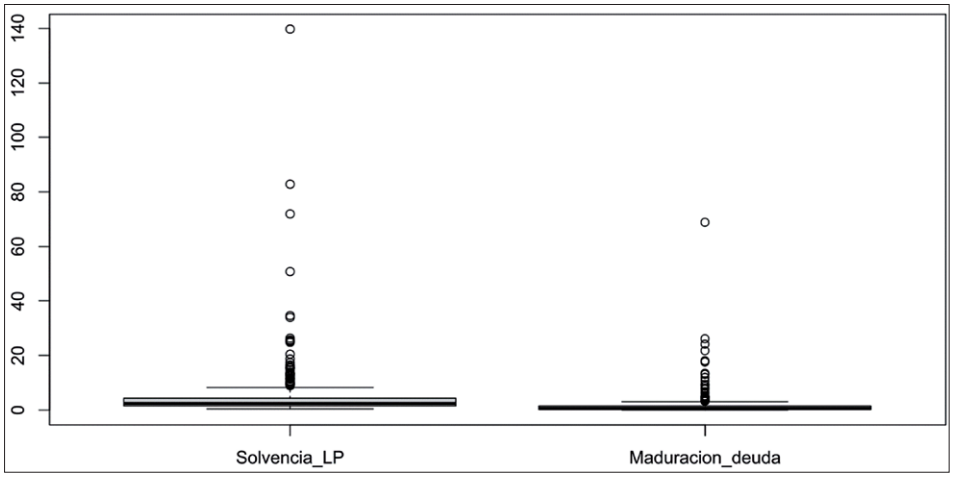
- El centro de una variable representado por la mediana. Las log-ratios con mediana positiva indican que, para una empresa típica del sector (la que tiene igual número de empresas por encima que por debajo), el numerador de la ratio supera el denominador. Valores negativos indican lo contrario.
- La dispersión, representada por la amplitud de la caja. Vemos, por ejemplo, que hay poca variación en la log-ratio por pares y_2 de margen (*x5_IE_X6_GE*) y mucha más en la log-ratio por pares y_5 de maduración de la deuda (*x3_PNC_x4_PC*).
- La asimetría, representada por la apariencia general del gráfico, lo que no es el caso con ninguna de estas log-ratios por pares.
- Las observaciones atípicas, que preocupan cuando son extremas, es decir, muy alejadas, lo que tampoco es el caso con estas log-ratios por pares.

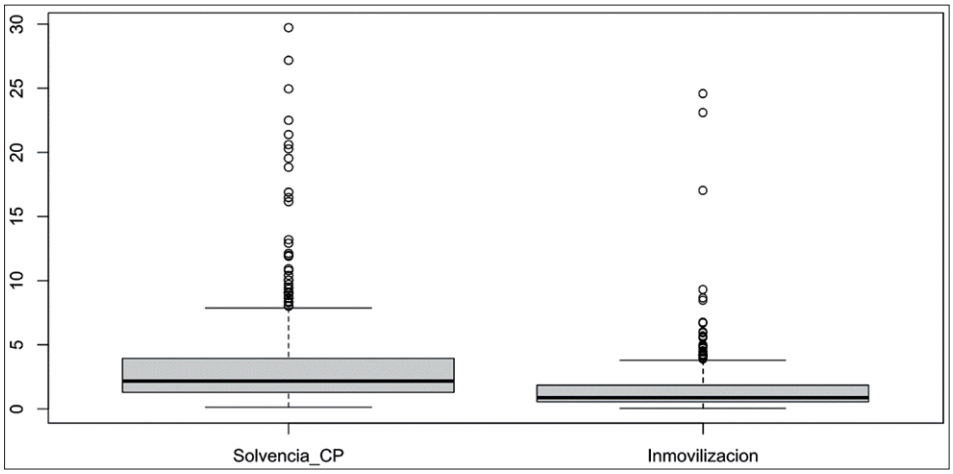
Con el menú *Graphs* > *Boxplot*, podemos representar los diagramas de caja de las ratios clásicas que ya están calculadas en el archivo de datos importado desde Excel. Para que los gráficos se vean bien, es mejor representar juntas las ratios que tomen valores mínimos y máximos parecidos.









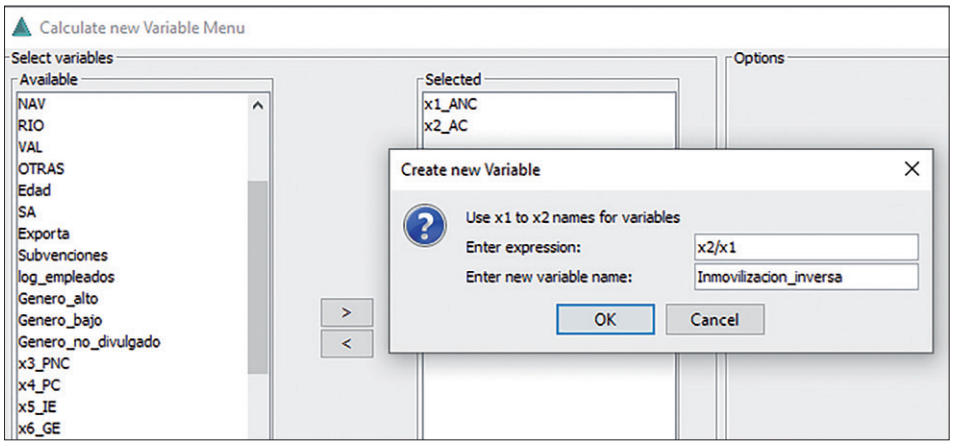


Observamos que el aspecto de los gráficos cambia completamente con las ratios clásicas. La rotación del activo corriente, el endeudamiento a corto plazo, la solvencia a largo plazo, la maduración de la deuda, la solvencia a corto plazo y la inmovilización del activo son marcadamente asimétricos. Todas las ratios clásicas tienen observaciones atípicas extremas, exceptuando con muy buena voluntad la rotación y el ROA. Recordemos que esto ocurre tras haber eliminado ya las (pocas) observaciones atípicas extremas sobre las log-ratios centradas. Es decir, algunas observaciones atípicas surgen atendiendo a empresas de características no habituales (las que eliminamos sobre las log-ratios centradas)

y otras surgen debido a los problemas estadísticos de las ratios clásicas aun con empresas de características no necesariamente extremas. Tomemos como ejemplo la ratio clásica *Inmovilización* y la ratio composicional *x1_ANC_x2_AC*. Se trata conceptualmente de la misma ratio de inmovilización del activo, a pesar de lo cual la versión composicional no presenta ninguna observación atípica extrema y la versión composicional presenta tres muy extremas. Estas observaciones no pueden atribuirse a características anómalas de las empresas, sino al funcionamiento anómalo de la ratio clásica.

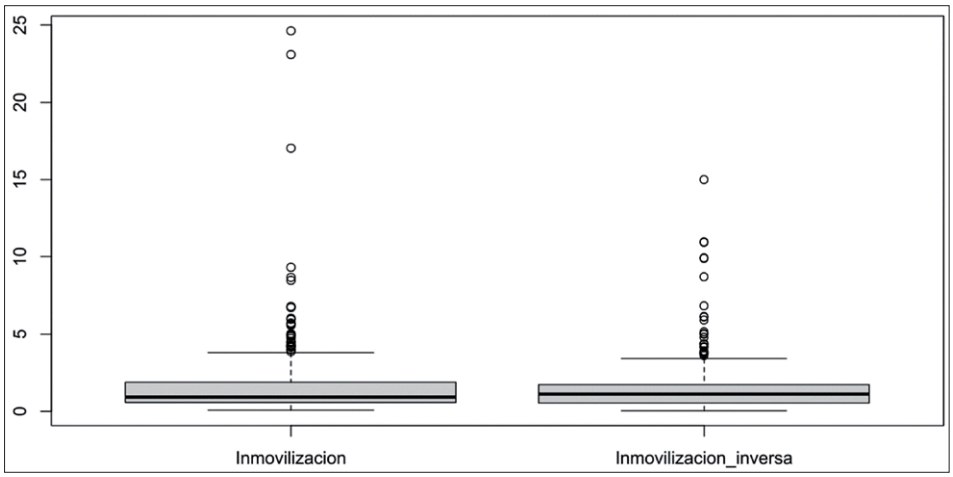
Observamos que el endeudamiento y la solvencia a largo plazo son, en realidad, la misma ratio con permutación del numerador y el denominador, pero el aspecto del gráfico es muy distinto. Muy a menudo no hay consenso sobre qué valor debe ir en el numerador y en el denominador de la ratio. Veamos otro ejemplo. Nosotros hemos definido la estructura del activo como el grado de inmovilización, es decir, la ratio de activo no corriente sobre activo corriente. Uno podría definir la ratio inversa que indicara no la importancia relativa del inmovilizado, sino la importancia relativa del realizable a corto plazo, las existencias y el disponible (en definitiva, los activos corrientes son más fáciles de convertir en líquido), lo que equivale a invertir la ratio.

Para calcular una ratio de activo corriente sobre activo no corriente, en el menú *Data > Manipulate > Calculate New Variable* introducimos las variables *x1_ANC* y *x2_AC* en el cuadro *Selected*. En el orden en el que se han introducido e independientemente de los nombres originales, CoDaPack llama a la primera variable *x1*; la segunda, *x2*; la tercera, *x3*, etc. (obsérvese que la letra *x* está en minúscula siempre). Tras clicar *Accept*, aparece un nuevo cuadro de diálogo donde hay que entrar la fórmula que se desea emplear para crear la nueva variable, en este caso el cociente de la segunda variable sobre la primera, *x2/x1* en la casilla superior y el nombre que queramos dar a la nueva variable, por ejemplo, *Inmovilizacion_inversa*, en la inferior.



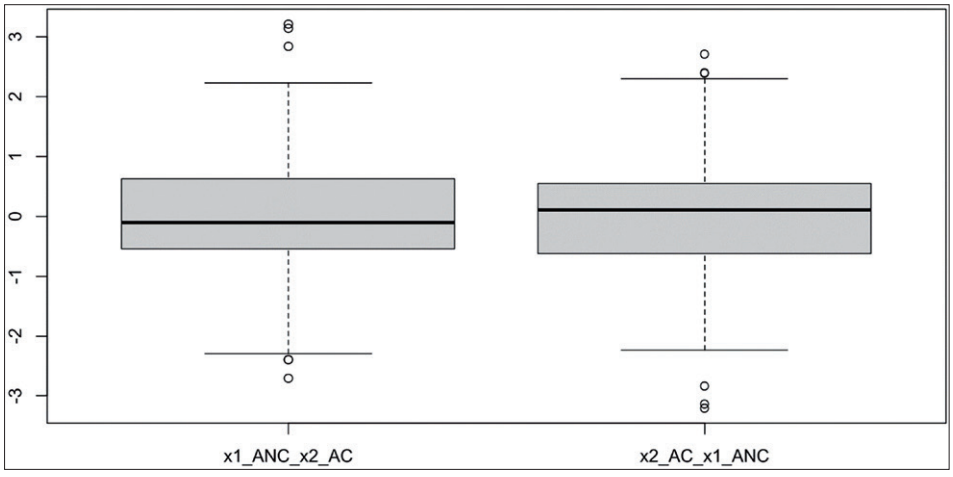
Nuevamente, el aspecto de dos ratios que son en realidad la misma tras invertir el numerador y el denominador no es igual. Cuando usamos la ratio original, aparecen observaciones atípicas por la parte alta (que tienen valores del activo no corriente muy superiores a los del activo corriente). Cuando usamos la ratio inversa, esperaríamos que estas observaciones atípicas aparecieran en la parte baja de la ratio invertida, pero no es así. Por el contrario, vuelven a salir

observaciones atípicas en la parte alta, pero ahora destacan las empresas con valores del activo corriente muy superiores a los del activo no corriente. Lo que ocurre en realidad es que en la ratio original aparecen como atípicas las empresas con activo corriente muy bajo, y en la ratio invertida. las empresas con el activo no corriente muy bajo. Recordemos que los valores bajos precisamente en el denominador son los que provocan las observaciones atípicas extremas en las ratios clásicas. En cambio, en la log-ratio por pares *x1_ANC_x2_AC* apenas aparecen observaciones atípicas y no son nada extremas.



Por el contrario, invertir una log-ratio por pares solo modifica su signo. Para calcular la log-ratio por pares de activo corriente sobre activo no corriente, en el menú *Data* > *Transformation* > *ALR* introducimos en este orden las partes *z.x2_AC* y *z.x1_ANC* con la opción *Raw-ALR*. En consonancia con el resto, cambiamos el nombre de la variable creada por *x2_AC_x1_ANC*.

Los gráficos de caja de las log-ratios por pares original e invertida son una simple imagen de espejo el uno del otro. Las observaciones atípicas no son nada extremas y corresponden a las mismas empresas.



4.6. Para saber más. Log-ratios isométricas y aditivas

Otras transformaciones por log-ratios también son útiles para el modelado estadístico, como las llamadas *coordenadas por log-ratios isométricas* (Egozcue y Pawlowsky-Glahn, 2005; Egozcue et al., 2003; Pawlowsky-Glahn y Egozcue, 2011) llamadas también *coordenadas por log-ratios ortonormales* (Martín-Fernández, 2019). En el análisis de los estados financieros han sido utilizadas por Arimany-Serrat et al. (2022; 2023), Carreras-Simó y Coenders (2021), Coenders (2025), Coenders y Arimany-Serrat (2023), Coenders et al. (2023a), Escaramís y Arbussà (2025), Linares-Mustarós et al. (2018; 2022) y Molas-Colomer et al. (2024).

Otra transformación muy conocida es la llamada de *log-ratios aditivas*, e históricamente fue la primera que se desarrolló (Aitchison, 1982). Se trata de un caso particular de las log-ratios por pares en la que todas las $D - 1$ log-ratios tienen la misma parte en el denominador. Su uso no parece muy prometedor en el análisis de los estados financieros, y, de hecho, no nos consta ninguna aplicación. Por el contrario, se adecuan mucho a los indicadores de las Normas Europeas de Información sobre Sostenibilidad. Por ejemplo, los consumos de energía y de agua, las emisiones de CO_2, etc. pueden expresarse como ratios con el denominador común de ingresos de explotación.

La elección de las D partes la realizan los propios investigadores según sus objetivos, y hay ejemplos tanto con más (Carreras-Simó y Coenders, 2020) como con menos (Arimany-Serrat y Sgorla, 2024; Coenders y Arimany-Serrat, 2023; Saus-Sala et al., 2021; 2023) que las usadas aquí.

5. Medias sectoriales fidedignas. De la media aritmética a la media geométrica

5.1. El centro composicional y las propiedades de la media geométrica

El uso estadístico más simple concebible de las ratios financieras es calcular las medias de las ratios para buscar valores representativos dentro de un sector. El *centro composicional* (Aitchison, 1997) se utiliza para calcular los valores medios de los datos composicionales. Se define como el conjunto de medias geométricas de todas las empresas para cada parte, normalizadas a la suma unitaria por conveniencia. Esto no debe confundirse con las medias geométricas de todas las partes para cada empresa utilizadas para calcular las log-ratios centradas en las ecuaciones (25) y (27).

Al igual que las ratios, las medias geométricas se centran en las diferencias relativas en lugar de en las absolutas. Las ratios y las medias geométricas son mutuamente compatibles y deben usarse juntas para las variables en una escala de razón. Si volvemos a tomar las empresas 3, 4 y 5 en el ejemplo de la tabla 1, la media geométrica de los valores de x_2 100, 1 000 y 10 000 es $g(x_2) = \sqrt[3]{100 \times 1\,000 \times 10\,000} = 1\,000$. Esto es así porque, en términos relativos, la diferencia entre 1 000 y 100, que es 1 000/100 = 10, es la misma que la diferencia relativa entre 10 000 y 1 000, que es 10 000/1 000 = 10. Por el contrario, la media aritmética está más cerca de los valores absolutos más altos, sin tener en cuenta las diferencias relativas: $\bar{x}_2 = (100 + 1\,000 + 10\,000)/3 = 3\,700$.

El centro calculado como una media geométrica bajo el enfoque CoDa permite calcular las ratios financieras clásicas medias a nivel sectorial (Arimany-Serrat y Coenders, 2025; Arimany-Serrat y Sgorla, 2024; Saus-Sala et al., 2021; 2023; 2024). La media geométrica tiene la propiedad atractiva de que la ratio entre las medias geométricas de dos partes es igual a la media geométrica de sus ratios. Sea $g(x_i)$ la media geométrica de la parte iésima sobre una muestra de empresas:

$$g\left(\frac{x_i}{x_j}\right) = \frac{g(x_i)}{g(x_j)} \tag{29}$$

En el mismo ejemplo de la tabla 1, la media geométrica de las ratios x_2/x_1 para las empresas 3, 4 y 5 es $g(x_2/x_1) = \sqrt[3]{0{,}01 \times 1 \times 100} = 1$, que es igual a la ratio de las medias geométricas de x_2 y x_1 $g(x_2)/g(x_1) = 1\,000/1\,000 = 1$.

La media aritmética no tiene esta propiedad. El cálculo primero de las medias aritméticas de los valores contables a nivel sectorial y luego de las ratios financieras clásicas entre dichas medias puede estar en contradicción con los resultados de calcular primero las ratios clásicas de cada empresa y luego las medias aritméticas de dichas ratios (Saus-Sala et al., 2021).

En el mismo ejemplo de la tabla 1, la media aritmética de las ratios x_2/x_1 para las empresas 3, 4 y 5 es $(0{,}01 + 1 + 100)/3 = 33{,}67$, que no es el cociente de las medias aritméticas de x_1 y x_2 $\bar{x}_1/\bar{x}_2 = 3\,700/3\,700 = 1$.

Las medias geométricas tienen otra propiedad atractiva en el análisis de los estados financieros. La media geométrica de una ratio con el numerador y el denominador permutados es la inversa de la media geométrica de la ratio original (Arimany-Serrat y Coenders, 2025; Arimany-Serrat y Sgorla, 2024):

$$g\left(\frac{x_i}{x_j}\right) = \frac{1}{g\left(\frac{x_j}{x_i}\right)} \quad (30)$$

Esta propiedad garantiza la consistencia de los resultados de dos investigadores utilizando versiones permutadas de la misma ratio. En el mismo ejemplo de la tabla 1, la media geométrica de las ratios x_2/x_1 para las empresas 4, 5 y 6 es $g(x_2/x_1) = \sqrt[3]{1 \times 100 \times 10\,000} = 100$ que es la inversa de la media geométrica de la ratio x_1/x_2 $g(x_1/x_2) = \sqrt[3]{1 \times 0{,}01 \times 0{,}0001} = 0{,}01$.

La media aritmética no tiene esta propiedad. En el mismo ejemplo de la tabla 1, la media aritmética de las ratios x_2/x_1 para las empresas 4, 5 y 6 es $(1 + 100 + 10\,000)/3 = 3\,367$, que no es la inversa de la media aritmética de las ratios x_1/x_2 $(1 + 0{,}01 + 0{,}0001)/3 = 0{,}3367$. El primer resultado sugiere que x_2 supera a x_1 en un factor de aproximadamente 3 000, mientras que el segundo resultado sugiere que x_1 está por debajo de x_2 en un factor de aproximadamente un tercio.

La tabla 2 muestra las medias geométricas en un sector ficticio solo con el objeto de practicar el cálculo de las ratios medias sectoriales a partir de las mismas ecuaciones (1) hasta (12), aplicadas ahora a las medias geométricas.

Tabla 2. Medias geométricas de x_1 a x_6 en un sector ficticio

x_1: activo no corriente	0,30
x_2: activo corriente	0,10
x_3: pasivo no corriente	0,05
x_4: pasivo corriente	0,10
x_5: ingresos de explotación	0,25
x_6: gastos de explotación	0,20

Usando las propiedades descritas, la media sectorial de la rotación puede calcularse como:

$$g(x_5)/(g(x_1) + g(x_2)) = 0{,}25/(0{,}30 + 0{,}10) = 0{,}625 \quad (31)$$

La de la rotación del activo corriente como:

$$(32)\ g(x_5)/g(x_2) = 0{,}25/0{,}10 = 2{,}5$$

El margen como:

$$(33)\ (g(x_5) - g(x_6))/g(x_5) = (0{,}25 - 0{,}20)/0{,}25 = 0{,}2$$

El apalancamiento como:

$$(34)\ (g(x_1) + g(x_2))/(g(x_1) + g(x_2) - g(x_3) - g(x_4)) = (0{,}30 + 0{,}10)/(0{,}30 + 0{,}10 - 0{,}05 - 0{,}10) = 1{,}6$$

El ROA como:

$$(35)\ (g(x_5) - g(x_6))/(g(x_1) + g(x_2)) = (0{,}25 - 0{,}20)/(0{,}30 + 0{,}10) = 0{,}125$$

El ROE como:

$$(36)\ (g(x_5) - g(x_6))/(g(x_1) + g(x_2) - g(x_3) - g(x_4)) = (0{,}25 - 0{,}20)/(0{,}30 + 0{,}10 - 0{,}05 - 0{,}10) = 0{,}2$$

El endeudamiento como:

$$(37)\ (g(x_3) + g(x_4))/(g(x_1) + g(x_2)) = (0{,}05 + 0{,}10)/(0{,}30 + 0{,}10) = 0{,}375$$

El endeudamiento a corto plazo como:

$$(38)\ g(x_4)/(g(x_1) + g(x_2)) = 0{,}10/(0{,}30 + 0{,}10) = 0{,}250$$

La solvencia a largo plazo es la permutación del numerador y el denominador del endeudamiento, con lo que resultado es 1/0,375 = 2,667 y se calcula como:

$$(39)\ (g(x_1) + g(x_2))/(g(x_3) + g(x_4)) = (0{,}30 + 0{,}10)/(0{,}05 + 0{,}10) = 2{,}667$$

La solvencia a corto plazo como:

$$(40)\ g(x_2)/g(x_4) = 0{,}10/0{,}10 = 1$$

La inmovilización del activo como:

$$(41)\ g(x_1)/g(x_2) = 0{,}30/0{,}10 = 3$$

La maduración de la deuda como:

$$(42)\ g(x_3)/g(x_4) = 0{,}05/0{,}10 = 0{,}5$$

Cada euro de activo corriente genera 2,5 euros de cifra de negocios, con lo que el período medio de rotación del activo corriente es 1/2,5 años, que equivale a unos cinco meses. En conjunto se trata de un sector con buena rentabilidad. El margen es el 20 % de las ventas, la rentabilidad económica o sobre los activos (ROA) del 12,5 % y la rentabilidad financiera o sobre los fondos propios (ROE) del 20 %. El

37,5 % de los activos están financiados por deuda, o, lo que es lo mismo, los activos exceden a los pasivos por un factor de 2,667. Se trata de un endeudamiento aceptable. El 25 % del activo está financiado por pasivo corriente (deuda a corto plazo), pero lo que es preocupante es que, según la ratio de solvencia a corto plazo, esta deuda a corto plazo apenas puede pagarse realizando el activo corriente. Es deseable que la ratio de solvencia a corto plazo sea mayor que 1. La mayor parte del activo es no corriente por un factor de 3 a 1. La mayor parte del pasivo es corriente, es decir, la deuda a largo plazo, es la mitad de la deuda a corto plazo. La recomendación más inmediata para una empresa media del sector sería reconvertir parte de la deuda de corto plazo a largo plazo, cosa que mejoraría, a la vez, la solvencia a corto plazo y la maduración de la deuda.

Las medias geométricas permiten presentar los resultados del análisis composicional de un sector en términos de ratios financieras clásicas, que son mejor comprendidas por la comunidad investigadora y profesional contable y financiera que las log-ratios que propone la metodología CoDa. El análisis se puede realizar para todo el sector o para subdivisiones previamente identificadas dentro del sector, por tipo de sociedad anónima o limitada, por comunidad autónoma, separando pymes frente a grandes empresas, etc.

Hasta aquí, hemos aprendido que cuando los datos están en una escala de razón, lo que significa que las diferencias relativas y no absolutas son de interés, las ratios, los logaritmos y las medias geométricas constituyen operaciones que tienen sentido y que deben usarse conjuntamente. No tiene sentido usar ratios pretendiendo que se están buscando diferencias relativas y luego no usar el logaritmo o no usar la media geométrica como si se estuvieran buscado diferencias absolutas. Estas tres operaciones son el núcleo del análisis CoDa.

5.2. Manos a la obra con CoDaPack. Calculamos medias representativas del sector o partes de este

Vamos a calcular las ratios medias sectoriales con los datos del mismo ejemplo del sector vitivinícola. Abrimos el archivo de datos en formato *.cdp* (menú *File* > *Open Workspace*). Comprobamos que la tabla activa sea *clrx1_2.5* en el desplegable *Tables*. El centro composicional se obtiene por medio del menú *Statistics* > *Compositional Statistics Summary*. Introducimos las partes *z.x1_ANC, z.x2_AC, z.x3_PNC, z.x4_PC, z.x5_IE* y *z.x6_GE* en el cuadro *Selected* solo con la opción *Center*.

Estos son los valores que hay que usar para calcular cualesquiera ratios financieras clásicas representativas de la empresa media del sector:

Compositional statistics summary:
NA's:
0

Sample size:
368

Statistics

	Center
z.x1_ANC	0.2372
z.x2_AC	0.2319
z.x3_PNC	0.0385
z.x4_PC	0.0975
z.x5_IE	0.1981
z.x6_GE	0.1968

Veamos algunos ejemplos. Según la ecuación (31), la ratio sectorial de rotación se calcula como $g(x_5)/(g(x_1) + g(x_2)) = 0{,}1981/(0{,}2372 + 0{,}2319) = 0{,}422$. Según la ecuación (32), la ratio sectorial de rotación del activo corriente se calcula como $g(x_5)/g(x_2) = 0{,}1981/0{,}2319 = 0{,}854$. Según la ecuación (33), la ratio sectorial de margen se calcula como $(g(x_5) - g(x_6))/g(x_5) = (0{,}1981 - 0{,}1968)/0{,}1981 = 0{,}007$. Todas las ratios clásicas se han calculado de este modo (ecuaciones (31) hasta (42)) en la tabla 3. Pueden calcularse incluso las ratios de apalancamiento y ROE porque, aunque algunas empresas individuales tengan patrimonio neto negativo, el patrimonio neto medio calculado a partir de las medias geométricas es positivo: $g(x_1) + g(x_2) - g(x_3) - g(x_4) = 0{,}2372 + 0{,}2319 - 0{,}0385 - 0{,}0975 = 0{,}3331$.

Tabla 3. Ratios clásicas representativas del sector vitivinícola a partir de las medias geométricas

Rotación	0,422
Rotación del activo corriente	0,854
Margen	0,007
Apalancamiento	1,408
ROA	0,003
ROE	0,004
Endeudamiento	0,290
Endeudamiento a corto plazo	0,208
Solvencia a largo plazo	3,449
Solvencia a corto plazo	2,378
Inmovilización del activo	1,023
Maduración de la deuda	0,395

Nos encontramos con un sector con unos márgenes y rentabilidades (ROA y ROE) muy modestos (ninguno de los indicadores alcanza 0,01 = 1 %), pero con endeudamientos bajos y una solvencia tanto a largo plazo como a corto plazo muy sólida. Según el endeudamiento, solo el 29 % de los activos están financiados por deuda y, según la solvencia a corto plazo, realizando el activo corriente se podría pagar el pasivo corriente 2,378 veces. El activo se divide casi a partes iguales entre el no corriente y el corriente. La maduración de la deuda indica que la deuda con vencimiento a largo plazo es muy inferior a la deuda con vencimiento a corto plazo, pero la elevada solvencia hace que este dato no sea preocupante. La rotación baja viene determinada por la venta de vinos añejos y es bastante inevitable. Dado

que el margen parece ser la variable clave, la recomendación obvia para el sector es aplicar técnicas del llamado *yield management* para aumentar los precios de venta en canales de distribución y segmentos de mercado seleccionados, y controlar mejor los costes de producción.

Veamos un ejemplo de lo que implicaría el cálculo con las medias aritméticas. Calculemos dichas medias aritméticas para las partes *z.x5_IE* y *z.x6_GE*, que son las necesarias para calcular el margen, junto con la media aritmética de la propia variable *Margen*. Usamos el menú *Statistics* > *Classical Statistics Summary*. Si marcamos las opciones *Mean* (media) y *Standard Deviation* (desviación típica) e introducimos *z.x5_IE*, *z.x6_GE* y *Margen* en el cuadro *Selected*, obtenemos:

Clasical statistics summary:
NA's:
0

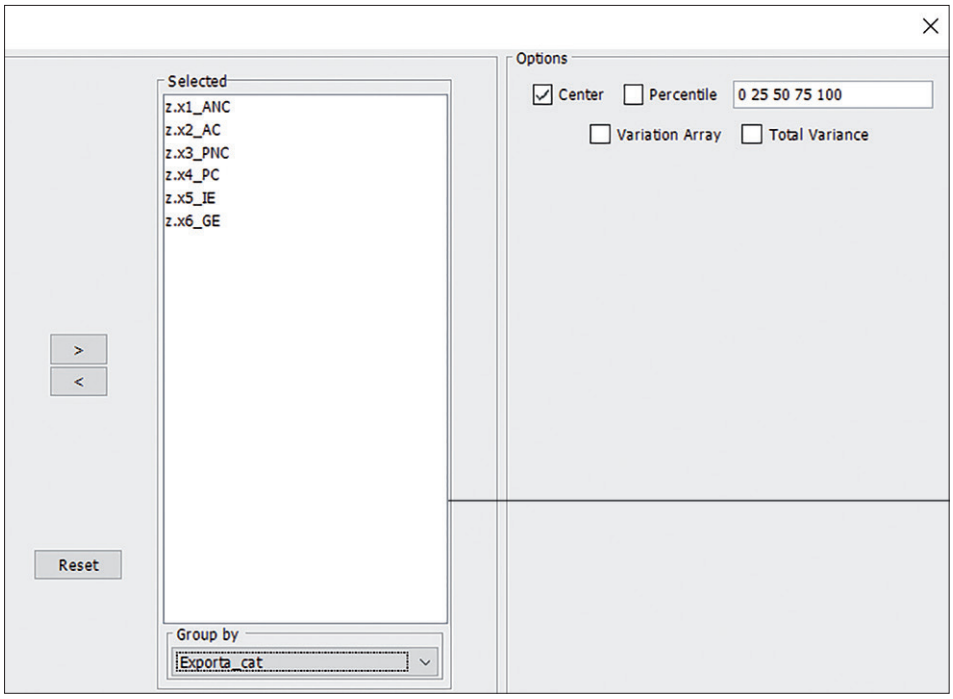
Sample size:
368

Statistics

	Mean	Std.Dev
z.x5_IE	7643.1749	16547.4875
z.x6_GE	7128.8180	16007.2985
Margen	-0.0505	0.5093

Los resultados son muy diferentes a los obtenidos con las medias geométricas (0,007 = + 0,7 %) y mutuamente contradictorios. Si calculamos la ratio de margen para cada empresa y luego la media aritmética de dichas ratios, obtenemos –0,0505 = –5,1 %, que es negativo. Si calculamos primero las medias aritméticas de los ingresos y los gastos de explotación y luego la ratio sectorial de margen sobre esas medias, obtenemos (7643,1749 – 7128,8180)/7643,1749 = 0,067 = + 6,7 %, positivo.

El análisis mediante las medias geométricas puede subdividirse por grupos de empresas definidos por una variable categórica. Por ejemplo, si queremos distinguir entre las empresas exportadoras y las que no lo son, entramos en el menú *Statistics* > *Compositional Statistics Summary*. Introducimos las partes *z.x1_ANC*, *z.x2_AC*, *z.x3_PNC*, *z.x4_PC*, *z.x5_IE* y *z.x6_GE* en el cuadro *Selected* solo con la opción *Center*. Seleccionamos *Exporta_cat* en el desplegable *Group by*.



Obtenemos los siguientes resultados, donde el grupo 0 identifica las empresas no exportadoras y el grupo 1, las empresas exportadoras (106 del total de 268).

Compositional statistics summary:
NA's:
0

-- Group 0.0 --
Sample size:
262

Statistics

	Center
z.x1_ANC	0.2531
z.x2_AC	0.2253
z.x3_PNC	0.0387
z.x4_PC	0.0913
z.x5_IE	0.1948
z.x6_GE	0.1968

-- Group 1.0 --
Sample size:
106

Statistics

	Center
z.x1_ANC	0.2007
z.x2_AC	0.2471
z.x3_PNC	0.0378
z.x4_PC	0.1140
z.x5_IE	0.2051
z.x6_GE	0.1953

Hacemos lo mismo con la variable *Subvenciones_cat*. Obtenemos los siguientes resultados, donde el grupo 0 identifica las empresas sin subvenciones y el grupo 1, las empresas con subvenciones (262 del total de 368).

Compositional statistics summary:
NA's:
0

-- Group 1.0 --
Sample size:
262

Statistics

	Center
z.x1_ANC	0.2479
z.x2_AC	0.2302
z.x3_PNC	0.0400
z.x4_PC	0.0937
z.x5_IE	0.1981
z.x6_GE	0.1901

-- Group 0.0 --
Sample size:
106

Statistics

	Center
z.x1_ANC	0.2119
z.x2_AC	0.2351
z.x3_PNC	0.0349
z.x4_PC	0.1071
z.x5_IE	0.1975
z.x6_GE	0.2136

A partir de estas medias geométricas por subgrupos, en la tabla 4 calculamos las ratios clásicas de la manera habitual con las ecuaciones (31) hasta (42). Observamos que las empresas exportadoras tienen mejores márgenes, ROA y ROE, menor solvencia tanto a corto como a largo plazo (aunque siguen siendo plenamente aceptables), un activo menos inmovilizado y una deuda de maduración más corta. Las empresas con subvenciones tienen mejores márgenes, ROA y ROE, tienen mejor solvencia tanto a corto como a largo plazo, un activo más inmovilizado y una deuda de maduración más a largo plazo. No se observan diferencias importantes en ninguna de las dos rotaciones.

Tabla 4. Ratios clásicas medias del sector vitivinícola según exportación y subvenciones a partir de las medias geométricas

	Total sector	No exportadoras	Exportadoras	Sin subvenciones	Con subvenciones
Rotación	0,422	0,407	0,458	0,442	0,414
Rotación del activo corriente	0,854	0,865	0,830	0,840	0,861
Margen	0,007	−0,010	0,048	−0,082	0,040
Apalancamiento	1,408	1,373	1,513	1,466	1,388
ROA	0,003	−0,004	0,022	−0,036	0,017
ROE	0,004	−0,006	0,033	−0,053	0,023

	Total sector	No exportadoras	Exportadoras	Sin subvenciones	Con subvenciones
Endeudamiento	0,290	0,272	0,339	0,318	0,280
Endeudamiento a corto plazo	0,208	0,191	0,255	0,240	0,196
Solvencia a largo plazo	3,449	3,680	2,950	3,148	3,576
Solvencia a corto plazo	2,378	2,468	2,168	2,195	2,457
Inmovilización del activo	1,023	1,123	0,812	0,901	1,077
Maduración de la deuda	0,395	0,424	0,332	0,326	0,427

5.3. Para saber más. Medias aritméticas de las log-ratios centradas

Ejemplos sencillos de trabajos publicados en los que se explican las medias geométricas son los de Arimany-Serrat y Coenders (2025); Arimany-Serrat y Sgorla (2024) y Saus-Sala et al. (2021).

Uno puede preguntarse por qué los promedios sectoriales no se calculan a partir de las log-ratios. Implícitamente es así. Se puede demostrar que las medias aritméticas calculadas sobre las log-ratios centradas son equivalentes a las medias geométricas calculadas a partir de los valores contables que se han presentado en el apartado 5.1. Lo único que hay que hacer es aplicar la función exponencial a las medias aritméticas de las log-ratios centradas (Aitchison, 1997).

Esto deriva de la siguiente propiedad que hace equivalente el logaritmo de la media geométrica a la media aritmética de los logaritmos, y la media geométrica a la exponencial de la media aritmética de los logaritmos, propiedad que reafirma el carácter complementario de los logaritmos y las medias geométricas:

$$\log\left(\sqrt[n]{x_1 x_2 \cdots x_n}\right) = \frac{1}{n}\left(\log\left(x_1\right) + \log\left(x_2\right) + \cdots + \log\left(x_n\right)\right)$$

$$\sqrt[n]{x_1 x_2 \cdots x_n} = e^{\frac{1}{n}\left(\log\left(x_1\right) + \log\left(x_2\right) + \cdots + \log\left(x_n\right)\right)} \quad (43)$$

En este libro hemos optado por utilizar directamente las medias geométricas de los valores contables por su carácter más intuitivo.

6. Todas las empresas y ratios en un único gráfico. El *biplot* composicional

6.1. Construcción, interpretación y proyecciones

Al igual que cualesquiera otros datos estadísticos, los datos composicionales requieren herramientas de visualización para ayudar a los investigadores a interpretar grandes tablas de datos con muchas empresas y partes. Con este fin, Aitchison (1983) extendió el conocido procedimiento clásico de *análisis en componentes principales* (Greenacre et al., 2022; Hotelling, 1933) al caso composicional. Este método pertenece a la familia del *análisis estadístico multivariante*, y la extensión se reduce a someter las D log-ratios centradas de la ecuación (27) como variables de entrada en un análisis en componentes principales estándar, cambiando solo las reglas de interpretación de los resultados, interpretación que va a ser puramente gráfica.

Un análisis composicional en componentes principales calcula un pequeño número de combinaciones lineales no correlacionadas de las log-ratios centradas, llamadas *dimensiones*, que explican la mayor parte posible de la suma de las varianzas de todas las log-ratios centradas. De esta manera, el conjunto de datos original con muchas log-ratios centradas se puede resumir con solo unas pocas dimensiones (con un poco de suerte solo dos) que sean adecuadas para una visualización gráfica.

Las dos primeras dimensiones se representan en el llamado *biplot CoDa* de covarianza (Aitchison y Greenacre, 2002, a partir de los trabajos de Gabriel, 1971), la primera como eje horizontal y la segunda como eje vertical. El *biplot* CoDa o biplot composicional puede entenderse como la representación gráfica más precisa posible de un conjunto de datos composicionales en dos dimensiones. La bondad del ajuste se indica por el porcentaje de varianza de las log-ratios centradas explicada por las dos primeras dimensiones. Según nuestra propia experiencia, los porcentajes superiores al 70 % se consideran aceptables; los porcentajes superiores al 80 %, buenos, y los porcentajes superiores al 90 %, muy buenos. Aun así, un grado de imprecisión subsiste siempre, y toda interpretación que se haga sobre el *biplot* debe entenderse como aproximada.

El *biplot* CoDa para los datos de los estados financieros traza cada log-ratio centrada como una línea llamada *rayo*. El *vértice* o extremo del rayo representa el valor contable en el numerador de la log-ratio centrada. Las empresas individuales aparecen como puntos.

Al contrario que en el *biplot* clásico (no composicional), aquí los ángulos entre los rayos no se interpretan. Tampoco se interpretan los ángulos entre los rayos y los ejes horizontal y vertical que representan las dimensiones. De hecho, el objetivo no es ni tratar las dimensiones como indicadores financieros compuestos, ni interpretar dichas dimensiones. El *biplot* se considera una herramienta puramente gráfica, y el protagonismo recae en las empresas y en las ratios financieras.

Carreras-Simó y Coenders (2020) y Saus-Sala et al. (2021; 2023) destacan la herramienta interpretativa más importante del *biplot* CoDa en el análisis de los estados financieros. Se pueden trazar líneas adicionales que unan los vértices de un par de rayos y representen las log-ratios por pares entre los dos valores contables correspondientes de los numeradores de las log-ratios centradas que se unen. Estas líneas adicionales se denominan *enlaces*. La *proyección ortogonal* de todas las empresas a lo largo de la dirección definida por el enlace entre los vértices de un par de rayos muestra un orden aproximado de las empresas de acuerdo con la log-ratio por pares entre los dos valores contables correspondientes. La proyección ortogonal se realiza dejando caer las empresas sobre el enlace de tal manera que la dirección en la que caen las empresas forma un ángulo de noventa grados con el enlace. Por ejemplo, y_1, que representa la rotación del activo corriente, es una línea que une los vértices de los rayos que representan los ingresos de explotación (x_5) y los activos corrientes (x_2). Al dejar caer todas las empresas sobre esta línea formando un ángulo de noventa grados, las empresas situadas en el lado correspondiente a x_5 tienen una mayor rotación del activo corriente que las empresas situadas en el lado correspondiente a x_2.

De esta manera, el *biplot* CoDa es también una representación visual de cualquiera de las $D(D - 1)/2$ posibles ratios financieras calculadas a partir de dos valores contables cualesquiera. El usuario o usuaria puede dibujar tantos enlaces como desee. Dado que el análisis se realiza de todos modos a partir de las log-ratios centradas, la redundancia no es un problema. Eso sí, solo los enlaces largos que muestran log-ratios por pares de alta varianza conducen a direcciones informativas. Por lo tanto, las log-ratios por pares no deben dibujarse cuando los enlaces son muy cortos, en otras palabras, cuando los vértices de los dos rayos de las log-ratios centradas involucradas están muy juntos.

Las empresas más cercanas al origen de coordenadas del *biplot* son las más cercanas a la media sectorial descrita en el capítulo 5. Proyectar dicha media sobre un enlace cualquiera permitiría situar la media de la log-ratio por pares, y distinguir así las empresas superiores e inferiores a la media de la log-ratio por pares.

La capacidad de interpretar visualmente las ratios entre dos valores contables cualesquiera es de gran interés en el análisis de los estados financieros en general (Carreras-Simó y Coenders, 2020; Saus-Sala et al., 2021; 2023). El *biplot* composicional se convierte así en una herramienta intuitiva y útil para el análisis estratégico (Carreras-Simó y Coenders, 2020), ya que permite a los investigadores identificar rápidamente las empresas individuales que compiten sobre la base del margen, sobre la base de la rotación, o sobre la base del apalancamiento, o con una solvencia o una inmovilización del activo diferentes. Los resultados de los *biplots* también son muy útiles para los gerentes de una sola empresa. En el *biplot*, los gerentes pueden comparar visualmente el perfil financiero de su empresa con el de la empresa media o con el de cualquier otra empresa del sector que consideren su competencia directa o a la que quieran emular.

No hay un tamaño de muestra mínimo para realizar un *biplot*. De hecho, cuantas menos empresas haya, más fácil es identificarlas individualmente sobre el gráfico y compararlas entre sí.

6.2. Manos a la obra con CoDaPack. Visualizamos las empresas individuales del sector

Abrimos el archivo de datos en formato *.cdp* (menú *File* > *Open Workspace*). Comprobamos que la tabla activa sea *clrx1_2.5* en el desplegable *Tables*. El menú *Graphs* > *CLR-biplot* representa el *biplot* de covarianza. El menú calcula internamente las log-ratios centradas, de modo que se deben introducir los valores originales *z.x1_ANC*, z.*x2_AC*, *z.x3_PNC*, *z.x4_PC*, *z.x5_IE* y *z.x6_GE* en el cuadro *Selected*, nunca las log-ratios centradas. Los puntos se pueden colorear de acuerdo con una variable categórica que defina clústeres o cualquier otra subdivisión dentro de la industria seleccionándola en el desplegable *Group by*. Seleccionamos la variable *CA* (comunidad autónoma). La variable de agrupación debe ser categórica (marcada en naranja en la tabla de datos) o bien debe haber sido transformada previamente con el menú *Data* > *Manipulate* > *Numeric to Categoric*.

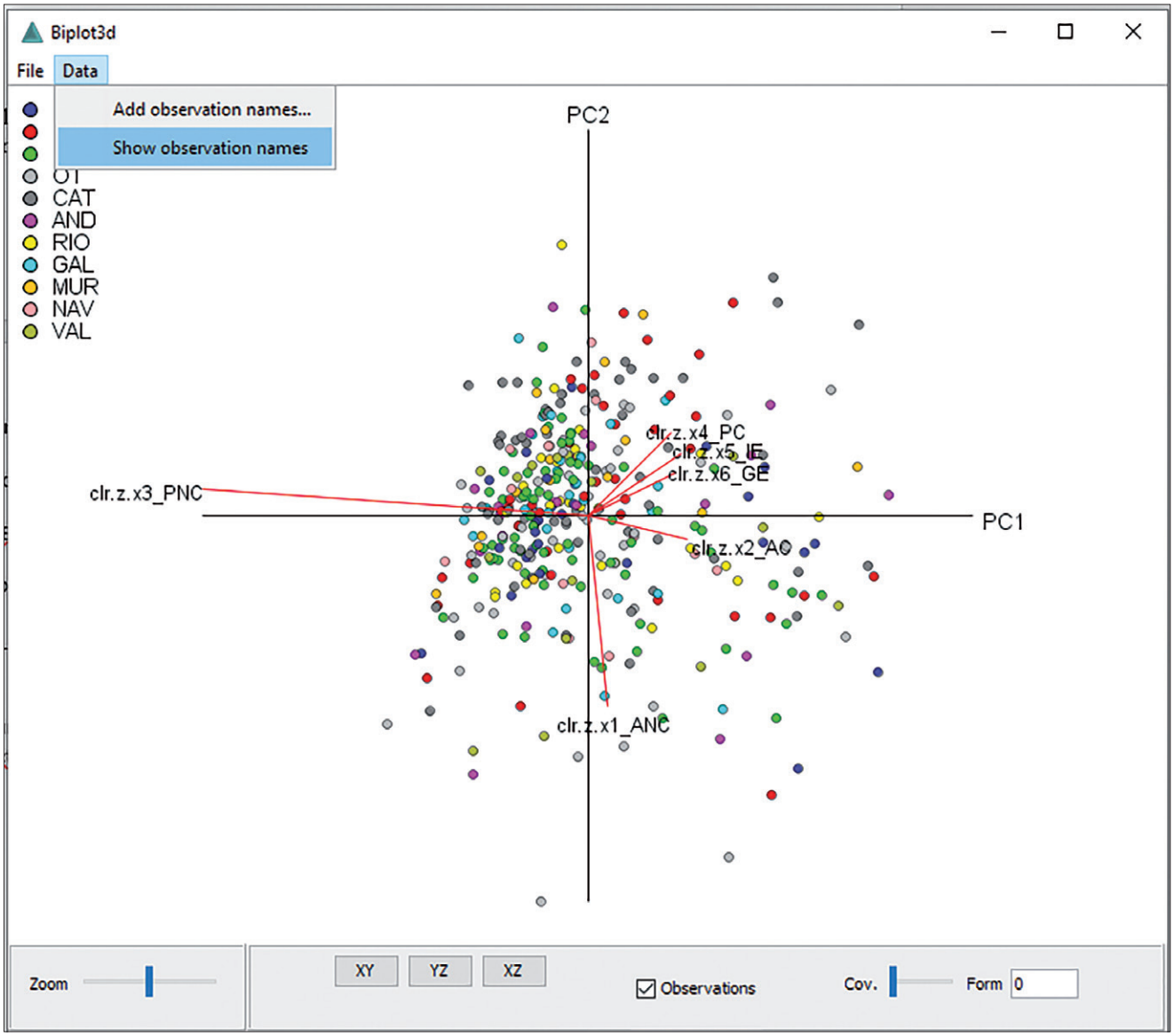
En primer lugar, obtenemos una tabla que nos informa, en su última columna, de los porcentajes de varianza explicados. En nuestro caso, las dos primeras dimensiones acumulan el 83,43 % de la varianza de las log-ratios centradas, resultado muy satisfactorio. La representación de las empresas y sus log-ratios por pares en un gráfico de solo dos dimensiones será fiable.

Biplot generated:
Data: z.x1_ANC z.x2_AC z.x3_PNC z.x4_PC z.x5_IE z.x6_GE
Principal Components:

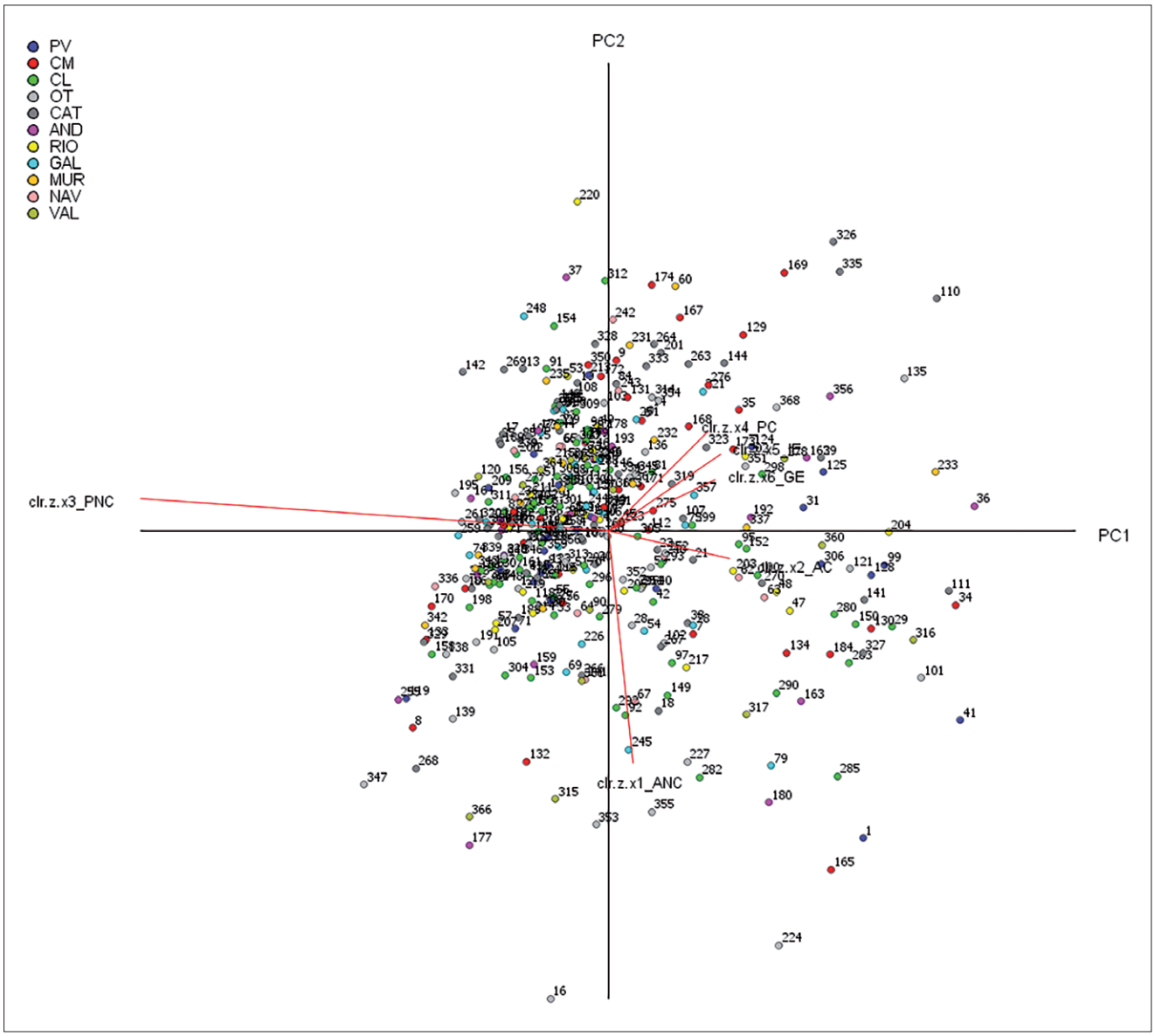
	clr.z.x1_ANC	clr.z.x2_AC	clr.z.x3_PNC	clr.z.x4_PC	clr.z.x5_IE	clr.z.x6_GE	Cum.Prop.Exp.
PC1	0.0478	0.2353	-0.9023	0.1940	0.2187	0.2065	0.6540
PC2	-0.8518	-0.1037	0.1182	0.3678	0.2803	0.1893	0.8343
PC3	0.1273	-0.0933	-0.0346	0.8042	-0.4483	-0.3553	0.9412
PC4	-0.2898	0.8704	0.0627	-0.1128	-0.2422	-0.2884	0.9914
PC5	0.0726	-0.0301	-0.0041	0.0314	0.6688	-0.7386	1.0000

El *biplot* aparece en una ventana gráfica. En realidad, CoDaPack recoge tres dimensiones llamadas *X*, *Y* y *Z* y unos botones en la parte inferior permitirían visualizar la tercera dimensión (*Z*) junto con la primera (botón *XZ*) o junto con la segunda (botón *YZ*). En la parte inferior derecha, una barra permitiría pasar del *biplot* de covarianza que conocemos (*Cov.*) a otros tipos de *biplot*. Por último, situando el ratón sobre el área del gráfico este se puede rotar para ver las tres dimensiones simultáneamente. Normalmente no usamos ninguna de estas opciones. Sí encontramos práctico el menú *Data* > *Show observation names*, que se puede utilizar para identificar empresas individuales por sus números de fila en el archivo de datos, útil sobre todo si hay pocas empresas.

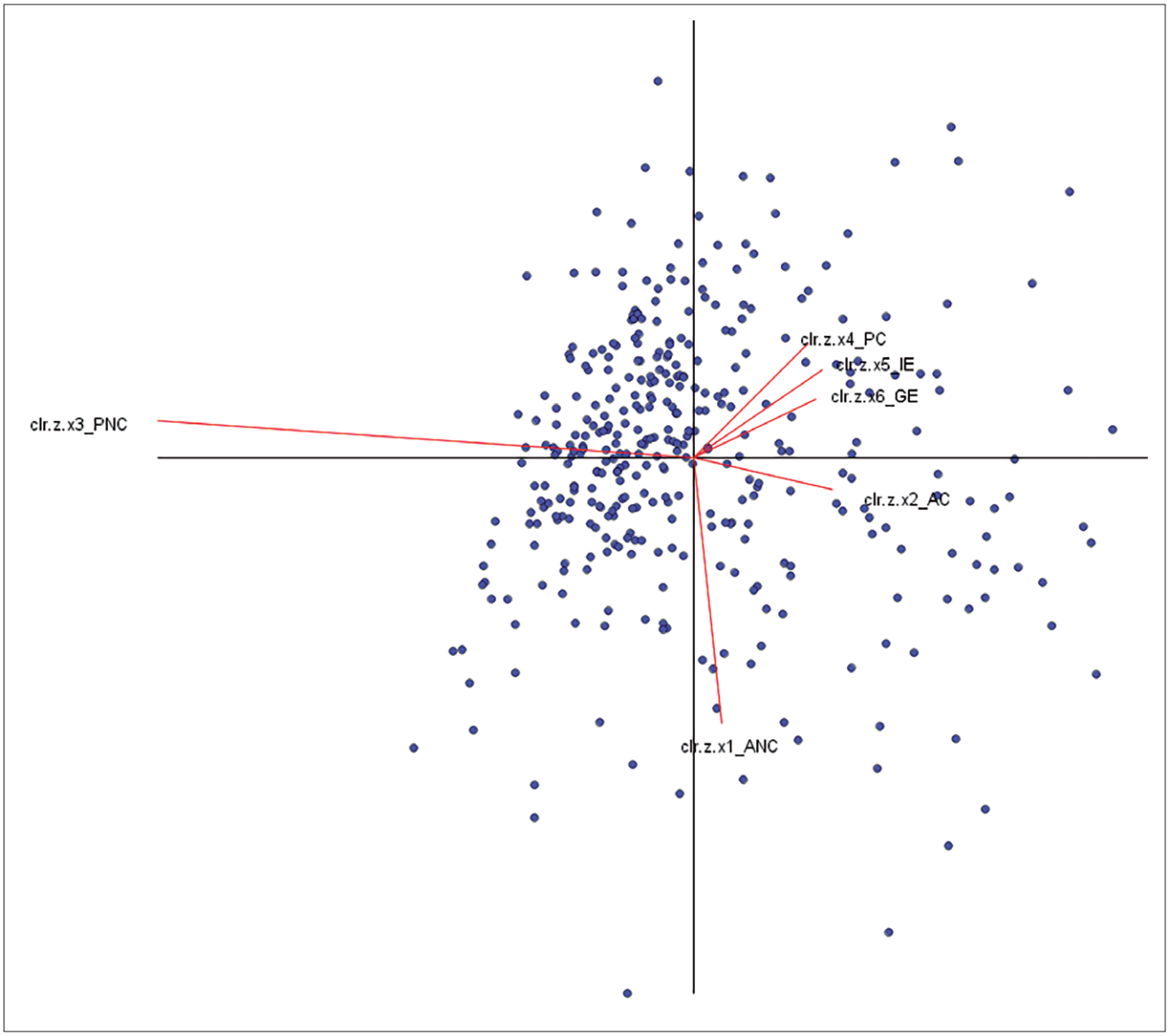
Si el usuario o usuaria desea que los puntos se etiqueten por una variable en el archivo de datos en lugar de por su fila, primero debe seleccionar dicha variable en la opción *Data > Add observation names*.



Tras hacerlo simplemente por el número de fila, obtenemos:

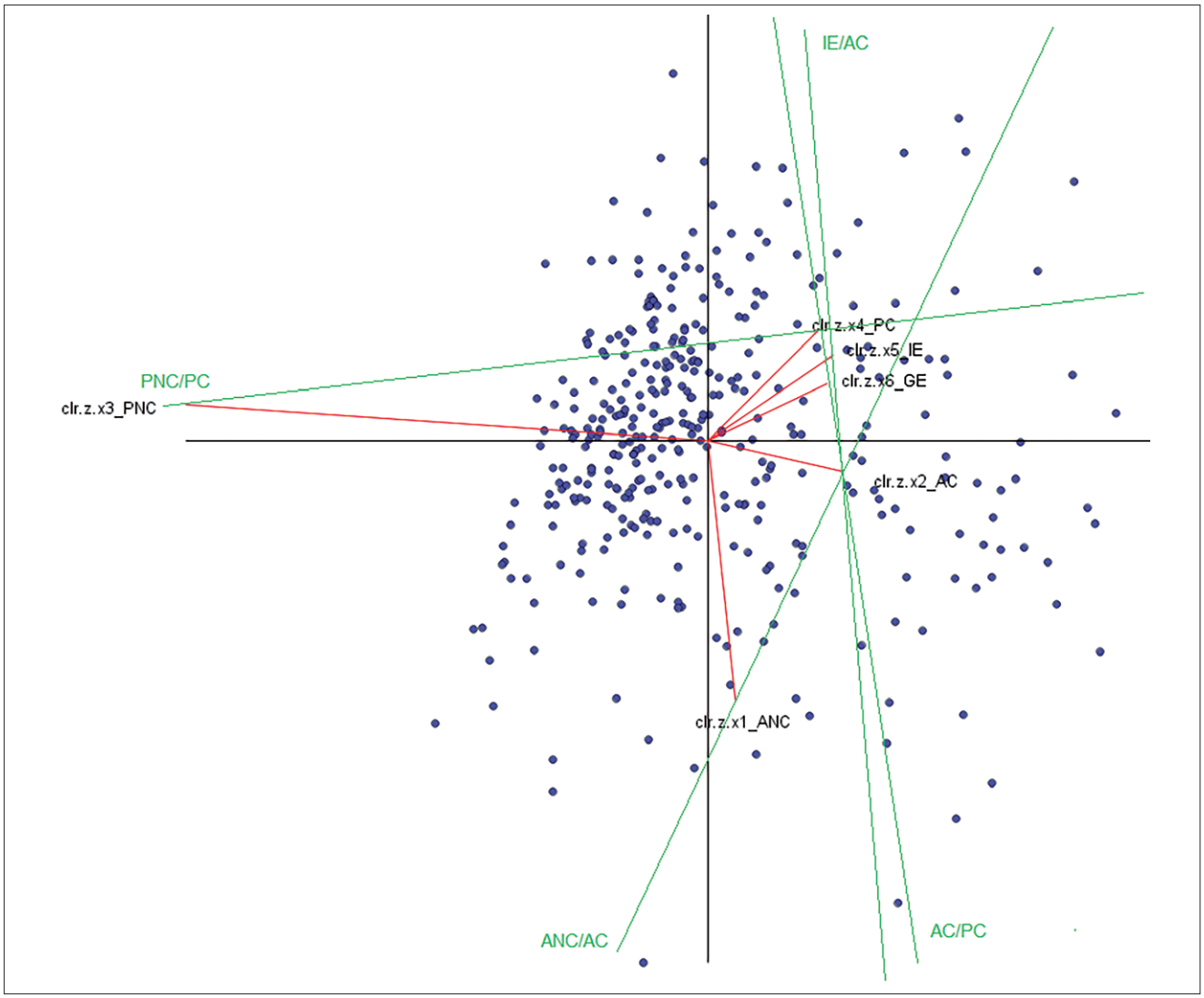


Etiquetar los puntos de las empresas es especialmente útil cuando hay pocas empresas fáciles de identificar. Como no es el caso, partimos de un *biplot* limpio, sin colores ni etiquetas de puntos, para añadir direcciones adicionales (enlaces) que representen ratios entre pares de valores contables. CoDaPack no lo hace y deben añadirse manualmente con un software de edición de gráficos.

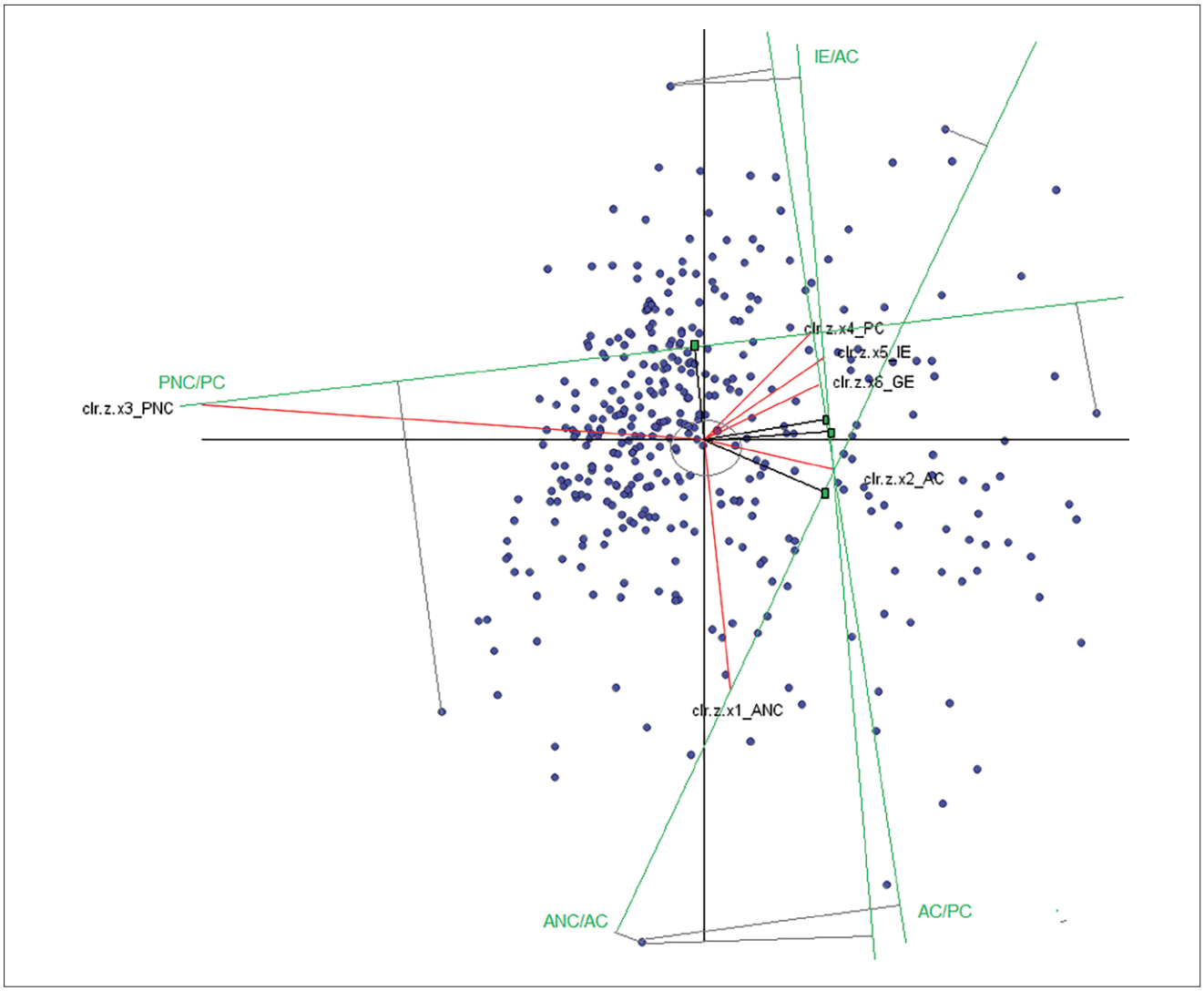


Representamos nuestras cinco log-ratios por pares, exceptuando la de margen $y_2 = \log(IE/GE)$ que relaciona los ingresos y los gastos de explotación, y que, al estar ambos vértices tan próximos entre sí, no sería fiable. Hay que tener en cuenta en qué sentido están orientados el numerador y el denominador de la ratio. Las empresas en la parte superior del *biplot* son las de mayor rotación del activo corriente $y_1 = \log(IE/AC)$. Las empresas en la inferior del *biplot* son las de mayor solvencia a corto plazo $y_3 = \log(AC/PC)$. Las empresas más hacia la izquierda son las de más larga maduración de la deuda $y_5 = \log(PNC/PC)$. Las empresas del cuadrante inferior izquierdo son las de mayor inmovilización del activo $y_4 = \log(ANC/AC)$. Para evitar confusiones, hemos etiquetado las

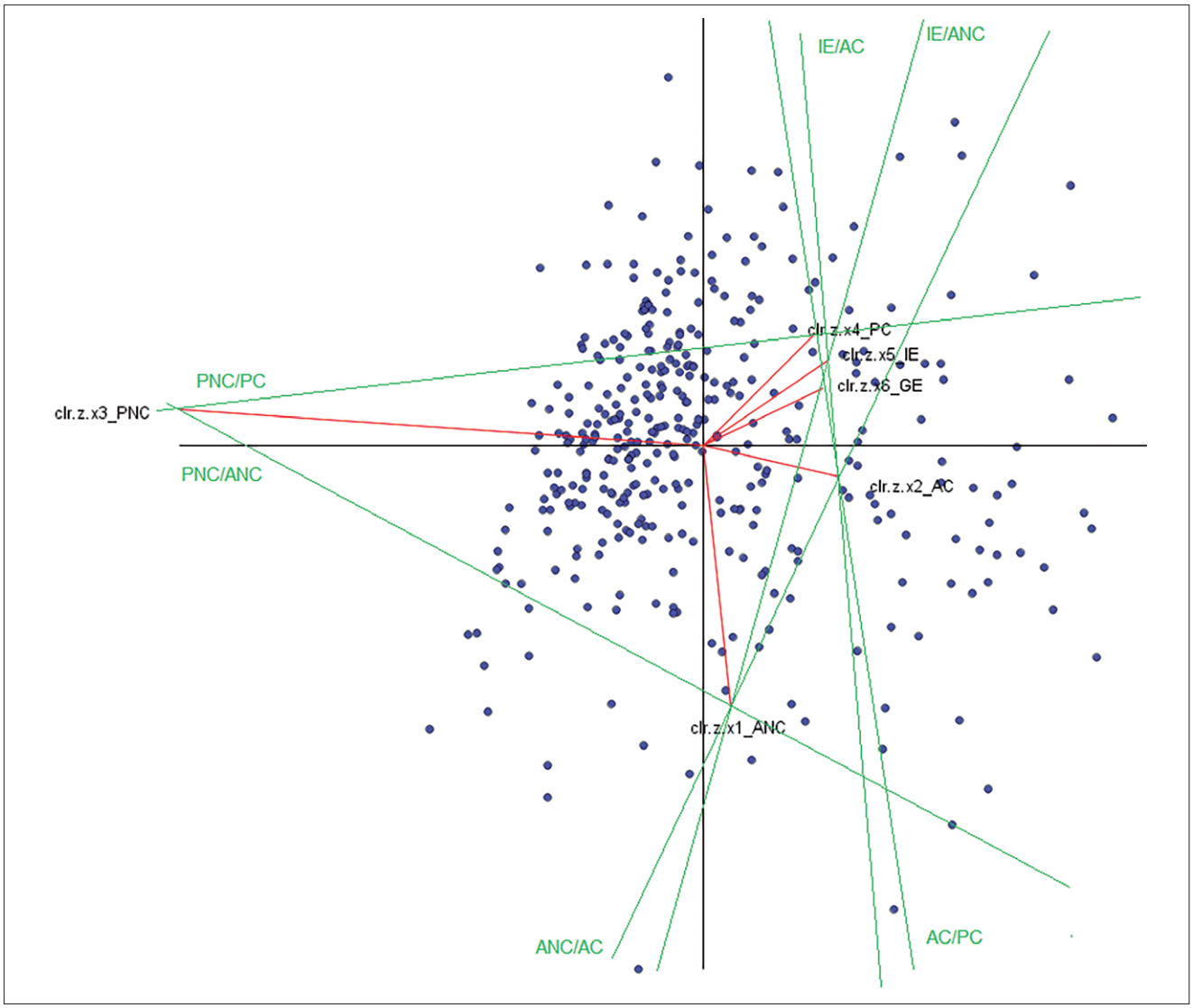
ratios del lado en que toman valores elevados, es decir, el lado más próximo al numerador.



Las posiciones relativas de las empresas sobre las log-ratios por pares se obtienen proyectándolas ortogonalmente (es decir, con un ángulo de noventa grados) sobre sus direcciones respectivas. Mostramos ahora las empresas con valores máximo y mínimo de cada una de las log-ratios. Las empresas próximas al origen de coordenadas (que hemos marcado dentro de un círculo) tienen log-ratios parecidas a las obtenidas con las medias geométricas. También hemos proyectado el origen de coordenadas sobre las log-ratios por pares para tener localizada la media de cada log-ratio por pares y, por comparación, saber qué empresas están por encima y por debajo de la media. Hemos representado dichas medias como puntos cuadrados en verde.



Puesto que el análisis se ha realizado con las log-ratios centradas, añadir más log-ratios por pares no va a crear ningún problema de redundancia. Como se ha dicho, el *biplot* permite representar las log-ratios de cualesquiera de los $D(D-1)/2$ pares posibles de valores x_1 a x_D, teniendo en cuenta que solo definen direcciones informativas las que unen valores cuyos vértices están a una cierta distancia en el *biplot*. De este modo, hemos añadido una log-ratio de rotación del activo no corriente log(*IE/ANC*) y otra que indica hasta qué punto las inversiones en activo no corriente están financiadas por pasivo no corriente log(*PNC/ANC*), log-ratio que ya había aparecido en la ecuación (24). Las empresas situadas en el cuadrante superior derecho son las que tienen una rotación más alta del activo no corriente. Las empresas situadas en el cuadrante superior izquierdo son las que en mayor grado financian su activo no corriente con pasivo no corriente.



6.3. Para saber más. Datos de más de un año

Un ejemplo sencillo de trabajo publicado en el que se explica el *biplot* es el de Saus-Sala et al. (2021).

También es posible tener en el archivo de datos información sobre más de un año. En este caso, cada empresa representa más de una fila del archivo. Es decir, los años deben disponerse unos por debajo de los otros, lo que obligará a editar cualquier archivo Excel plurianual de SABI, pues dicha base de datos dispone los años unos al lado de otros. Para distinguirlas, cada fila suele estar codificada con una variable identificadora con el número de la empresa y algún indicativo temporal. Por ejemplo, el código «10i» se referiría al año inicial de la empresa número 10 y «2f» al año final de la empresa número 2. No se suelen poner más de dos años en el archivo para no sobrecargar el *biplot*. Lo ideal es que todas las empresas tengan datos disponibles en ambos años. La figura muestra un ejemplo de cómo se vería el archivo *vinicolas.xls* en caso de existir diez empresas

y dos años. Nótese que la columna *id* identifica la empresa, la columna *a* el año y la columna *id_a* la mencionada combinación de empresa y año.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	id	a	Id_a	CA	AND	PV	CAT	CL	CM	GAL	MUR	NAV	RIO	VAL	OTRAS	Edad	SA	Exporta	Subvenciones	log_empleados	Genero	Genero_alto	Genero_bajo	Genero_no_divulgado	x1_ANC
2	1	2019	1i	PV	0	1	0	0	0	0	0	0	0	0	0	45	1	0	1	2,77	Bajo	0	1	0	9404,660
3	2	2019	2i	PV	0	1	0	0	0	0	0	0	0	0	0	36	1	1	1	2,89	Bajo	0	1	0	2459,198
4	3	2019	3i	PV	0	1	0	0	0	0	0	0	0	0	0	34	1	0	1	3,18	ND	0	0	1	6088,926
5	4	2019	4i	PV	0	1	0	0	0	0	0	0	0	0	0	23	1	1	1	3,97	Alto	1	0	0	17727,090
6	5	2019	5i	PV	0	1	0	0	0	0	0	0	0	0	0	22	1	1	1	3,04	ND	0	0	1	11897,929
7	6	2019	6i	CM	0	0	0	0	1	0	0	0	0	0	0	34	1	1	1	4,52	Alto	1	0	0	28875,990
8	7	2019	7i	CM	0	0	0	0	1	0	0	0	0	0	0	31	1	0	1	2,56	Bajo	0	1	0	11290,243
9	8	2019	8i	CM	0	0	0	0	1	0	0	0	0	0	0	25	1	0	1	2,48	Bajo	0	1	0	7763,159
10	9	2019	9i	CM	0	0	0	0	1	0	0	0	0	0	0	24	1	1	1	3,30	ND	0	0	1	11482,582
11	10	2019	10i	CM	0	0	0	0	1	0	0	0	0	0	0	19	1	0	1	2,83	ND	0	0	1	6661,268
12	1	2020	1f	PV	0	1	0	0	0	0	0	0	0	0	0	46	1	1	1	3,18	Bajo	0	1	0	9707,005
13	2	2020	2f	PV	0	1	0	0	0	0	0	0	0	0	0	37	1	1	1	2,89	Alto	1	0	0	2425,086

El *biplot* se realiza una sola vez con los datos de ambos años. Al salir cada empresa dos veces en el mismo *biplot*, se puede trazar su evolución de mejoría o empeoramiento en rotación, solvencia, etc. Ejemplos con más de un año de datos se encuentran en Carreras-Simó y Coenders (2020) y Saus-Sala et al. (2023).

7. ¿Es homogéneo el sector? Análisis clúster composicional

7.1. Cuántos grupos, cómo extraerlos y cómo interpretarlos

Muy raramente se puede suponer que una industria representa un único patrón homogéneo de estados financieros. El *análisis clúster*, llamado también *análisis de conglomerados*, es otro método de análisis estadístico multivariante muy popular que tiene como objetivo extraer *grupos*, *conglomerados* o *clústeres* de objetos (es decir, empresas) de tal manera que los objetos del mismo grupo sean lo más similares (homogéneos) posible de acuerdo con las variables de interés. En otras palabras, las empresas del mismo grupo deben tener distancias mutuas bajas. Del mismo modo, las empresas de los diferentes grupos deben ser lo más diferentes posible, es decir, tener grandes distancias mutuas (Dolnicar et al., 2018; Kaufman y Rousseeuw, 1990).

En el contexto del análisis de los estados financieros, el análisis de conglomerados composicional se puede utilizar para identificar grupos de empresas con estructuras de estados financieros similares dentro de un sector (Arimany-Serrat y Coenders, 2025; Arimany-Serrat y Sgorla, 2024; Coenders, 2025; Dao et al., 2024; Jofre-Campuzano y Coenders, 2022; Linares-Mustarós et al., 2018; Molas-Colomer et al., 2024; Saus-Sala et al., 2021; 2023; 2024). A esto a veces se le ha llamado «elaboración de perfiles de rendimiento financiero y dificultades financieras» (Dao et al., 2024; Linares-Mustarós et al., 2018).

El análisis de conglomerados composicional se reduce a realizar un análisis de conglomerados estándar usando las D log-ratios centradas de la ecuación (27) como variables de entrada (Ferrer-Rosell y Coenders, 2018; Martín-Fernández et al., 1998).

Si se utilizan log-ratios centradas como datos, las distancias euclidianas se vuelven iguales a las distancias de Aitchison, que son las de uso estándar en CoDa (Aitchison, 1983; Aitchison et al., 2000). Por lo tanto, la distancia entre las empresas m y l se calcula con la fórmula de la distancia euclídea a partir de las diferencias entre sus respectivas D log-ratios centradas como:

$$d_{ml} = \sqrt{\left(clr_{1m} - clr_{1l}\right)^2 + \left(clr_{2m} - clr_{2l}\right)^2 + \ldots + \left(clr_{Dm} - clr_{Dl}\right)^2} \quad (44)$$

Por lo tanto, con las log-ratios centradas se puede utilizar cualquier método de análisis de conglomerados estándar que maneje distancias euclidianas (Ferrer-Rosell y Coenders, 2018). Esto incluye, entre otros, dos métodos muy populares en el análisis de los estados financieros (Linares-Mustarós et al., 2018): el método de Ward (Ward, 1963) y el método de las *k*-medias (MacQueen, 1967). CoDaPack usa este último.

Para clasificar las empresas en *k* conglomerados, grupos o clústeres, el método de *k*-medias:

- Toma al azar *k* empresas como centros iniciales de clúster.
- Cada una de las empresas restantes se asigna al clúster con el centro más cercano (es decir, el clúster cuyo centro esté a la distancia euclidiana más baja a la empresa).
- Los centros de los clústeres se vuelven a calcular como las medias aritméticas de las log-ratios centradas para las empresas a ellos asignadas. Como los datos están en log-ratios, las medias aritméticas son válidas.

Se vuelve a ejecutar la reasignación de las empresas al clúster con el nuevo centro más cercano, y se vuelve a actualizar el cálculo de los centros de clúster. El proceso continúa hasta que ninguna empresa se mueve de clúster entre una iteración del proceso y la siguiente, momento en que se dice que *el algoritmo de clasificación ha convergido*.

Dado que el resultado final puede depender de las empresas que se tomen como centros iniciales, el proceso completo se repite varias veces con diferentes centros de clúster iniciales elegidos al azar. CoDaPack realiza veinticinco de estas repeticiones. De las veinticinco, solo se presenta al usuario o usuaria la solución con la mayor homogeneidad de clústeres (es decir, la que tiene la suma más baja de las varianzas de las log-ratios centradas dentro de los clústeres).

Rara vez se conoce de antemano el número adecuado de clústeres en los que hay que dividir un sector empresarial. Existen varios índices estadísticos para decidir cuál es la mejor *k* después de hacer clasificaciones con un número razonable de clústeres, por ejemplo, de $k = 2$ a $k = 8$. CoDaPack tiene disponibles *la amplitud media de la silueta* (Kaufman y Rousseeuw, 1990) y el *índice de Caliński-Harabasz* (Caliński y Harabasz, 1974). El valor más alto de ambos índices muestra la mejor *k*, estadísticamente hablando, que a veces difiere de un índice a otro. El número de clústeres también se puede elegir de acuerdo con la interpretabilidad: agregar un clúster tiene sentido si agrega un perfil de los estados financieros sustancialmente diferente, sin que ninguno de los clústeres existentes sea muy pequeño. Se comienza con dos clústeres y se siguen añadiendo clústeres mientras se mantenga la afirmación anterior. Según nuestra propia experiencia, el número ideal de clústeres suele estar entre tres y cinco. En una solución de dos clústeres, cada clúster simplemente tiene las características opuestas al otro en todas las ratios, lo cual es bastante poco interesante. Una solución con más de cinco clústeres tiende a ser muy difícil de interpretar.

Esto nos lleva a hablar de cómo se interpretan los clústeres. La forma más obvia de hacerlo es a partir de las medias geométricas de las ratios clásicas. A partir de los centros composicionales de cada clúster, las ratios financieras clásicas de las ecuaciones (31) a (42) se pueden calcular para representar a una empresa promedio en el clúster. Así, la interpretación de una clasificación

composicional no implica mayores conocimientos contables ni estadísticos que la familiaridad con las propias ratios clásicas, y queda al alcance de cualquier profesional del campo.

Una segunda posibilidad atractiva es interpretar los clústeres a partir de su situación en el *biplot* en referencia a las direcciones definidas por las log-ratios por pares. Así, el *biplot* se puede redibujar con las empresas coloreadas según el clúster al que pertenecen.

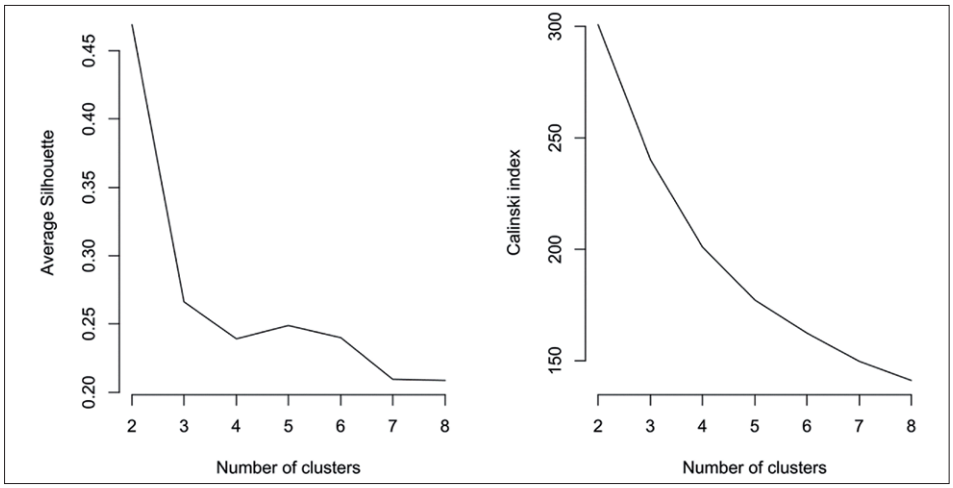
También se pueden obtener ideas útiles para interpretar los clústeres a partir de las características no financieras de las empresas. Los *diagramas de mosaico* se pueden utilizar para relacionar la pertenencia al clúster con características no financieras de la empresa de tipo categórico, como tener o no tener exportaciones o subvenciones (Dolnicar et al., 2018). Los diagramas de caja se pueden utilizar para relacionar la pertenencia al clúster con las características no financieras de la empresa de tipo numérico. Veremos estos gráficos en detalle a partir de los datos del ejemplo.

Después del análisis clúster, los gerentes pueden comparar el perfil financiero de su empresa con el perfil promedio no de todo el sector, sino de un subconjunto de empresas similares, teniendo en cuenta la heterogeneidad del sector. La comparación se puede realizar con respecto al clúster más cercano en el momento de realizar el análisis o con el clúster al que la empresa aspiraría a pertenecer en el futuro. A veces, estos clústeres de empresas pueden identificarse con grupos estratégicos, que compiten sobre la base del margen o la rotación, o con una determinada elección para sus estructuras productivas, reflejadas en su estructura de activos, y de tipos de financiación, reflejadas en su estructura de pasivos.

Según nuestra experiencia, el tamaño de muestra necesario para trabajar cómodamente con análisis clúster es de unas treinta empresas por grupo o clúster.

7.2. Manos a la obra con CoDaPack. Interpretamos los grupos homogéneos dentro del sector con análisis clúster

Abrimos el archivo de datos en formato *.cdp* (menú *File* > *Open Workspace*). Comprobamos que la tabla activa sea *clrx1_2.5* en el desplegable *Tables*. Para realizar un análisis clúster por el método de las *k* medias, entramos en el menú *Statistics* > *Multivariate Analysis* > *Cluster* > *K-means*. El menú calcula internamente las log-ratios centradas, de modo que se deben introducir los valores originales *z.x1_ANC*, z.*x2_AC*, *z.x3_PNC*, *z.x4_PC*, *z.x5_IE* y *z.x6_GE* en el cuadro *Selected*. El menú busca el número de grupos óptimo entre el mínimo y el máximo especificados en *Minimum* y *Maximum*. Recomendamos explorar las soluciones entre un mínimo de dos y un máximo de ocho o diez grupos. El desplegable *Optimality* permite escoger como criterio de decisión entre la amplitud media de la silueta (*Average Silhouette*) y el índice de Caliński-Harabasz (*Calinski Index*). Aunque el programa decide el número de grupos basado en el índice escogido por el usuario o usuaria, muestra gráficamente la evolución de ambos. En este caso, una solución en dos grupos maximiza ambos criterios.



CoDaPack crea una nueva variable categórica llamada *Cluster* al final de la tabla de datos, que identifica cada empresa con su clúster correspondiente según el valor *k* óptimo hallado. En este caso, *k* = 2 y los clústeres están etiquetados como «1» y «2». Clicando dos veces sobre el nombre de la variable podemos cambiarlo, por ejemplo, a *Cluster2*, teniendo en cuenta que más adelante exploraremos soluciones con más clústeres. El menú *File* > *Save as* permite guardar el archivo así ampliado.

Los resultados incluyen, en primer lugar, una tabla con la evolución de la amplitud media de la silueta (`AS.index`) y el índice de Caliński-Harabasz (`CH.index`). En segundo lugar, tenemos el centro composicional para cada uno de los grupos y el número de empresas en cada uno (`Size`). Por ejemplo, el grupo 1 tiene 79 empresas y la media geométrica de su activo no corriente es 0,2018516.

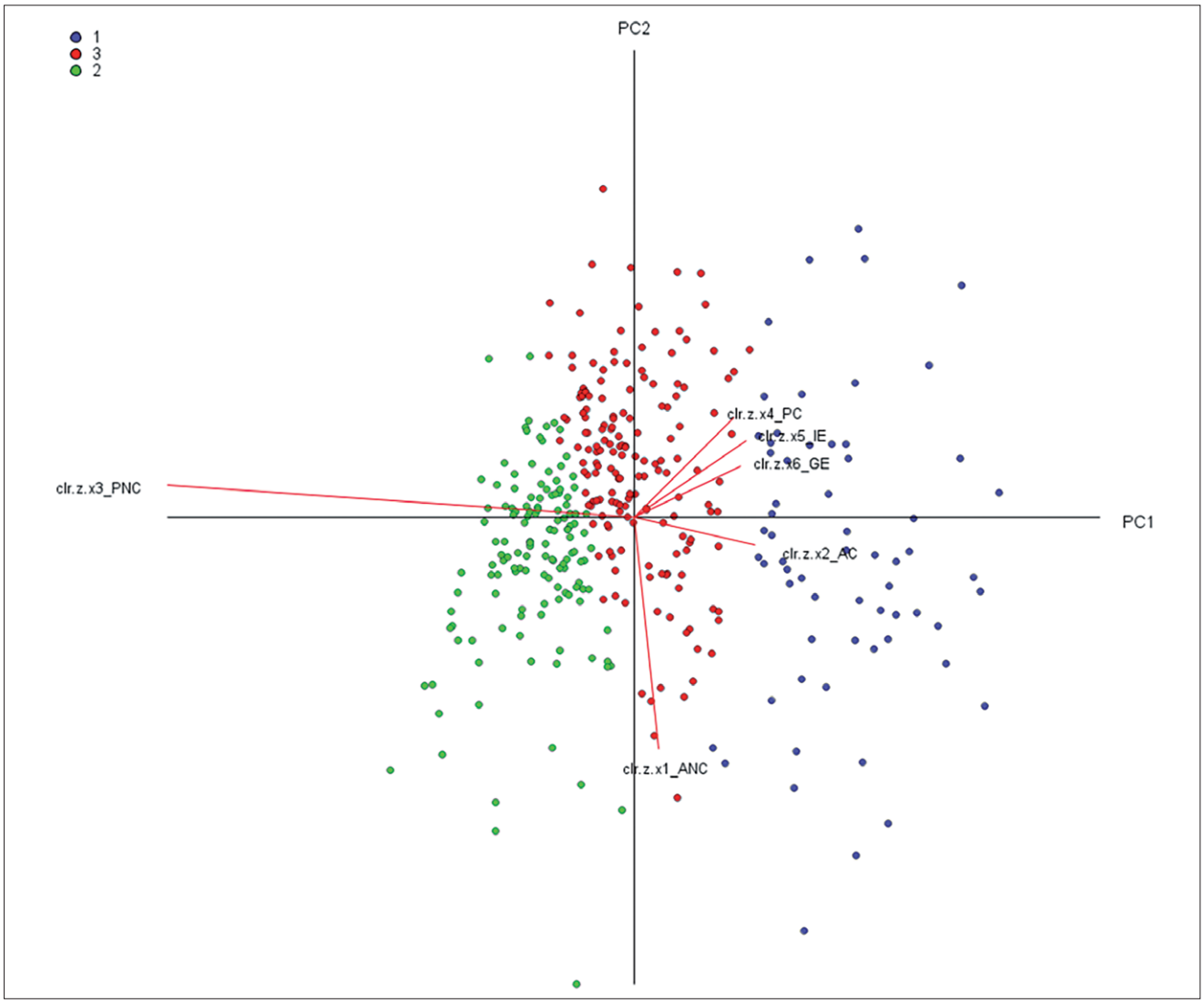
```
K-means
  Centers CH.index  AS.index
1       2 300.7916 0.4691126
2       3 240.0477 0.2661490
3       4 200.9396 0.2390934
4       5 177.2329 0.2490289
5       6 162.4801 0.2397766
6       7 149.7679 0.2096726
7       8 141.1417 0.2086176

Centers:
   z.x1_ANC   z.x2_AC    z.x3_PNC    z.x4_PC   z.x5_IE   z.x6_GE Size Within SS
1 0.2018516 0.2825624 0.002196683 0.09572384 0.2102379 0.2074275   79  252.5017
2 0.2386828 0.2114885 0.081050622 0.09436265 0.1876713 0.1867441  289  607.4087
(between_SS / total_SS =  45.11 %)
```

Las soluciones con solo dos clústeres suelen carecer de interés. Por construcción, en cada valor de x_1 a x_D, uno de los grupos está por arriba y otro simétricamente por debajo.

Aunque la amplitud media de la silueta y el índice de Caliński-Harabasz apunten a una solución determinada, recomendamos siempre explorar otras a partir de tres clústeres para buscar la(s) más interpretable(s). Para obtener una solución con tres clústeres, en el menú *Statistics* > *Multivariate Analysis* > *Cluster* > *K-means* introducimos el valor 3 tanto en la casilla *Minimum* como en *Maximum*. Renombramos la nueva variable creada *Cluster3*.

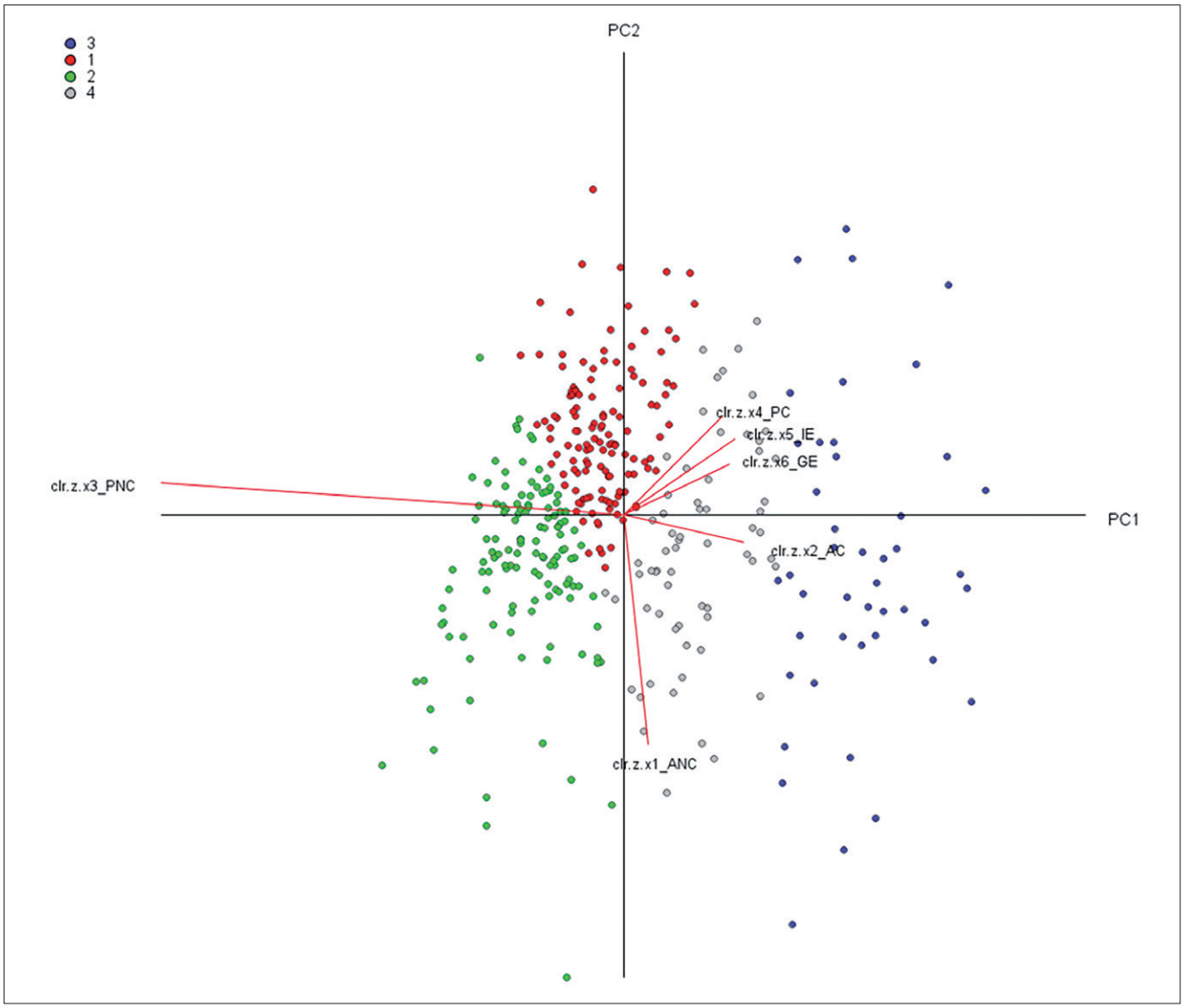
```
K-means
  Centers CH.index AS.index
1       3 240.0477 0.266149
Centers:
   z.x1_ANC   z.x2_AC    z.x3_PNC    z.x4_PC   z.x5_IE  z.x6_GE  Size Within SS
1 0.2050690 0.2963634 0.001470017 0.09608237 0.1989236 0.2020916   63  189.6820
2 0.3234479 0.1740071 0.142591834 0.07425316 0.1404920 0.1452081  136  218.3008
3 0.1713518 0.2340917 0.039786005 0.10721974 0.2290578 0.2184930  169  268.6442
(between_SS / total_SS =  56.81 %)
```

Una primera manera de ver la interpretación de una solución con tres o más clústeres es representarlos sobre el *biplot*. En el menú *Graphs* > *CLR-biplot*, introducimos los valores contables originales *z.x1_ANC*, z.*x2_AC*, *z.x3_PNC*, *z.x4_PC*, *z.x5_IE* y *z.x6_GE* en el cuadro *Selected* y escogemos *Cluster3* en el desplegable *Group by*. Observamos que los clústeres están ordenados respecto al eje horizontal del *biplot* que contenía sobre todo diferencias en la ratio de maduración de la deuda (*PNC*/*PC*), más a largo plazo cuanto más a la izquierda (grupo 2), y más a corto cuanto más a la derecha (grupo 1); el grupo 3 tiene valores intermedios. Clústeres así construidos apenas tendrían diferencias en las ratios que se trazan en dirección aproximadamente vertical en el *biplot*, en especial la rotación del activo corriente (*IE*/*AC*) y la solvencia a corto plazo (*AC*/*PC*).



Desechamos, pues, la solución en tres clústeres y pasamos a cuatro. Obtenemos ahora un grupo 1 básicamente con valores positivos del eje vertical, que se caracterizará por una alta rotación del activo corriente (*IE*/*AC*) y una baja solvencia a corto plazo (*AC*/*PC*). Renombramos la nueva variable creada *Cluster4*.

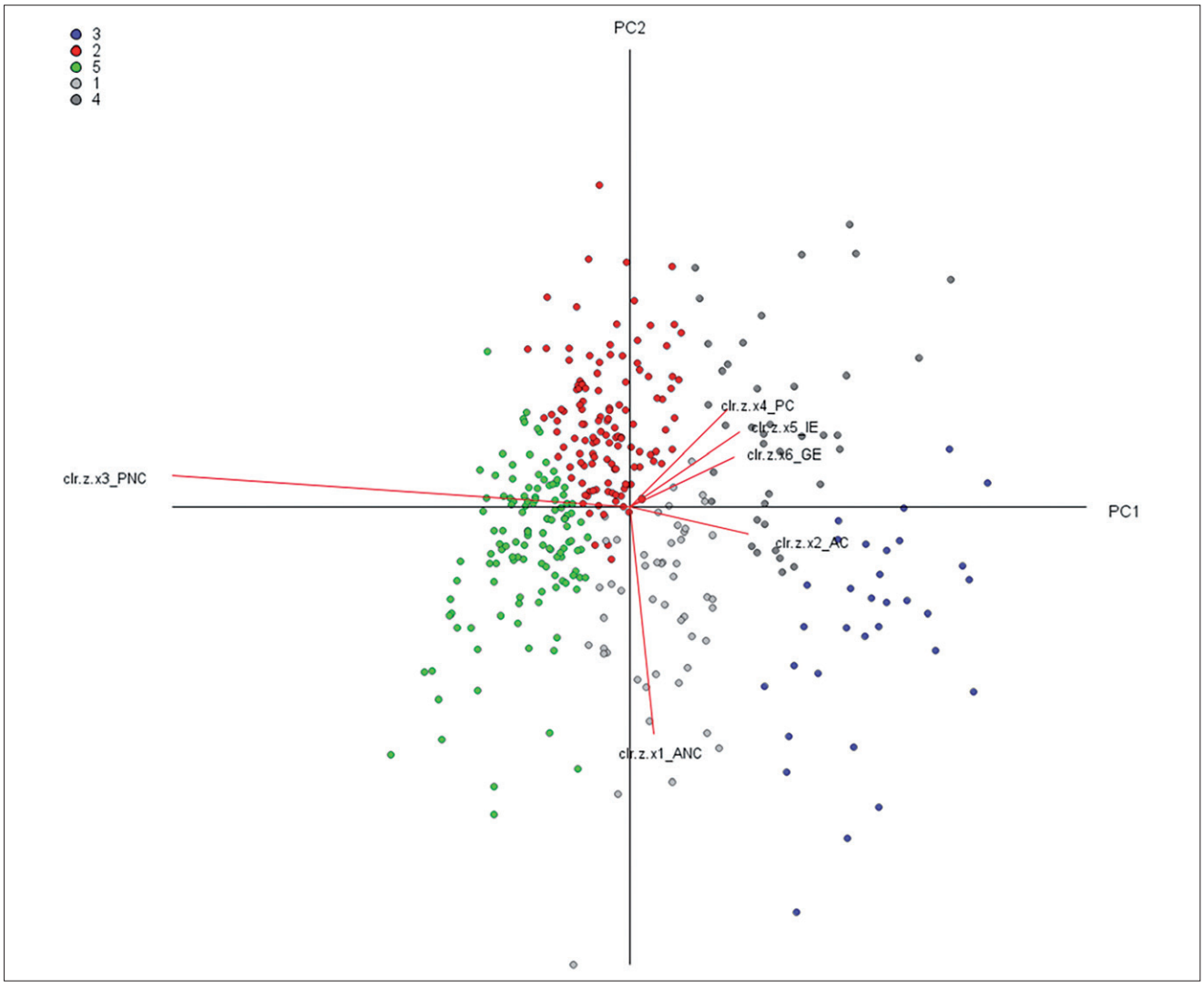
```
K-means
  Centers CH.index  AS.index
1       4 200.9396 0.2390934
Centers:
   z.x1_ANC   z.x2_AC     z.x3_PNC    z.x4_PC   z.x5_IE  z.x6_GE  Size Within SS
1 0.1419681 0.2264742 0.0582559720 0.12792313 0.2276331 0.2177455  131  138.8286
2 0.3366802 0.1711959 0.1422014130 0.07053060 0.1370068 0.1423850  128  205.6570
3 0.2036071 0.3108906 0.0008985565 0.09182314 0.1933598 0.1994207   45  132.3703
4 0.2522436 0.2443173 0.0113969682 0.07506354 0.2133658 0.2036127   64  112.9629
(between_SS / total_SS =  62.35 %)
```



Visto que la solución en cuatro clústeres sería una primera candidata razonable, exploramos la de cinco. Aparece un grupo 1 con valores esencialmente

negativos en el eje vertical y el grupo 3 de la solución anterior se divide en los nuevos grupos 3 y 4, según el eje vertical del *biplot*. Renombramos la variable creada *Cluster5*.

```
K-means
  Centers CH.index  AS.index
1       5 177.2329 0.2490289
Centers:
   z.x1_ANC   z.x2_AC    z.x3_PNC    z.x4_PC   z.x5_IE  z.x6_GE  Size Within SS
1 0.3628707 0.2351056 0.019647802 0.05256703 0.1614136 0.1683953   51  84.03141
2 0.1426633 0.2240385 0.059388324 0.13135532 0.2257386 0.2168159  126 126.93865
3 0.3118003 0.2999236 0.000684434 0.07396699 0.1536642 0.1599605   32  75.58485
4 0.1032250 0.2406803 0.003962629 0.12303878 0.2715189 0.2575744   39  72.15598
5 0.3190783 0.1709441 0.154421909 0.07179092 0.1411738 0.1425910  120 171.80918
(between_SS / total_SS =  66.14 %)
```



La solución en seis clústeres ya conduce a un grupo 4 muy pequeño (19 empresas, solo el 5 % del total de empresas del sector). Esta situación se tiende a evitar, con lo que interrumpimos la búsqueda aquí. En resumen, se añaden grupos adicionales mientras ofrezcan distinciones interesantes y relevantes entre las empresas sin conducir a grupos muy pequeños.

```
K-means
  Centers CH.index  AS.index
1       6 162.4801 0.2397766
Centers:
   z.x1_ANC   z.x2_AC    z.x3_PNC    z.x4_PC    z.x5_IE   z.x6_GE  Size Within SS
1 0.2922115 0.1751010 0.131817035 0.08998769 0.15840349 0.15247928  113 121.8259
2 0.3574686 0.2383750 0.019842286 0.04892037 0.16899243 0.16640127   45  59.7040
3 0.1170714 0.2431739 0.005026823 0.12258975 0.26384472 0.24829346   41  66.0599
4 0.4822047 0.1299226 0.212176108 0.02012052 0.06122609 0.09434992   19  30.8425
5 0.1353552 0.2270592 0.057482345 0.13202309 0.22837107 0.21970917  116 115.5845
6 0.2809257 0.3037497 0.000672350 0.07889483 0.16474901 0.17100833   34  88.8803
(between_SS / total_SS =  69.18 %)
```

En nuestro caso, la solución elegida es la de cinco clústeres. Para interpretar cada uno de los grupos, a partir del centro composicional para cada uno de ellos calculamos las ratios clásicas (tabla 5). Como el análisis se ha hecho sobre las log-ratios centradas, añadir un mayor número de ratios clásicas no conlleva problemas de redundancia. Acostumbra a ser útil identificar los grupos con valores destacadamente altos (azul) o bajos (rojo) de cada ratio.

	Total sector	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Rotación	0,422	0,270	0,616	0,251	0,790	0,288
Rotación del activo corriente	0,854	0,687	1,008	0,512	1,128	0,826
Margen	0,007	−0,043	0,040	−0,041	0,051	−0,010
Apalancamiento	1,408	1,137	2,084	1,139	1,586	1,857
ROA	0,003	−0,012	0,024	−0,010	0,041	−0,003
ROE	0,004	−0,013	0,051	−0,012	0,064	−0,005
Endeudamiento	0,290	0,121	0,520	0,122	0,369	0,462
Endeudamiento a corto plazo	0,208	0,088	0,358	0,121	0,358	0,147
Solvencia a largo plazo	3,449	8,281	1,922	8,194	2,708	2,166
Solvencia a corto plazo	2,378	4,472	1,706	4,055	1,956	2,381
Inmovilización del activo	1,023	1,543	0,637	1,040	0,429	1,867
Maduración de la deuda	0,395	0,374	0,452	0,009	0,032	2,151

Tabla 5. Ratios clásicas medias del sector vitivinícola por clústeres

Por ejemplo, la rotación del grupo 1 se calcula como 0,1614136/(0,3628707 + 0,2351056). El grupo 1 (51 empresas, que son el 14 % de las empresas del sector) se caracteriza por baja rotación, bajo margen, ROA y ROE, bajo endeudamiento, por lo tanto, bajo apalancamiento, y bajo endeudamiento a corto plazo; en otras palabras, alta solvencia a largo y a corto plazo. El grupo 3 (9 % del sector) hasta aquí tiene características similares. La diferencia entre ambos es que el grupo 1 destaca por una alta inmovilización del activo y el grupo 3 por tener casi toda la deuda a corto plazo, según indica la ratio de maduración de la deuda. Cuando el endeudamiento es bajo, tener la deuda a corto plazo no es un problema. El problema principal de ambos grupos es el margen negativo traducido al ROA y al ROE también negativos. A nuestro juicio, esos dos grupos son los peores. Son los que están en el cuadrante inferior derecho del *biplot*.

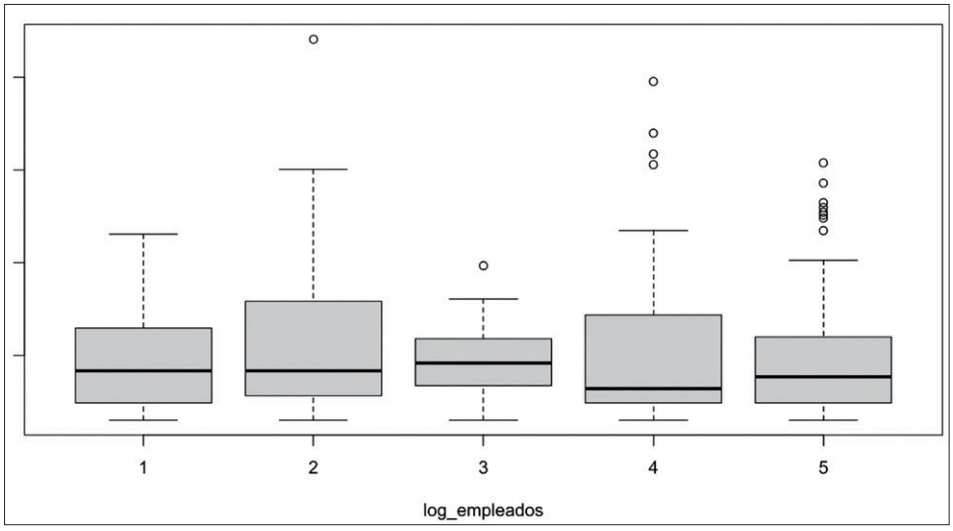
El grupo 2 (34 % del sector) destaca por unos buenos datos de margen, rotación y ROE, acompañados de baja solvencia a largo y a corto plazo y los correspondientes altos endeudamiento y apalancamiento. Aun así, las cifras

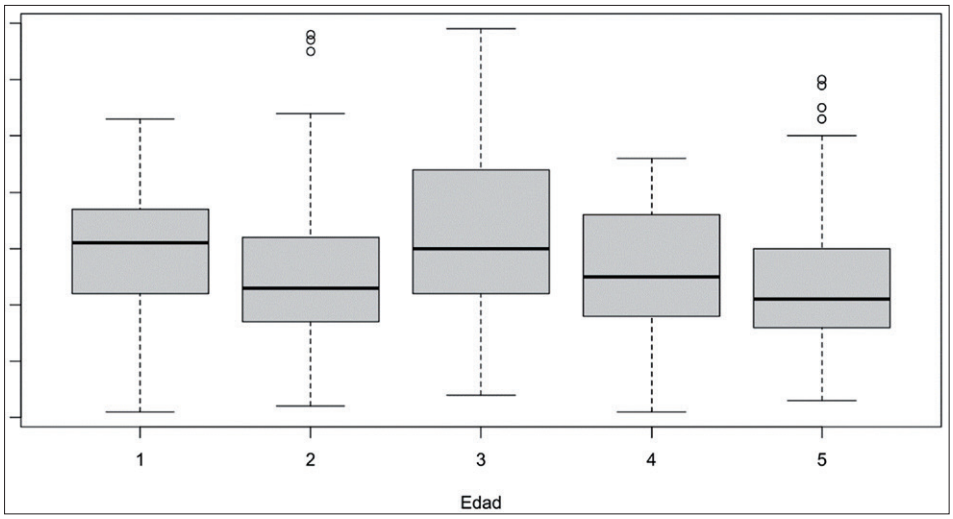
de endeudamiento no son nada preocupantes. La ratio de solvencia a corto plazo (activo corriente sobre pasivo corriente) supera en mucho la unidad, y casi la mitad del activo está financiado con fondos propios según la ratio de endeudamiento. La estructura del activo contiene sobre todo activo corriente. Es el grupo que está en la parte superior central del *biplot*.

El grupo 4 (11 % del sector) se diferencia del grupo 2 por una mayor solvencia a largo plazo y una maduración de la deuda muy a corto plazo. Dada la reducida deuda total, su maduración en el corto plazo no representa un problema. Los márgenes, ROA y ROE son ligeramente mejores. En conjunto es, a nuestro juicio, el mejor de los cuatro clústeres. Es el grupo que está en el cuadrante superior derecho del *biplot*.

El grupo 5 (33 % del sector) tiene baja solvencia a largo plazo y los correspondientes altos endeudamiento y apalancamiento. Aun así, el endeudamiento no es preocupante porque más de la mitad del activo está financiado con fondos propios. Tiene una baja rotación y márgenes, ROA y ROE próximos a cero. La inmovilización del activo y la maduración de la deuda toman valores máximos. El hecho de que los activos no corrientes se financien con pasivos también no corrientes es financieramente ortodoxo. La alta maduración de la deuda reduce aún más el impacto del endeudamiento. Es el grupo que está a la izquierda del *biplot*.

La interpretación de los clústeres se completa poniéndolos en relación con las variables no financieras del archivo. Para variables numéricas empleamos gráficos de caja. En el menú *Graphs* > *Boxplot*, introducimos el logaritmo de los empleados (*log_empleados*) en el cuadro *Selected*. Marcamos *Cluster5* en el desplegable *Group by*. A continuación, hacemos lo mismo con el número de años que la empresa ha estado activa (*Edad*). El mejor grupo, 4, se caracteriza por la mediana del número de empleados más baja, mediana que, en cambio, es máxima para el grupo 3, que era uno de los peores grupos. Parecería que un número de empleados reducido favorece el buen desenvolvimiento financiero. En lo que respecta a la edad de la empresa, los peores grupos, 1 y 3, son los que tienen una mayor mediana de años de actividad.

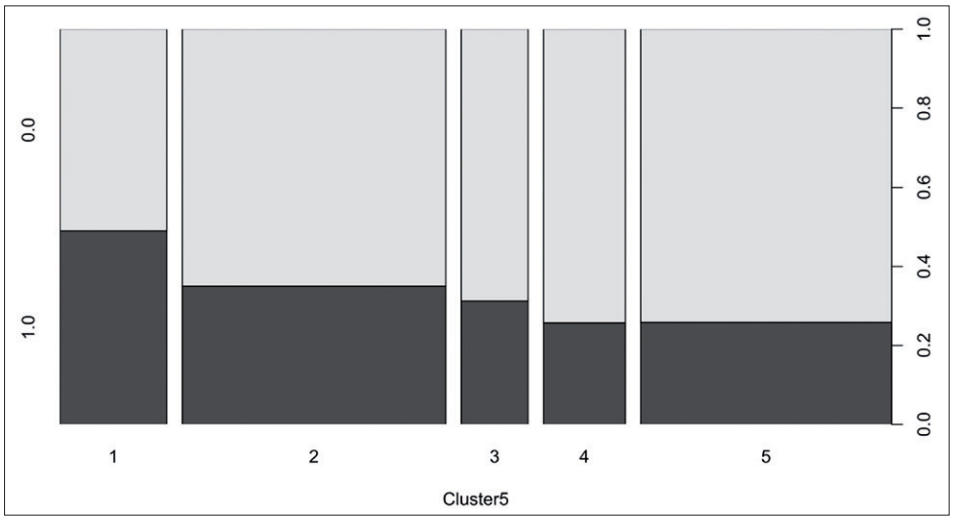




Para relacionar los clústeres con variables categóricas, usamos diagramas de mosaico (menú *Graphs* > *Mosaic plot*). Entramos *Cluster5* y una variable categórica que nos interese, en ese orden, en el cuadro *Selected*. Para la variable *SA_cat* sobre tipo de sociedad («1» SA, «0» SL) obtenemos una tabla con el número de empresas de cada tipo dentro de cada grupo. Por ejemplo, de las 51 empresas del grupo 1, 26 son SL y 25 son SA. En el gráfico propiamente dicho, la anchura de las barras representa el tamaño de cada grupo y la altura la importancia relativa de las SA y las SL dentro de cada grupo. El grupo 1 tiene la máxima proporción de unos (SA) y los grupos 4 y 5, la máxima proporción de ceros (SL).

Mosaic plot

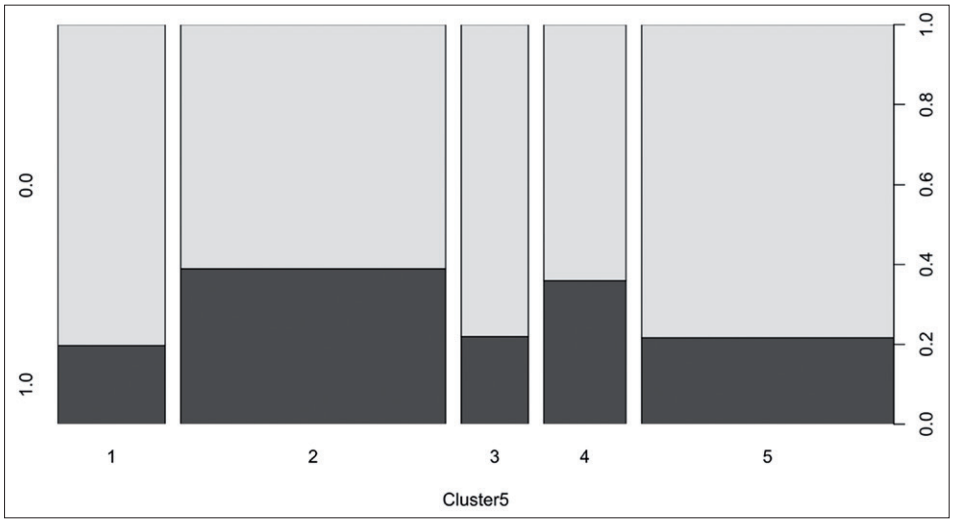
	0.0	1.0	Sum
1	26	25	51
2	82	44	126
3	22	10	32
4	29	10	39
5	89	31	120
Sum	248	120	368



Lo repetimos para otras variables categóricas. En la variable *Exporta_cat*, el código 1 representa a las empresas exportadoras, que predominan en los grupos 2 y 4, que eran los de los márgenes, ROA y ROE más favorables.

Mosaic plot

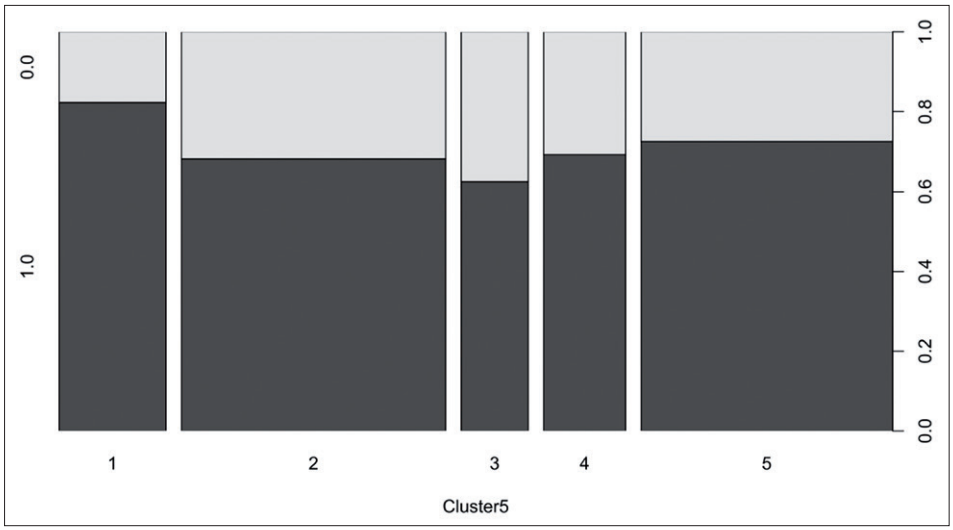
	0.0	1.0	Sum
1	41	10	51
2	77	49	126
3	25	7	32
4	25	14	39
5	94	26	120
Sum	262	106	368



En la variable *Subvenciones_cat*, el código 1 representa a las empresas con subvenciones, que predominan en el grupo 1, que es uno de los de peor desempeño financiero.

Mosaic plot

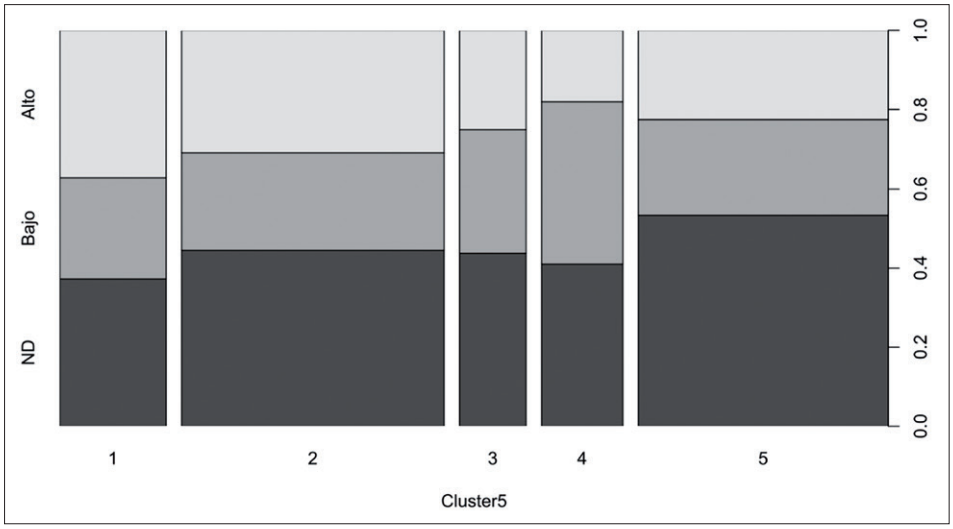
	0.0	1.0	Sum
1	9	42	51
2	40	86	126
3	12	20	32
4	12	27	39
5	33	87	120
Sum	106	262	368



En la variable *Genero_cat* distinguimos las empresas con porcentaje de mujeres superior a la mediana del sector (*Alto*), que predominan en el grupo 1; las empresas con porcentaje de mujeres inferior a la mediana del sector (*Bajo*), que predominan en el grupo 4, y las empresas que no divulgan el dato (*ND*), que predominan en el grupo 5.

Mosaic plot

	Alto	Bajo	ND	Sum
1	19	13	19	51
2	39	31	56	126
3	8	10	14	32
4	7	16	16	39
5	27	29	64	120
Sum	100	99	169	368



Repitiendo la clasificación en cinco clústeres con las ratios clásicas de rotación del activo corriente, margen, solvencia a corto plazo, inmovilización del activo y maduración de la deuda (este análisis no está disponible en CoDaPack),

obtenemos tres grupos construidos en torno a observaciones atípicas (con 0,3 %, 2,4 % y 2,4 % de las empresas del sector respectivamente), un grupo de tamaño moderado (16,6 %) y un grupo que contiene la grandísima mayoría del sector (78,3 %). Como es habitual, el análisis clúster con ratios clásicas da un resultado muy poco atractivo para su interpretación, con grupos de tamaños o muy grandes o muy pequeños, algunos formados solo por observaciones atípicas.

7.3. Para saber más. Datos de más de un año

Un ejemplo sencillo de trabajo publicado en el que se explica la clasificación composicional es el de Saus-Sala et al. (2021).

También es posible tener, en el archivo de datos, información sobre más de un año. En este caso, cada empresa representa más de una fila del archivo de datos. Una de las columnas del archivo de datos, de tipo categórico, tiene que identificar el año. Al contrario que en el *biplot*, no hay límite en el número de años que se pueden poner. Los años pueden ser consecutivos o no según la pregunta de investigación. Se puede estudiar, por ejemplo, el período de los últimos diez años, o los años inmediatamente antes y después de un fenómeno relevante (COVID-19, guerra de Ucrania, etc.). Lo ideal es que todas las empresas tengan datos disponibles todos los años. Para la estructura del archivo, véase la figura del apartado 6.3, donde la columna *a* identifica el año.

La clasificación se realiza una sola vez con los datos de todos los años. A continuación, se puede relacionar la variable donde se ha almacenado la clasificación con la variable que contiene el año por medio del correspondiente diagrama de mosaico, para ver si determinados grupos se encuentran más presentes en unos años u otros. Entrando más en detalle, examinar las empresas una a una permite determinar si han pertenecido al mismo clúster todos los años o han ido migrando de un clúster a otros. Ejemplos de trabajos de análisis clúster con más de un año de datos se encuentran en Arimany-Serrat y Coenders (2025), Arimany-Serrat y Sgorla (2024), Dao et al. (2024) y Saus-Sala et al. (2023; 2024).

Sea con un año o con más de uno, la clasificación composicional no permite estandarizar las variables (es decir, las clr) como a veces se hace en el análisis clúster clásico. Al hacerlo, las distancias euclidianas dejarían de corresponder con las distancias de Aitchison. Sí existe la posibilidad de ponderar las partes y, con ello, definir distancias de Aitchison ponderadas (Dao et al., 2024; Jofre-Campuzano y Coenders, 2022).

También es muy prometedora la clasificación borrosa o *fuzzy clustering*, que permite la existencia de empresas híbridas que pertenezcan en cierto grado a más de un clúster (Molas-Colomer et al., 2024).

8. ¿Existen relaciones con otras variables? La regresión composicional

Hasta ahora nos hemos ocupado de los métodos estadísticos descriptivos, univariantes y multivariantes. Este capítulo está dedicado al *modelado estadístico*, la *inferencia estadística* y los *contrastes de hipótesis*. Para ello, las ratios financieras composicionales actúan como variables en cualquier *modelo estadístico* junto con indicadores no financieros, características de la empresa, características del empresario, estilos de gestión, etc.

Las log-ratios centradas se recomiendan para los análisis estadísticos descriptivos multivariantes (por ejemplo, análisis clúster, *biplot* y análisis en componentes principales, como se han utilizado en los capítulos 6 y 7), pero no para ciertos tipos de modelos estadísticos, para los que son preferibles transformaciones de log-ratios alternativas. Incluso de mayor importancia práctica es el hecho de que las log-ratios centradas no son directamente interpretables financieramente, mientras que en los modelos estadísticos la interpretación de las variables incluidas es una cuestión crucial, lo que convierte un conjunto de $D - 1$ log-ratios por pares en una opción mucho más atractiva. Como se ha indicado en el apartado 4.2, para incluir toda la información en las D partes y evitar la redundancia, las log-ratios por pares deben formar un grafo acíclico conexo. Las log-ratios y_1 a y_5 de las ecuaciones (19) a (23) son, por lo tanto, válidas y se usarán en casi todo el presente capítulo.

En este capítulo nos centramos en los conocidos *modelos de regresión lineal* o *modelos de regresión por mínimos cuadrados ordinarios*. En ellos, las log-ratios por pares o bien pueden usarse como variables que dependan de la información no financiera (apartado 8.1), o bien pueden usarse para explicar, prever o predecir una variable no financiera (apartado 8.3). En cuanto al tamaño de muestra necesario para una regresión, nuestra experiencia indica que suelen bastar entre diez y veinte empresas por cada variable explicativa, con tal que se alcancen al menos entre cincuenta y cien empresas en total. Esta recomendación vale para ambos apartados.

8.1. Ratios como variables dependientes

Primero consideramos el caso en el que las log-ratios desempeñan el papel de *variables dependientes,* también conocidas como *variables explicadas*, *variables endógenas* o *variables de respuesta*. El caso contrario se encuentra en el apartado 8.3.

Una vez que se han calculado las log-ratios adecuadas, se puede ajustar un modelo estadístico con métodos estándar, comenzando con el caso más sencillo, que es la regresión lineal por mínimos cuadrados ordinarios en la que la composición (es decir, las ratios financieras transformadas) se hace dependiente de una o más *variables independientes* no composicionales y no financieras, también conocidas como *variables predictoras*, *variables explicativas* o *variables exógenas*. Los conceptos estadísticos de la regresión composicional se desarrollan en Aitchison (1982); Egozcue et al. (2012) y Tolosana-Delgado y Van den Boogaart (2011). Las aplicaciones a los estados financieros se encuentran en Arimany-Serrat et al. (2023); Coenders (2025); Escaramís y Arbussà (2025) y Mulet-Forteza et al. (2024). Los predictores no solo pueden ser numéricos, sino también cualitativos con dos categorías (es decir, binarios), siempre que las dos categorías estén codificadas numéricamente como «0» y «1». Recordemos que así es como están codificadas las variables cualitativas del archivo *vinicolas.xls*. Esto permite predecir las ratios financieras composicionales contenidas en la composición de los estados financieros a partir de indicadores no financieros y otras características de la empresa o de su gestión. Se aconseja a los lectores que no estén familiarizados con la regresión lineal por mínimos cuadrados ordinarios y con los contrastes de hipótesis estadísticos que recurran a cualquier manual de introducción a la estadística o a la econometría.

Antes de la modelización, es muy útil una representación gráfica que relacione las log-ratios con las variables explicativas no financieras. Es lo que se denomina *análisis exploratorio de datos*. Para variables explicativas numéricas se usan *diagramas bivariantes de dispersión*, llamados también simplemente *gráficos bivariantes* o *gráficos de dispersión*. Para variables explicativas cualitativas se usan gráficos de caja, tal como veremos en los datos del ejemplo. Este paso es crucial para tener una primera idea de las relaciones entre las variables y para anticipar posibles problemas en los datos, como relaciones no lineales y observaciones atípicas extremas.

Sin pérdida de generalidad, supondremos que tenemos un predictor numérico z_1 y un predictor cualitativo z_2 que distingue unas empresas de tipo A de otras empresas de tipo B. Como corresponde, la variable cualitativa z_2 está codificada numéricamente de modo binario como «0» (empresas A) y «1» (empresas B). Se dice que la variable z_2 es un indicador de la categoría B (la codificada con «1»), mientras que la categoría A (codificada como «0») se denomina *categoría de referencia* en la interpretación, tal como enseguida veremos. Las log-ratios por pares $y_1,\ldots, y_{D-1}$ son las variables dependientes en D – 1 ecuaciones de regresión lineal especificadas como

$$
\begin{aligned}
y_1 &= \alpha_1 + \beta_{11} z_1 + \beta_{12} z_2 + \varepsilon_1 \\
y_2 &= \alpha_2 + \beta_{21} z_1 + \beta_{22} z_2 + \varepsilon_2 \\
y_3 &= \alpha_3 + \beta_{31} z_1 + \beta_{32} z_2 + \varepsilon_3 \\
y_4 &= \alpha_4 + \beta_{41} z_1 + \beta_{42} z_2 + \varepsilon_4 \\
y_5 &= \alpha_5 + \beta_{51} z_1 + \beta_{52} z_2 + \varepsilon_5
\end{aligned} \tag{45}
$$

donde y_1 a y_5 son las log-ratios por pares en las ecuaciones (19) a (23), z_1 y z_2 son las variables predictoras, los parámetros $\boldsymbol{\alpha}$ son los términos constantes y los parámetros $\boldsymbol{\beta}$ son los efectos de cada uno de los predictores z sobre cada

una de las log-ratios por pares. Estos efectos se interpretan manteniendo constantes todos los demás predictores de la misma log-ratio. Por ejemplo, si β_{11} es positivo, aumentar la característica z_1 tiende a aumentar la log-ratio por pares y_1, manteniendo el tipo de empresa constante. Si β_{21} es negativo, aumentar la característica z_1 tiende a reducir la log-ratio por pares y_2, manteniendo el tipo de empresa constante. Si β_{32} es positivo, las empresas de tipo B tienden a una mayor log-ratio por pares y_3 comparado con las empresas tipo A (que son la referencia de la comparación), manteniendo z_1 constante. Si β_{42} es negativo, las empresas tipo B tienden a una menor log-ratio por pares y_4 que las empresas tipo A, manteniendo z_1 constante. Los términos $\boldsymbol{\varepsilon}$ contienen los *residuos*, que representan la parte de las log-ratios por pares que los predictores no explican.

Los predictores utilizados deben ser los mismos para todas las log-ratios por pares. Esto es así porque hay que considerar la composición de los estados financieros como una sola variable multivariante con partes interrelacionadas. Por ejemplo, y_1 e y_2 tienen el mismo numerador. No sería concebible que una variable explicativa pertenezca a la ecuación que predice y_1 y no a la ecuación que predice y_2.

No es posible incluir log-ratios financieras en el lado derecho de las ecuaciones de regresión para predecir otra log-ratio financiera. Esto es así porque las ratios que involucran las mismas partes de la composición (es decir, los mismos valores obtenidos de los mismos estados financieros) son propensas a *correlaciones espurias* (es decir, falsas), un hecho que ya fue revelado por el propio Karl Pearson en el momento en que estaba desarrollando el concepto de correlación (Pearson, 1897), que ha sido reconocido desde hace mucho tiempo en la literatura contable y financiera (Lev y Sunder, 1979) y que afecta por igual a las ratios financieras clásicas y a las composicionales.

Se contrastan las siguientes *hipótesis estadísticas* (llamadas *hipótesis nulas*, *hipótesis cero*, o H_0), correspondientes a los parámetros $\boldsymbol{\beta}$ de las ecuaciones de regresión en (45). Primero se realiza un *contraste global* para cada una de las $D - 1$ ecuaciones:

H_0: $\beta_{11}=\beta_{12}=0$ (ninguna de las variables explicativas afecta la rotación del activo corriente)
H_0: $\beta_{21}=\beta_{22}=0$ (ninguna de las variables explicativas afecta el margen)
H_0: $\beta_{31}=\beta_{32}=0$ (ninguna de las variables explicativas afecta la solvencia a corto plazo)
H_0: $\beta_{41}=\beta_{42}=0$ (ninguna de las variables explicativas afecta la inmovilización del activo)
H_0: $\beta_{51}=\beta_{52}=0$ (ninguna de las variables explicativas afecta la maduración de la deuda)

A continuación, se realiza un *contraste individual* para cada coeficiente $\boldsymbol{\beta}$:

H_0: $\beta_{11}=0$ (la variable numérica z_1 no afecta la rotación del activo corriente)
H_0: $\beta_{21}=0$ (z_1 no afecta el margen)
H_0: $\beta_{31}=0$ (z_1 no afecta la solvencia a corto plazo)
H_0: $\beta_{41}=0$ (z_1 no afecta la inmovilización del activo)
H_0: $\beta_{51}=0$ (z_1 no afecta la maduración de la deuda)
H_0: $\beta_{12}=0$ (z_2, es decir, el tipo de empresa, no afecta la rotación del activo corriente)
H_0: $\beta_{22}=0$ (z_2 no afecta el margen)

H_0: β_{32}=0 (z_2 no afecta la solvencia a corto plazo)
H_0: β_{42}=0 (z_2 no afecta la inmovilización del activo)
H_0: β_{52}=0 (z_2 no afecta la maduración de la deuda)

El *valor p* asociado a cada contraste estadístico indica el riesgo que implica rechazar la hipótesis nula. Si dicho riesgo es bajo (por ejemplo, inferior a 0,05 = 5 % o inferior a 0,01 = 1 %), la hipótesis nula puede rechazarse. Si se rechaza la hipótesis nula del contraste global, se llega a la conclusión de que al menos una de las variables explicativas es útil para predecir la log-ratio en cuestión. A continuación, los contrastes individuales indican cuál o cuáles. Si se rechaza la hipótesis nula de un contraste individual, se llega a la conclusión de que el predictor en cuestión afecta a la log-ratio implicada (manteniendo constantes todos los demás predictores), en otras palabras, que el efecto del predictor es *estadísticamente significativo*. Si no se llega a rechazar la hipótesis nula del contraste global, no podemos concluir sobre la significación de ninguna de las variables. Esto lleva a interrumpir el análisis con el resultado que ninguna conclusión es posible. Ni siquiera se realizan los contrastes individuales.

Los *coeficientes de determinación*, *R-cuadradas* o R^2, son medidas de bondad de ajuste que indican los porcentajes de varianza de cada log-ratio por pares explicados por el conjunto de las variables *z*, es decir, su poder predictivo.

Por último, es de sobra conocida la sensibilidad del análisis de regresión al cumplimiento de unos *supuestos estadísticos*. Ninguna regresión está completa sin gráficos de los residuos de cada ecuación para verificar el cumplimiento de dichos supuestos:

- El diagrama de dispersión de los residuos frente a los valores ajustados debe exhibir un patrón lineal y horizontal para el cumplimiento del llamado *supuesto de linealidad*.
- El diagrama de dispersión de la raíz cuadrada de los residuos estandarizados absolutos frente a los valores ajustados debe exhibir un patrón horizontal con dispersión constante para el cumplimiento del llamado *supuesto de homocedasticidad* (llamado también *supuesto de igualdad de varianzas*).
- El *diagrama probabilístico* normal de los residuos debe exhibir un patrón aproximadamente lineal para que se cumpla el llamado *supuesto de normalidad*. Sin embargo, la violación del supuesto de normalidad solo tiene consecuencias graves para muestras pequeñas. Se suelen considerar pequeñas las muestras inferiores a treinta o cincuenta empresas. Para muestras grandes, una propiedad estadística llamada *teorema del límite central* hace que muchos resultados estadísticos, entre ellos los de la regresión, sean robustos ante la falta de normalidad.
- Se utiliza un diagrama de dispersión de los residuos frente al *apalancamiento* para detectar si hay *observaciones atípicas influyentes* (susceptibles de modificar sustancialmente los resultados de la regresión), que, si las hubiera, se encontrarían en las esquinas superior derecha o inferior derecha más allá de una frontera de 0,5 para la llamada *distancia de Cook*. El apalancamiento aquí es un concepto estadístico que no tiene nada que ver con el apalancamiento financiero.

Se aconseja a los lectores que no estén familiarizados con los supuestos del modelo de regresión que recurran a un manual introductorio de estadística o econometría.

8.2. Manos a la obra con CoDaPack. Explicamos las ratios a partir de variables no financieras

Abrimos el archivo de datos en formato *.cdp* (menú *File* > *Open Workspace*). Comprobamos que la tabla activa sea *clrx1_2.5* en el desplegable *Tables*. En este análisis de regresión vamos a tratar la información financiera como dependiente y las variables no financieras como explicativas. En concreto, queremos predecir las cinco log-ratios por pares:

- y_1 = rotación del activo corriente = log(*IE/AC*) = *x5_IE_x2_AC*.
- y_2 = margen = log(*IE/GE*) = *x5_IE_x6_GE*.
- y_3 = solvencia a corto plazo = log(*AC/PC*) = *x2_AC_x4_PC*.
- y_4 = inmovilización del activo = log(*ANC/AC*) = *x1_ANC_x2_AC*.
- y_5 = maduración de la deuda = log(*PNC/PC*) = *x3_PNC_x4_PC*.

Lo haremos a partir del resto de las variables numéricas y binarias del archivo de datos:

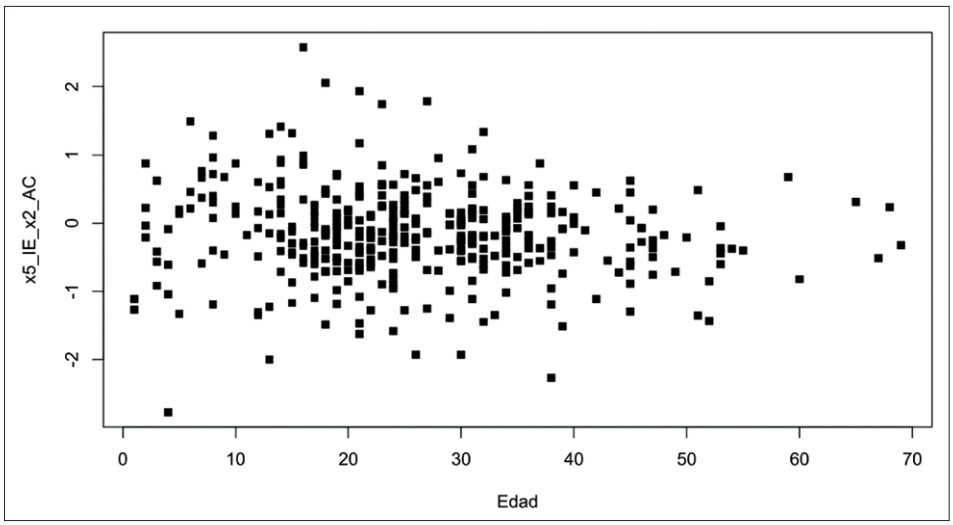
- z_1 = *AND*: variable binaria que indica las empresas sitas en Andalucía.
- z_2 = *PV*: variable binaria que indica las empresas sitas en el País Vasco.
- z_3 = *CAT*: variable binaria que indica las empresas sitas en Cataluña.
- z_4 = *CL*: variable binaria que indica las empresas sitas en Castilla y León.
- z_5 = *CM*: variable binaria que indica las empresas sitas en Castilla-La Mancha.
- z_6 = *GAL*: variable binaria que indica las empresas sitas en Galicia.
- z_7 = *MUR*: variable binaria que indica las empresas sitas en Murcia.
- z_8 = *NAV*: variable binaria que indica las empresas sitas en Navarra.
- z_9 = *RIO*: variable binaria que indica las empresas sitas en La Rioja.
- z_{10} = *VAL*: variable binaria que indica las empresas sitas en la Comunidad Valenciana.
- z_{11} = *Edad*: número de años que la empresa ha estado activa.
- z_{12} = *SA*: variable binaria que indica las sociedades anónimas.
- z_{13} = *Exporta*: variable binaria que indica las empresas que exportan al menos una parte de su producción.
- z_{14}= *Subvenciones*: variable binaria que indica las empresas que reciben alguna subvención.
- z_{15} = *log_empleados*: logaritmo del número de empleados.
- z_{16} = *Genero_bajo*: variable binaria que indica las empresas con porcentaje de mujeres inferior a la mediana del sector.
- z_{17} = *Genero_no_divulgado*: variable binaria que indica las empresas que no divulgan el porcentaje de mujeres.

Hay que recordar que en las variables cualitativas una de las categorías no se incluye en el modelo y actúa como categoría de referencia, contra la que se compararán las demás. En nuestro caso, hemos omitido las categorías contenidas en la variable *OTRAS* en *CA*, y la variable *Genero_alto* en *Genero*. Las categorías de sociedad limitada, empresas no exportadoras y empresas sin subvenciones ya no estaban incluidas como variables en el archivo de datos. El modelo tiene cinco ecuaciones construidas como:

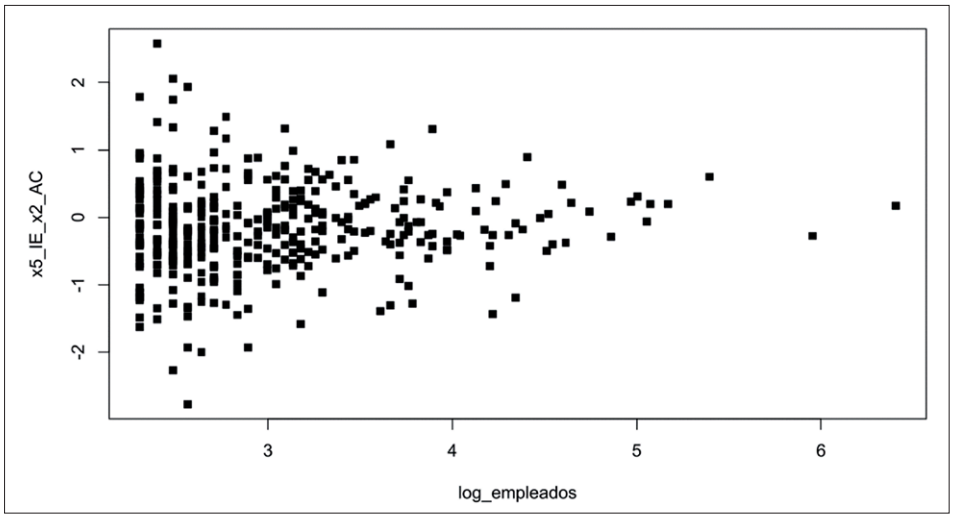
$$
\begin{aligned}
y_1 &= \alpha_1 + \beta_{11} z_1 + \beta_{12} z_2 + \cdots + \beta_{117} z_{17} + \varepsilon_1 \\
y_2 &= \alpha_2 + \beta_{21} z_1 + \beta_{22} z_2 + \cdots + \beta_{217} z_{17} + \varepsilon_2 \\
&\vdots \\
y_5 &= \alpha_5 + \beta_{51} z_1 + \beta_{52} z_2 + \cdots + \beta_{517} z_{17} + \varepsilon_5
\end{aligned} \tag{46}
$$

Se trata de ver si esas cinco log-ratios por pares (variables dependientes, explicadas, endógenas o respuesta) dependen significativamente del resto de las variables (variables independientes, explicativas, predictoras o exógenas). En otras palabras, si la localización de la empresa, su edad, su tipo social, sus exportaciones, sus subvenciones, su tamaño en empleados o su diversidad de género condicionan los resultados financieros. Por ejemplo, ¿podemos afirmar que hay unas comunidades mejores que otras en términos de margen, rotación o solvencia? ¿Las empresas exportadoras tienen mejores márgenes, rotaciones y solvencias? ¿Y las empresas que reciben subvenciones? ¿De qué depende que una empresa tenga una estructura del activo con mayor parte de inmovilizado? ¿Y una maduración de la deuda más a largo plazo? ¿Es la diversidad de género una variable que hay que tener en cuenta para el desempeño financiero?

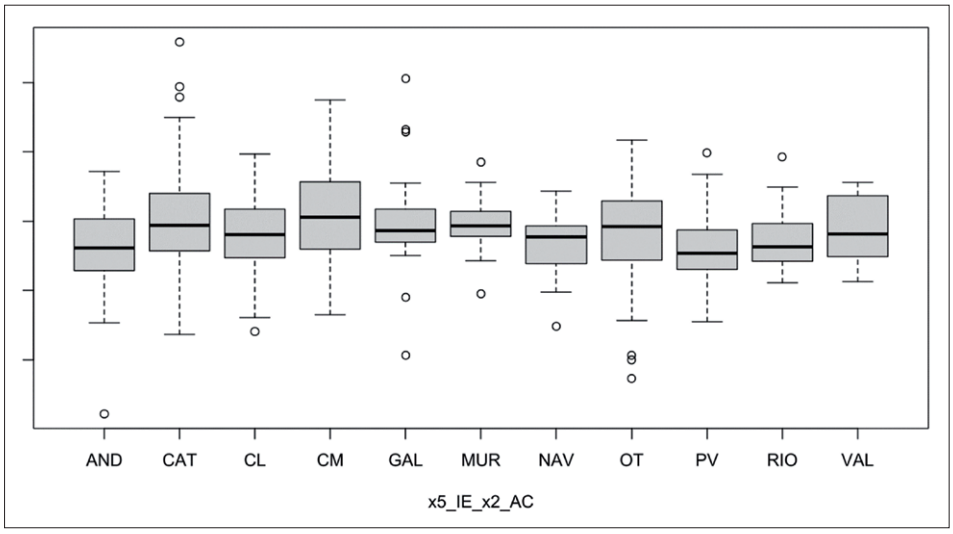
En primer lugar, realizamos gráficos de cada una de las log-ratios dependientes sobre cada una de las variables explicativas. Para ahorrar espacio, mostramos solo los de la primera log-ratio y_1 = rotación del activo corriente = log(*IE*/*AC*) = *x5_IE_x2_AC*. Para las variables numéricas explicativas, entramos en el menú *Graphs* > *Scatterplot*, que produce diagramas de dispersión mediante la introducción de dos variables numéricas en el cuadro *Selected*. La variable introducida en primer lugar aparece en el eje horizontal y corresponde con una variable explicativa en la regresión, en nuestro caso *Edad*. La variable introducida en el segundo lugar es la dependiente en la regresión, en nuestro caso, *x5_IE_x2_AC*.



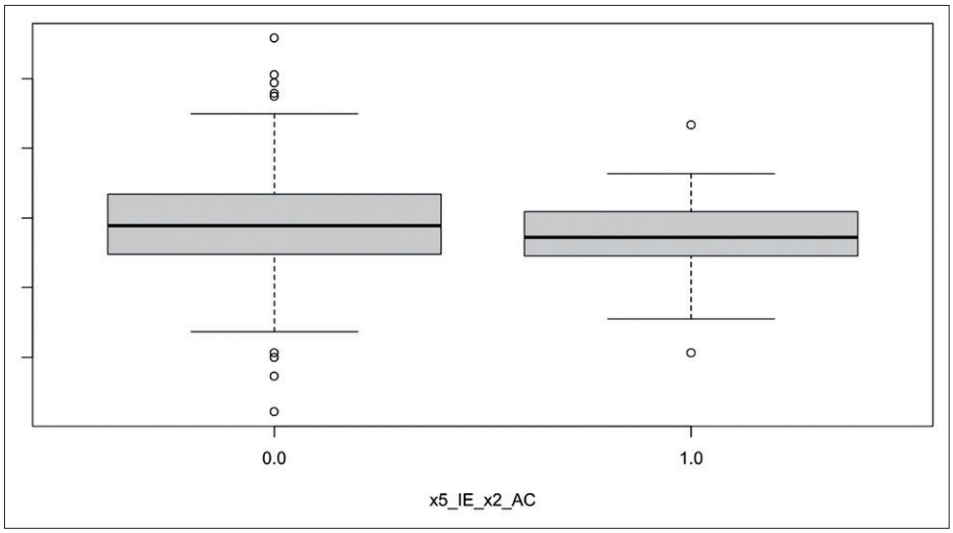
No se observa una relación apreciable ni positiva ni negativa. Tampoco se ven curvaturas u observaciones atípicas extremas, que invalidarían el análisis de regresión. Repetimos el gráfico con *log_empleados*, variable para la cual las conclusiones son las mismas:



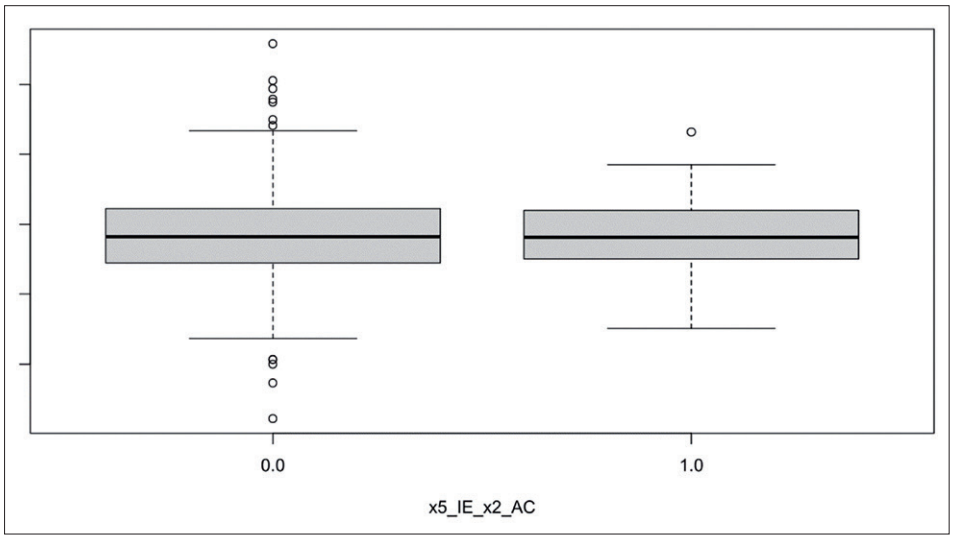
Para las variables explicativas cualitativas, debemos tomar su versión categórica (en naranja sobre el archivo de datos). En el menú *Graphs* > *Boxplot*, introducimos *x5_IE_x2_AC* en el cuadro *Selected*. Marcamos en el desplegable *Group by* una de las variables categóricas. Empezamos por *CA*. Según la mediana (la línea horizontal de trazo grueso en el interior de la caja), destaca por una alta rotación del activo corriente la comunidad *CM*, y por una baja rotación destacan *AND*, *PV* y *RIO*. No hay observaciones atípicas extremas.



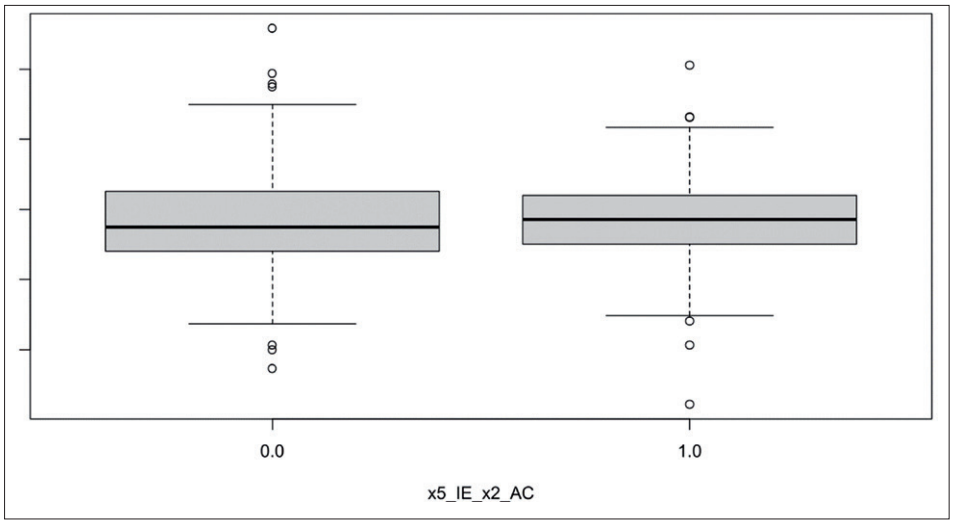
Seguimos con el tipo de sociedad (*SA_cat*). Las sociedades limitadas (categoría 0) tienen una rotación mediana muy ligeramente superior.



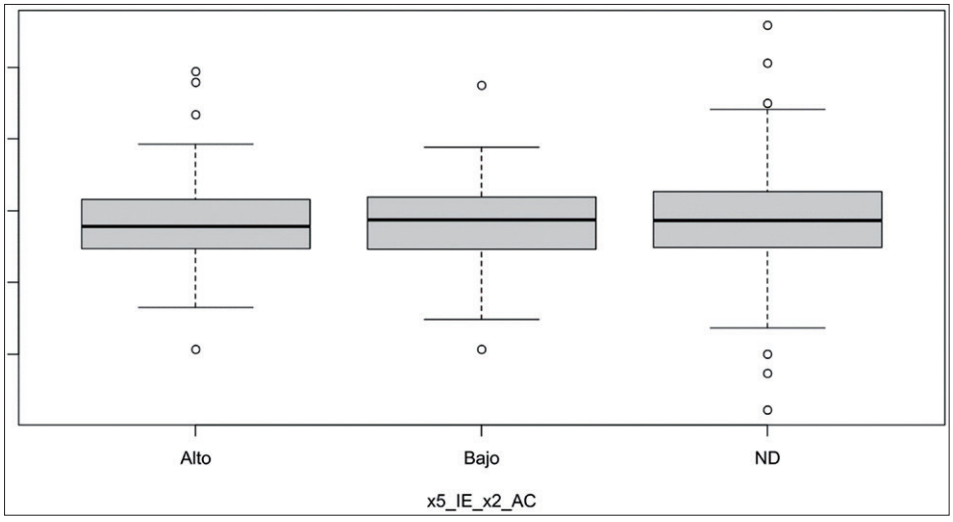
Seguimos con la exportación (*Exporta_cat*). Las empresas exportadoras (categoría 1) y no exportadoras (categoría 0) tienen rotaciones medianas virtualmente idénticas.



Seguimos con las subvenciones (*Subvenciones_cat*). Las empresas con subvenciones (categoría 1) tienen una rotación mediana muy ligeramente superior.



Finalmente, con la variable *Genero*, las empresas con un porcentaje de mujeres empleadas superior a la mediana del sector (categoría *Alto*) tienen una rotación mediana muy ligeramente inferior. En ninguno de los gráficos de caja realizados hay observaciones atípicas extremas.



Pasemos a estimar el modelo de regresión. Como partimos de las log-ratios previamente calculadas, todas las variables tienen la consideración de reales y el menú que se usa es *Statistics* > *Multivariate Analysis* > *Regression* > *X real Y real*. Introducimos las diecisiete variables explicativas en su versión numérica simultáneamente en el cuadro *Explanatory variables*. La variable dependiente *x5_IE_x2_AC* se introduce en el cuadro *Response variable*.

```
X real, Y real
LINEAR REGRESSION
Dependent variable
x5_IE_x2_AC
Explanatory variables
AND, PV, CAT, CL, CM, GAL, MUR, NAV, RIO, VAL, Edad, SA, Exporta, Subvenciones, Ln_
empleados, Genero_bajo, Genero_no_divulgado

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.549784   0.207852  -2.645  0.00854 **
XAND                 -0.220370   0.175109  -1.258  0.20906
XPV                  -0.101598   0.160362  -0.634  0.52679
XCAT                  0.309995   0.129826   2.388  0.01748 *
XCL                   0.063794   0.123766   0.515  0.60657
XCM                   0.301044   0.145162   2.074  0.03882 *
XGAL                  0.270321   0.158155   1.709  0.08830 .
XMUR                  0.148736   0.197470   0.753  0.45183
XNAV                 -0.049392   0.192329  -0.257  0.79747
XRIO                  0.028684   0.161217   0.178  0.85889
XVAL                  0.157415   0.196741   0.800  0.42419
XEdad                -0.004160   0.003202  -1.299  0.19479
XSA                  -0.151037   0.083934  -1.799  0.07281 .
XExporta             -0.051074   0.085011  -0.601  0.54837
XSubvenciones         0.022625   0.081856   0.276  0.78241
Xlog_empleados        0.133203   0.058722   2.268  0.02392 *
XGenero_bajo         -0.024470   0.094834  -0.258  0.79654
XGenero_no_divulgado  0.093197   0.084984   1.097  0.27356
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6586 on 350 degrees of freedom
Multiple R-squared:  0.09065,    Adjusted R-squared:  0.04649
F-statistic: 2.052 on 17 and 350 DF,  p-value: 0.008486
```

La ecuación de previsión estimada (`Estimate`) es:

$$\begin{aligned}&-0{,}550-0{,}220z_1-0{,}102z_2+0{,}310z_3+0{,}064z_4+0{,}301z_5+0{,}270z_6+0{,}149z_7-0{,}049z_8\\&+0{,}029z_9+0{,}157z_{10}-0{,}004z_{11}-0{,}151z_{12}-0{,}051z_{13}+0{,}022z_{14}+0{,}133z_{15}-0{,}024z_{16}+0{,}093z_{17}\end{aligned}\tag{47}$$

Por ejemplo, la previsión para la log-ratio por pares de rotación del activo circulante log(*IE/AC*) para una empresa de Andalucía, que lleva diez años en el sector, que es sociedad limitada, exportadora, sin subvenciones con cien empleados y que no divulga el género de sus empleados es:

$$-0{,}550-0{,}220\times 1-0{,}004\times 10-0{,}051\times 1+0{,}133\log(100)+0{,}093\times 1=-0{,}156\tag{48}$$

Para obtener la previsión de la ratio clásica *IE/AC* de rotación del activo circulante (sin logaritmos), simplemente calculamos la exponencial del resultado y obtenemos una previsión cuyo signo es siempre positivo:

$$e^{-0{,}156}=0{,}856\tag{49}$$

Nótese que no se introduce ningún término en la ecuación cuando la empresa pertenece a la categoría de referencia de alguna variable cualitativa, en este caso las categorías de sociedad limitada y ausencia de subvenciones.

La hipótesis nula del contraste global es:

$$H_0: \beta_{11} = \beta_{12} = \ldots = \beta_{117} = 0$$

y con riesgo $\alpha = 5\,\%$ se rechaza con un valor p de 0,008486 < 0,05 (`p-value`). Concluimos que por lo menos una de las variables de la ecuación se relaciona con la rotación del activo corriente, con lo que merece la pena proseguir el análisis de los contrastes individuales para ver cuál o cuáles. El porcentaje de varianza R^2 de la rotación del activo corriente explicado por el conjunto de las variables es 9,065 % (`Multiple R-squared`).

La primera línea de la tabla de coeficientes contiene el término *constante*, cuya significación no suele considerarse relevante. Los valores p de las variables *CAT*, *CM* y *log_empleados* (`Pr(>|t|)`) son inferiores a 0,05 (0,01748, 0,03882 y 0,02392, respectivamente). Para una más rápida identificación de las variables significativas, la tabla de coeficientes marca con un asterisco los valores p inferiores a 0,05, con un segundo asterisco los inferiores a 0,01 y con un tercer asterisco los inferiores a 0,001. Los resultados permiten rechazar las siguientes hipótesis nulas con riesgo $\alpha = 5\,\%$:

$$H_0: \beta_{13} = 0$$
$$H_0: \beta_{15} = 0$$
$$H_0: \beta_{115} = 0$$

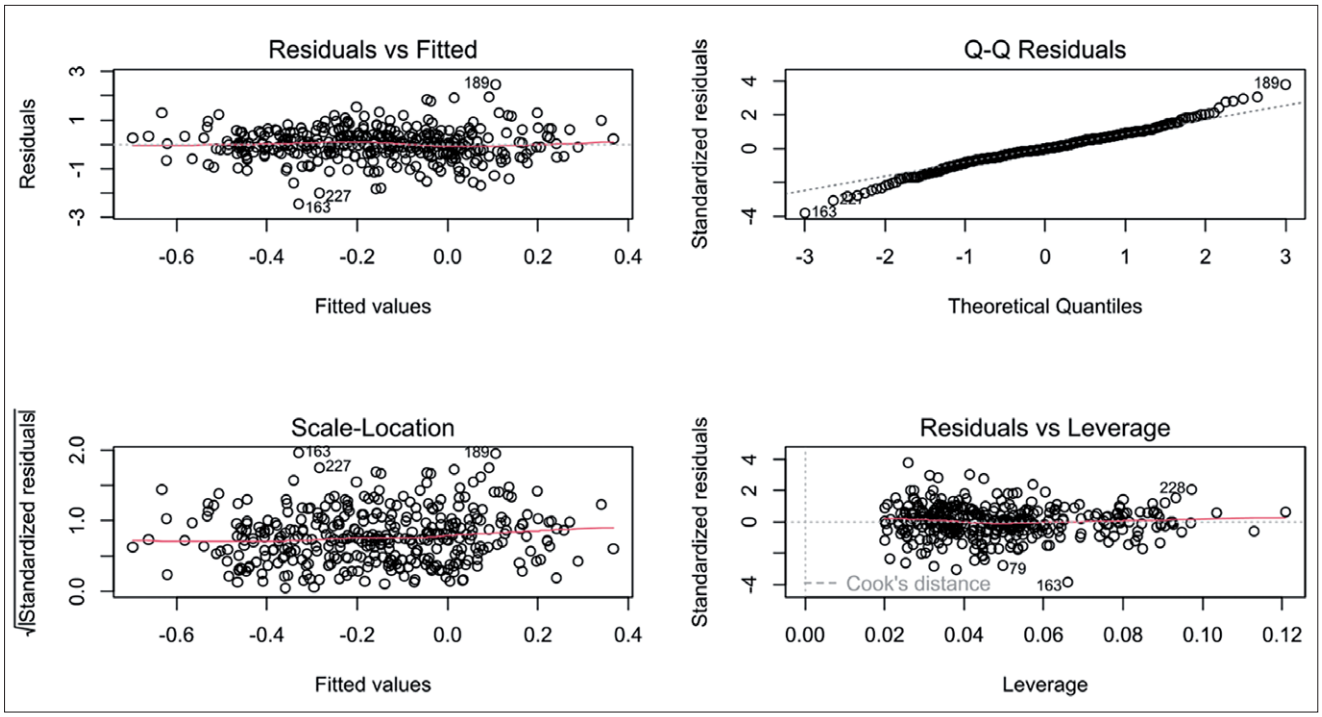
Según los signos de los coeficientes para las hipótesis rechazadas:

- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en Cataluña (*CAT*) tienen mayor rotación del activo corriente (signo positivo del coeficiente en `Estimate`), manteniendo constantes la edad de la empresa, su tipo social, la exportación o no de su producción, la recepción o no de subvenciones, el número de empleados y su diversidad de género.
- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en Castilla-La Mancha (*CM*) tienen mayor rotación del activo corriente (signo positivo del coeficiente en `Estimate`), manteniendo constantes la edad de la empresa, su tipo social, la exportación o no de su producción, la recepción o no de subvenciones, el número de empleados y su diversidad de género.
- Si aumenta el número de empleados (*log_empleados*), tiende a aumentar la rotación del activo corriente (signo positivo del coeficiente en `Estimate`), manteniendo constantes la comunidad autónoma, la edad de la empresa, su tipo social, la exportación o no de su producción, la recepción o no de subvenciones y la diversidad de género.

No podemos afirmar que exista relación entre la rotación del activo corriente y la edad de la empresa, su tipo social, la exportación o no de su producción, la recepción o no de subvenciones y la diversidad de género.

CoDaPack muestra un conjunto de gráficos de los residuos para verificar los supuestos del modelo de regresión lineal.

- El diagrama de dispersión de los residuos frente a los valores ajustados (*Residuals vs fitted*) exhibe un patrón lineal y horizontal, lo que muestra que el supuesto de linealidad se cumple al menos aproximadamente.
- El diagrama de dispersión de la raíz cuadrada de los residuos estandarizados absolutos frente a los valores ajustados (*Scale-Location*) exhibe un patrón horizontal con dispersión constante, lo que muestra que el supuesto de homocedasticidad (es decir, igualdad de varianzas) se cumple al menos aproximadamente.
- El diagrama probabilístico normal (*Q-Q Residuals*) de los residuos exhibe un patrón aproximadamente lineal, lo que muestra que se cumple el supuesto de normalidad. Sin embargo, y como hemos dicho anteriormente, la violación del supuesto de normalidad solo tiene consecuencias graves para muestras pequeñas, por lo que, en caso de haberse incumplido, no nos habría preocupado. Se suelen considerar pequeñas las muestras inferiores a treinta o cincuenta empresas; en nuestro caso tenemos 368.
- Se utiliza un diagrama de dispersión de los residuos frente al apalancamiento (*Residuals vs Leverage*) para detectar si hay observaciones atípicas influyentes (susceptibles de modificar sustancialmente los resultados de la regresión). Como hemos dicho anteriormente, si las hubiera, se encontrarían en las esquinas superior derecha o inferior derecha más allá de una frontera de 0,5 para la llamada *distancia de Cook*. Cuando no hay ninguna distancia de Cook próxima a 0,5, puede ocurrir que dichas fronteras queden fuera del área del gráfico y no se vean, como es el caso aquí. Concluimos, pues, que no hay ninguna observación atípica influyente.



Repetimos el análisis con la siguiente log-ratio por pares y_2 = margen = log(*IE/GE*) = *x5_IE_x6_GE*.

La hipótesis nula del contraste global es:

$$H_0: \beta_{21} = \beta_{22} = \ldots = \beta_{217} = 0$$

y se rechaza con un valor p de $1{,}181 \times 10^{-6} = 0{,}000001181 < 0{,}05$ (`p-value`). El porcentaje de varianza del margen explicado por el conjunto de las variables predictoras es 15,38 % (`Multiple R-squared`).

Los valores p de las variables *CAT*, *CL*, *GAL*, *RIO*, *Subvenciones* y *log_empleados* (`Pr(>|t|)`) son inferiores a 0,05 (0,02237, 0,00000219, 0,00978, 0,01161, 0,00446 y 0,00510, respectivamente), lo que permite rechazar las siguientes hipótesis con riesgo $\alpha = 5\,\%$:

$$H_0: \beta_{23} = 0$$
$$H_0: \beta_{24} = 0$$
$$H_0: \beta_{26} = 0$$
$$H_0: \beta_{29} = 0$$
$$H_0: \beta_{214} = 0$$
$$H_0: \beta_{215} = 0$$

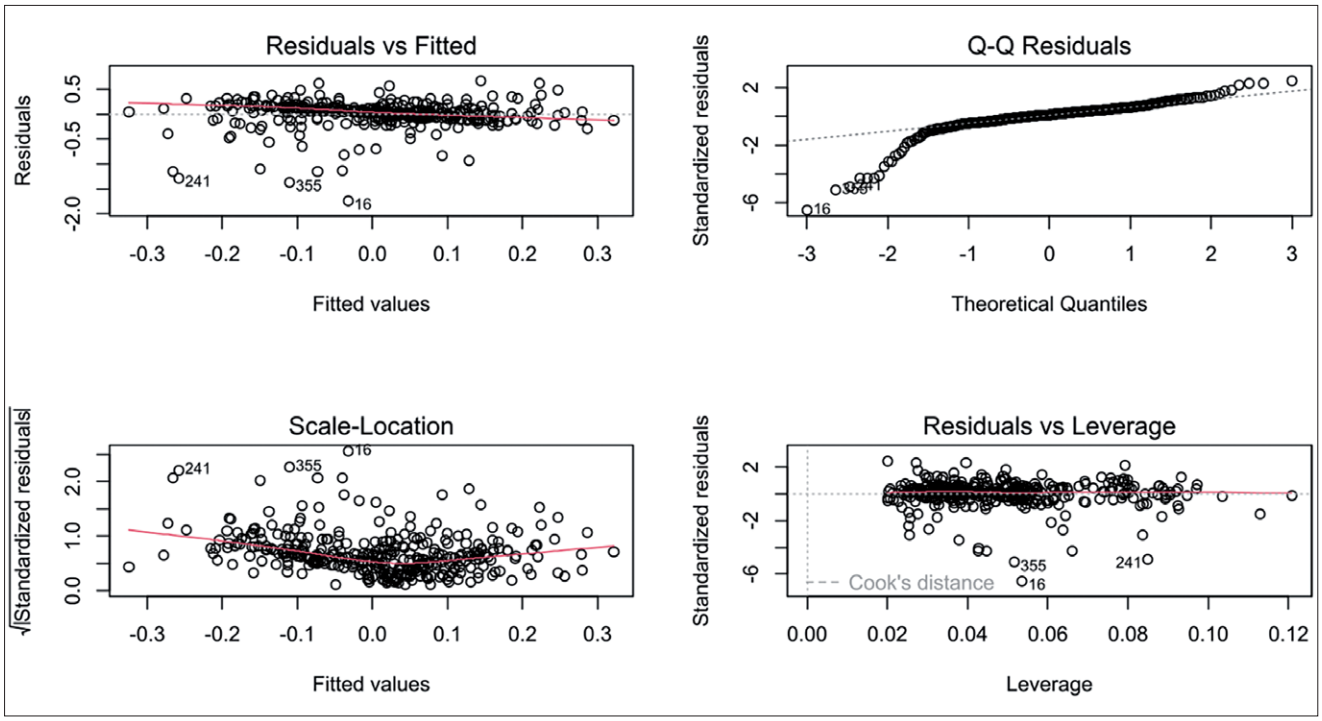
Según los signos de los coeficientes para las hipótesis rechazadas:

- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en Cataluña (*CAT*) tienen mayor margen (signo positivo del coeficiente en `Estimate`), manteniendo constantes la edad de la empresa, su tipo social, la exportación o no de su producción, la recepción o no de subvenciones, el número de empleados y su diversidad de género.
- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en Castilla y León (*CL*) tienen mayor margen (signo positivo del coeficiente), manteniendo constantes las mismas restantes variables de la ecuación.
- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en Galicia (*GAL*) tienen mayor margen (signo positivo del coeficiente), manteniendo constantes las restantes variables.
- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en La Rioja (*RIO*) tienen mayor margen (signo positivo del coeficiente), manteniendo constantes las restantes variables.
- Comparado con las empresas sin subvenciones, las empresas que sí las reciben (*Subvenciones*) tienen mayor margen (signo positivo del coeficiente), manteniendo constantes las restantes variables.
- Si aumenta el número de empleados (*log_empleados*), tiende a aumentar el margen (signo positivo del coeficiente), manteniendo constantes las restantes variables.

```
X real, Y real
LINEAR REGRESSION
Dependent variable
x5_IE_x6_GE
Explanatory variables
AND, PV, CAT, CL, CM, GAL, MUR, NAV, RIO, VAL, Edad, SA, Exporta, Subvenciones, Ln_
empleados, Genero_bajo, Genero_no_divulgado
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.4048562  0.0862180  -4.696 3.82e-06 ***
XAND                  0.0543277  0.0726360   0.748  0.45500
XPV                   0.1163250  0.0665190   1.749  0.08121 .
XCAT                  0.1235523  0.0538524   2.294  0.02237 *
XCL                   0.2472409  0.0513389   4.816 2.19e-06 ***
XCM                   0.0664297  0.0602141   1.103  0.27069
XGAL                  0.1704317  0.0656035   2.598  0.00978 **
XMUR                  0.1590895  0.0819117   1.942  0.05291 .
XNAV                 -0.0718720  0.0797790  -0.901  0.36827
XRIO                  0.1696660  0.0668736   2.537  0.01161 *
XVAL                  0.0980541  0.0816091   1.202  0.23037
XEdad                -0.0004519  0.0013283  -0.340  0.73391
XSA                   0.0345389  0.0348163   0.992  0.32187
XExporta              0.0115660  0.0352629   0.328  0.74311
XSubvenciones         0.0971902  0.0339541   2.862  0.00446 **
Xlog_empleados        0.0686591  0.0243582   2.819  0.00510 **
XGenero_bajo         -0.0256641  0.0393375  -0.652  0.51457
XGenero_no_divulgado  0.0388421  0.0352520   1.102  0.27129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2732 on 350 degrees of freedom
Multiple R-squared:  0.1538,     Adjusted R-squared:  0.1127
F-statistic: 3.743 on 17 and 350 DF,  p-value: 1.181e-06
```

En lo que respecta a los gráficos de los residuos:

- El diagrama *Residuals vs fitted* exhibe un patrón lineal, lo que muestra que el supuesto de linealidad se cumple aproximadamente.
- El diagrama *Scale-Location* exhibe un patrón horizontal con dispersión constante, lo que muestra que el supuesto de homocedasticidad se cumple aproximadamente.
- El diagrama *Q-Q Residuals* exhibe un patrón no lineal, lo que muestra que se viola el supuesto de normalidad. Sin embargo, la violación del supuesto de normalidad solo tiene consecuencias graves para muestras pequeñas, por lo que no nos preocupa.
- El diagrama *Residuals vs Leverage* no muestra distancias de Cook más allá de una frontera de 0,5.



Repetimos el análisis con la siguiente log-ratio por pares: y_3 = solvencia a corto plazo = log(*AC*/*PC*) = *x2_AC_x4_PC*.

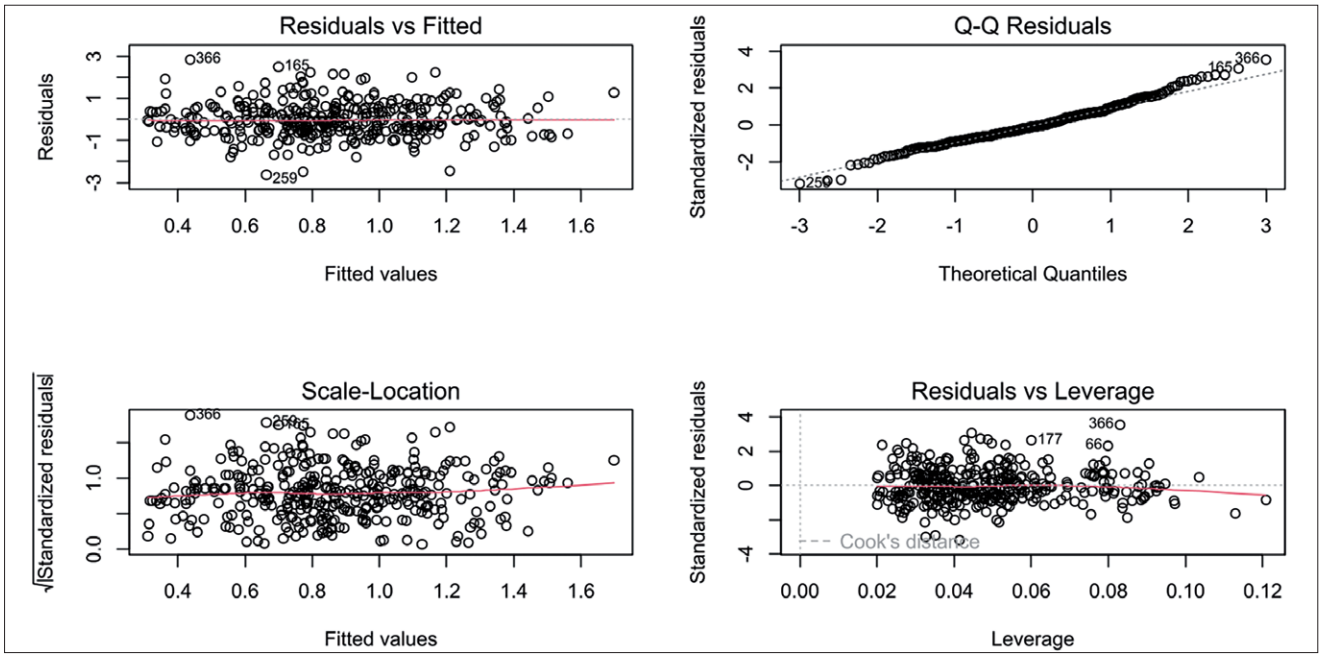
La hipótesis nula del contraste global se rechaza con un valor *p* de 0,003765 < 0,05. El porcentaje de varianza de la solvencia a corto plazo explicado por el conjunto de las variables predictoras es 9,74 %. Los valores *p* de las variables *Edad*, *SA* y *Exporta* son inferiores a 0,05, lo que permite rechazar sus hipótesis nulas respectivas con riesgo α = 5 %. Según los signos de los coeficientes para las hipótesis rechazadas:

- Si aumenta el número de años de actividad (*Edad*), tiende a aumentar solvencia a corto plazo (signo positivo del coeficiente), manteniendo constantes las restantes variables.
- Comparado con las sociedades limitadas (categoría de referencia), las sociedades anónimas (*SA*) tienen mayor solvencia a corto plazo (signo positivo del coeficiente), manteniendo constantes las restantes variables.
- Comparado con las empresas que no exportan, las empresas que sí lo hacen tienen menor solvencia a corto plazo (signo negativo del coeficiente), manteniendo constantes las restantes variables.

En lo que respecta a los gráficos de los residuos:

- El diagrama *Residuals vs fitted* exhibe un patrón lineal.
- El diagrama *Scale-Location* exhibe un patrón horizontal con dispersión constante.
- El diagrama *Q-Q Residuals* exhibe un patrón lineal.
- El diagrama *Residuals vs Leverage* no muestra distancias de Cook más allá de una frontera de 0,5.

```
X real, Y real
LINEAR REGRESSION
Dependent variable
x2_AC_x4_PC
Explanatory variables
AND, PV, CAT, CL, CM, GAL, MUR, NAV, RIO, VAL, Edad, SA, Exporta, Subvenciones, Ln_empleados,
Genero_bajo, Genero_no_divulgado
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.612883   0.265990   2.304  0.02180 *
XAND                 -0.033908   0.224089  -0.151  0.87981
XPV                   0.197704   0.205217   0.963  0.33602
XCAT                 -0.020736   0.166140  -0.125  0.90074
XCL                   0.159144   0.158385   1.005  0.31569
XCM                  -0.131792   0.185766  -0.709  0.47852
XGAL                 -0.099168   0.202393  -0.490  0.62446
XMUR                 -0.097449   0.252705  -0.386  0.70001
XNAV                 -0.102474   0.246125  -0.416  0.67741
XRIO                 -0.099340   0.206311  -0.482  0.63046
XVAL                 -0.103715   0.251771  -0.412  0.68063
XEdad                 0.012978   0.004098   3.167  0.00168 **
XSA                   0.238936   0.107411   2.224  0.02675 *
XExporta             -0.225698   0.108789  -2.075  0.03875 *
XSubvenciones         0.131504   0.104751   1.255  0.21018
Xlog_empleados       -0.061412   0.075147  -0.817  0.41436
XGenero_bajo          0.103343   0.121360   0.852  0.39505
XGenero_no_divulgado -0.062532   0.108755  -0.575  0.56568
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8428 on 350 degrees of freedom
Multiple R-squared:  0.09744,    Adjusted R-squared:  0.0536
F-statistic: 2.223 on 17 and 350 DF,  p-value: 0.003765
```



Repetimos el análisis con la siguiente log-ratio por pares y_4 = inmovilización del activo = log(*ANC*/*AC*) = *x1_ANC_x2_AC*.

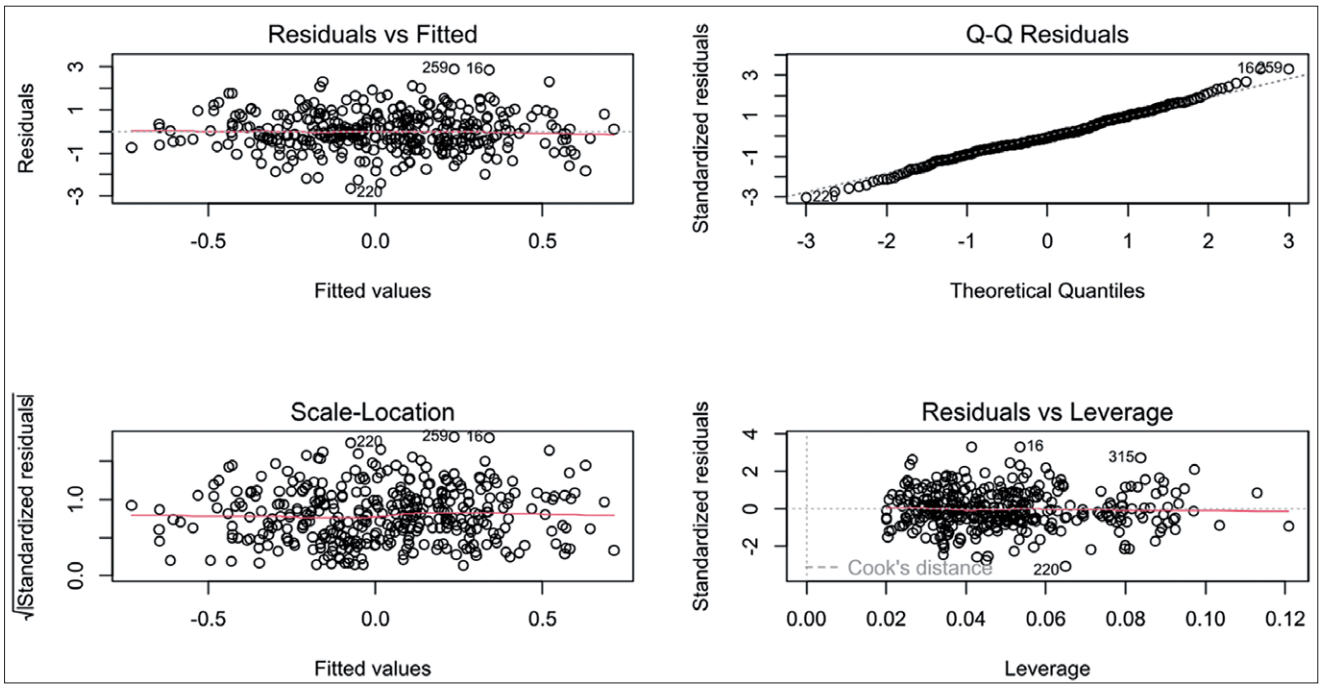
La hipótesis nula del contraste global se rechaza con un valor p de 0,008535 < 0,05. El porcentaje de varianza de la inmovilización del activo explicado por el conjunto de las variables predictoras es de 9,06 %. Los valores p de las variables *CAT*, *CM*, *MUR*, *RIO* y *Exporta* son inferiores a 0,05, lo que permite rechazar sus hipótesis nulas respectivas con $\alpha = 5\,\%$. Según los signos de los coeficientes para las hipótesis rechazadas:

- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en Cataluña (*CAT*) tienen menor inmovilización del activo (signo negativo del coeficiente), manteniendo constantes las restantes variables.
- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en Castilla-La Mancha (*CM*) tienen menor inmovilización del activo (signo negativo del coeficiente), manteniendo constantes las restantes variables.
- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en Murcia (*MUR*) tienen menor inmovilización del activo (signo negativo del coeficiente), manteniendo constantes las restantes variables.
- Comparado con las comunidades en la categoría de referencia (*OTRAS*), las empresas en La Rioja (*RIO*) tienen menor inmovilización del activo (signo negativo del coeficiente), manteniendo constantes las restantes variables.
- Comparado con las empresas que no exportan, las empresas que sí lo hacen tienen menor inmovilización del activo (signo negativo del coeficiente), manteniendo constantes las restantes variables.

En lo que respecta a los gráficos de los residuos:

- El diagrama *Residuals vs fitted* exhibe un patrón lineal.
- El diagrama *Scale-Location* exhibe un patrón horizontal con dispersión constante.
- El diagrama *Q-Q Residuals* exhibe un patrón lineal.
- El diagrama *Residuals vs Leverage* no muestra distancias de Cook más allá de una frontera de 0,5.

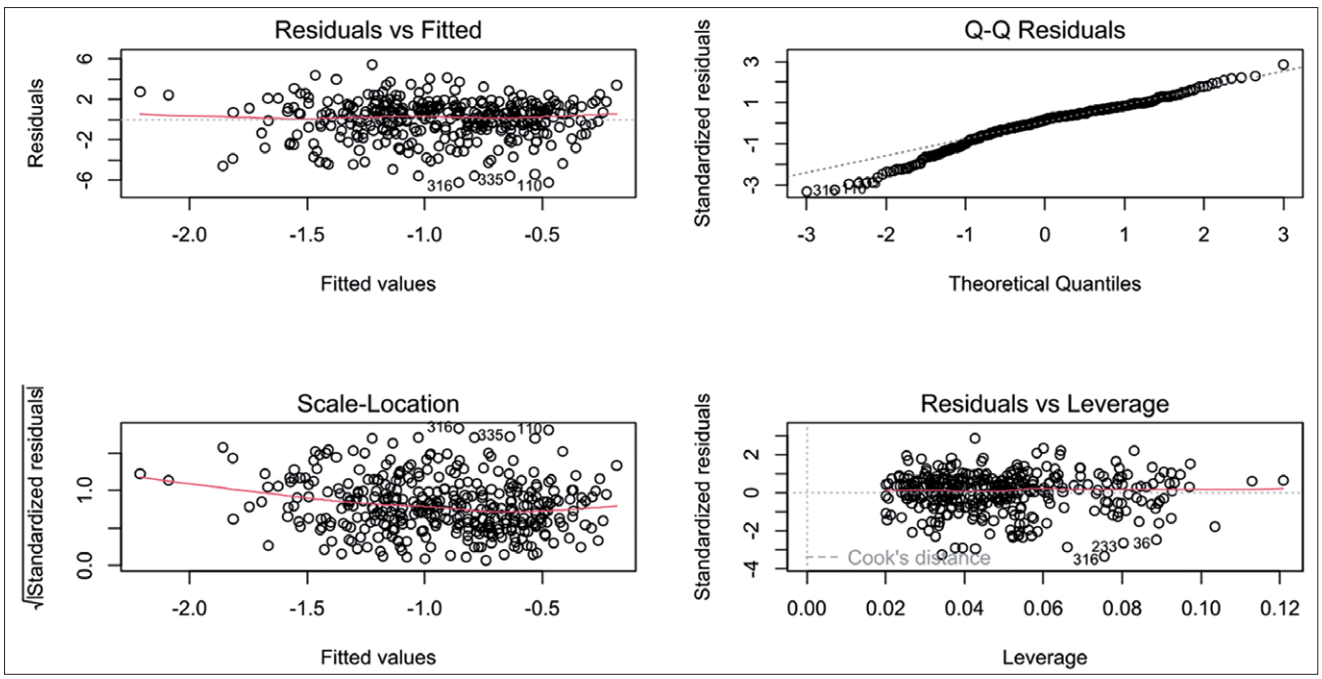
```
X real, Y real
LINEAR REGRESSION
Dependent variable
x1_ANC_x2_AC
Explanatory variables
AND, PV, CAT, CL, CM, GAL, MUR, NAV, RIO, VAL, Edad, SA, Exporta, Subvenciones, Ln_
empleados, Genero_bajo, Genero_no_divulgado
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.460305   0.282419   1.630  0.10403
XAND                 -0.387788   0.237929  -1.630  0.10403
XPV                  -0.386251   0.217892  -1.773  0.07715 .
XCAT                 -0.527729   0.176401  -2.992  0.00297 **
XCL                  -0.308553   0.168168  -1.835  0.06738 .
XCM                  -0.397039   0.197239  -2.013  0.04488 *
XGAL                 -0.302040   0.214893  -1.406  0.16075
XMUR                 -0.677614   0.268313  -2.525  0.01200 *
XNAV                 -0.301565   0.261327  -1.154  0.24930
XRIO                 -0.489688   0.219053  -2.235  0.02602 *
XVAL                  0.110781   0.267322   0.414  0.67883
XEdad                -0.003732   0.004351  -0.858  0.39164
XSA                  -0.136262   0.114046  -1.195  0.23297
XExporta             -0.295702   0.115509  -2.560  0.01089 *
XSubvenciones         0.182653   0.111221   1.642  0.10144
Xlog_empleados        0.004397   0.079789   0.055  0.95608
XGenero_bajo          0.009082   0.128855   0.070  0.94385
XGenero_no_divulgado -0.048807   0.115473  -0.423  0.67280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8949 on 350 degrees of freedom
Multiple R-squared:  0.0906,     Adjusted R-squared:  0.04643
F-statistic: 2.051 on 17 and 350 DF,  p-value: 0.008535
```



Repetimos el análisis con la siguiente log-ratio por pares: y_5 = maduración de la deuda = log(*PNC*/*PC*) = *x3_PNC_x4_PC*. La hipótesis nula del contraste global

no se rechaza con un valor *p* de 0,7536 > 0,05. Con ello, no está claro que ninguna de las variables explicativas consideradas ayude a predecir la maduración de la deuda, y ya no proseguimos con el análisis de los contrastes estadísticos, aunque los gráficos de los residuos muestren que los supuestos se cumplen.

```
X real, Y real
LINEAR REGRESSION
Dependent variable
x3_PNC_x4_PC
Explanatory variables
AND, PV, CAT, CL, CM, GAL, MUR, NAV, RIO, VAL, Edad, SA, Exporta, Subvenciones, Ln_
empleados, Genero_bajo, Genero_no_divulgado
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.757701   0.613234  -1.236   0.2174
XAND                 -0.219773   0.516631  -0.425   0.6708
XPV                  -0.441863   0.473123  -0.934   0.3510
XCAT                  0.124932   0.383031   0.326   0.7445
XCL                   0.141332   0.365153   0.387   0.6990
XCM                  -0.496236   0.428279  -1.159   0.2474
XGAL                  0.087922   0.466611   0.188   0.8507
XMUR                 -0.126651   0.582604  -0.217   0.8280
XNAV                  0.172849   0.567436   0.305   0.7608
XRIO                 -0.154377   0.475645  -0.325   0.7457
XVAL                 -0.158569   0.580452  -0.273   0.7849
XEdad                -0.018642   0.009448  -1.973   0.0493 *
XSA                   0.198703   0.247635   0.802   0.4229
XExporta             -0.227399   0.250811  -0.907   0.3652
XSubvenciones         0.339890   0.241502   1.407   0.1602
Xlog_empleados        0.028980   0.173250   0.167   0.8673
XGenero_bajo         -0.053854   0.279791  -0.192   0.8475
XGenero_no_divulgado  0.107316   0.250733   0.428   0.6689
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.943 on 350 degrees of freedom
Multiple R-squared:  0.035,      Adjusted R-squared:  -0.01187
F-statistic: 0.7468 on 17 and 350 DF,  p-value: 0.7536
```

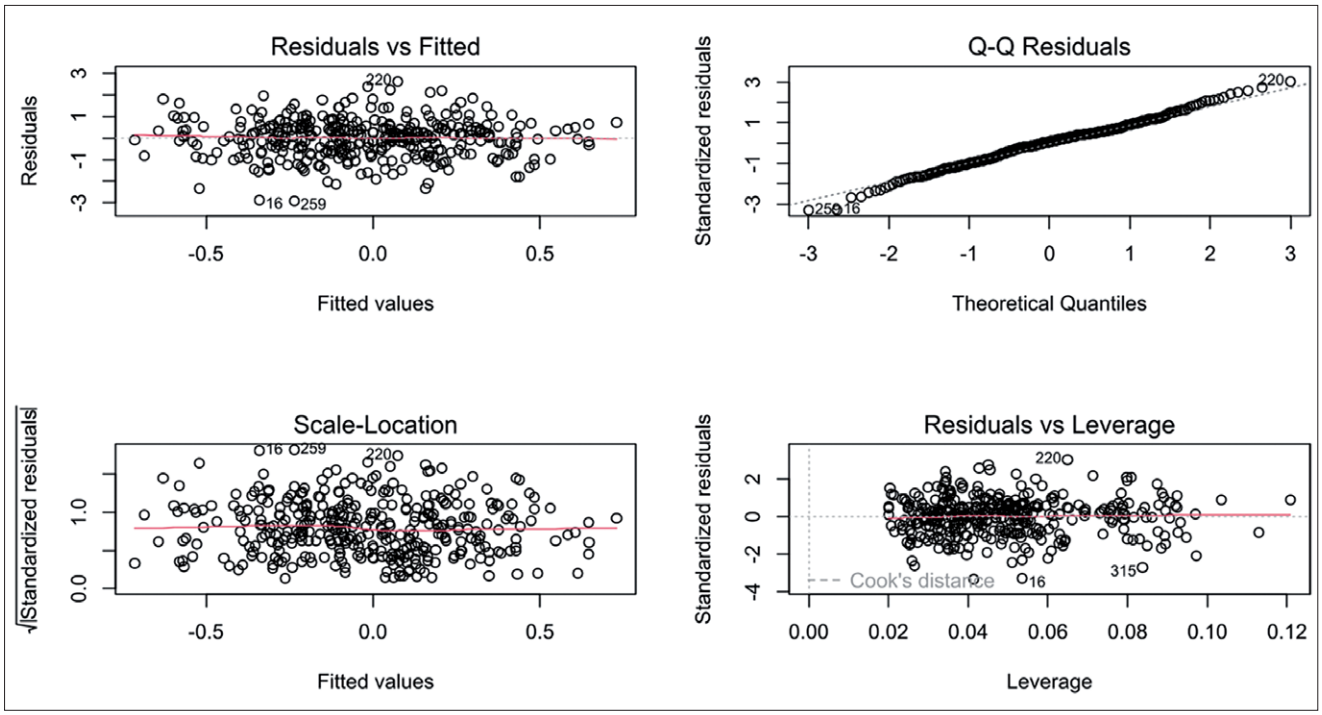


Recapitulando, todas las variables predictoras consideradas explican al menos una de las log-ratios por pares, excepto la diversidad de género. Sería legítimo repetir los cinco análisis sin *Genero_bajo* y *Genero_no_divulgado*, cosa que dejamos para el lector o lectora.

A continuación, mostramos cómo en el análisis composicional la permutación del numerador y del denominador no tiene ningún efecto sobre los resultados. Repetimos el análisis de y_4 = inmovilización del activo = log(*ANC/AC*) = *x1_ANC_x2_AC* permutando dichos numerador y denominador, es decir, calculando log(*AC/ANC*). Esto nos lleva simplemente a un cambio de signo, es decir, a $-y_4$. Esta variable ya había sido creada en el apartado 4.5 con el nombre de *x2_AC_x1_ANC*.

Lo que ocurre en la tabla de resultados es que el porcentaje de varianza explicado es idéntico, los valores *p* tanto global como individual son idénticos a los obtenidos con log(*ANC/AC*), y las estimaciones de los coeficientes (`Estimate`) son iguales cambiadas de signo. Asimismo, los gráficos de los residuos son copias simétricas de los obtenidos con log(*ANC/AC*). Los análisis de log(*ANC/AC*) y log(*AC/ANC*) llevan, pues, a conclusiones idénticas en todos los aspectos.

```
X real, Y real
LINEAR REGRESSION
Dependent variable
x2_AC_x1_ANC
Explanatory variables
AND, PV, CAT, CL, CM, GAL, MUR, NAV, RIO, VAL, Edad, SA, Exporta, Subvenciones, log_
empleados, Genero_bajo, Genero_no_divulgado
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.460305   0.282419  -1.630  0.10403
XAND                  0.387788   0.237929   1.630  0.10403
XPV                   0.386251   0.217892   1.773  0.07715 .
XCAT                  0.527729   0.176401   2.992  0.00297 **
XCL                   0.308553   0.168168   1.835  0.06738 .
XCM                   0.397039   0.197239   2.013  0.04488 *
XGAL                  0.302040   0.214893   1.406  0.16075
XMUR                  0.677614   0.268313   2.525  0.01200 *
XNAV                  0.301565   0.261327   1.154  0.24930
XRIO                  0.489688   0.219053   2.235  0.02602 *
XVAL                 -0.110781   0.267322  -0.414  0.67883
XEdad                 0.003732   0.004351   0.858  0.39164
XSA                   0.136262   0.114046   1.195  0.23297
XExporta              0.295702   0.115509   2.560  0.01089 *
XSubvenciones        -0.182653   0.111221  -1.642  0.10144
Xlog_empleados       -0.004397   0.079789  -0.055  0.95608
XGenero_bajo         -0.009082   0.128855  -0.070  0.94385
XGenero_no_divulgado  0.048807   0.115473   0.423  0.67280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8949 on 350 degrees of freedom
Multiple R-squared:  0.0906,     Adjusted R-squared:  0.04643
F-statistic: 2.051 on 17 and 350 DF,  p-value: 0.008535
```



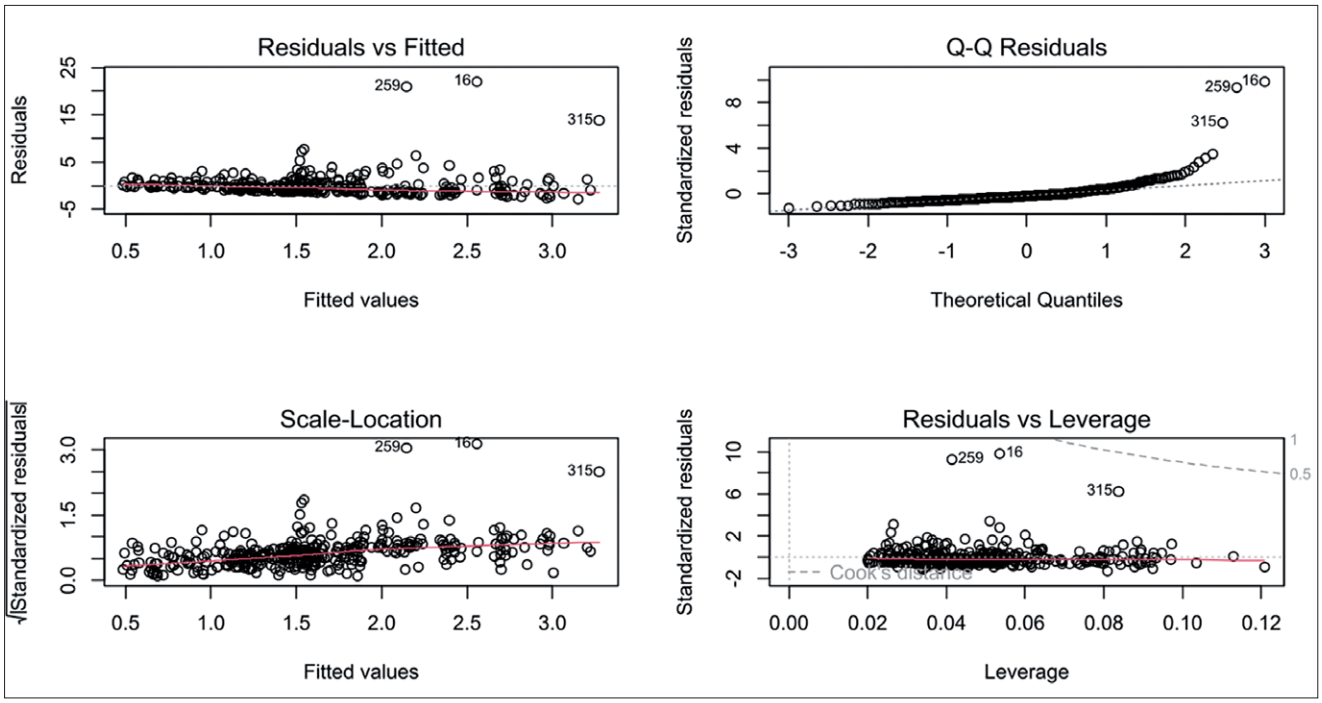
Trabajamos, a continuación, sobre la ratio clásica de inmovilización del activo, es decir, *Inmovilizacion* = *ANC/AC*. El menú de CoDaPack es el mismo. Entramos en *Statistics* > *Multivariate Analysis* > *Regression* > *X real Y real*. Introducimos las diecisiete variables explicativas en su versión numérica simultáneamente en el cuadro *Explanatory variables*. La variable dependiente *Inmovilizacion* se introduce en el cuadro *Response variable*.

Recordemos que con la log-ratio por pares $y_4 = \log(ANC/AC)$ eran significativas las variables *CAT*, *CM*, *MUR*, *RIO* y *Exporta*. Ahora tienen valores p (`Pr(>|t|)`) inferiores a 0,05 las variables *PV*, *CAT*, *CL*, *CM*, *MUR* y *RIO*, es decir, que se han añadido dos comunidades más, pero han desaparecido las exportaciones. Uno podría pensar que pasar de ratios clásicas a ratios composicionales afecta los resultados en su detalle o mejora la precisión en sus decimales, pero no es así. Afecta los resultados más básicos y modifica sustancialmente las conclusiones del estudio.

```
X real, Y real
LINEAR REGRESSION
Dependent variable
Inmovilizacion
Explanatory variables
AND, PV, CAT, CL, CM, GAL, MUR, NAV, RIO, VAL, Edad, SA, Exporta, Subvenciones, Ln_
empleados, Genero_bajo, Genero_no_divulgado
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.900563   0.726293   3.994 7.93e-05 ***
XAND                 -0.894912   0.611879  -1.463  0.14448
XPV                  -1.174129   0.560350  -2.095  0.03686 *
XCAT                 -1.240025   0.453648  -2.733  0.00659 **
XCL                  -1.144728   0.432475  -2.647  0.00849 **
XCM                  -1.012867   0.507238  -1.997  0.04662 *
XGAL                 -0.569723   0.552638  -1.031  0.30329
XMUR                 -1.586892   0.690017  -2.300  0.02205 *
XNAV                 -0.880160   0.672052  -1.310  0.19117
XRIO                 -1.191435   0.563337  -2.115  0.03514 *
XVAL                  0.234132   0.687468   0.341  0.73363
XEdad                -0.004057   0.011190  -0.363  0.71714
XSA                  -0.253698   0.293290  -0.865  0.38763
XExporta             -0.532600   0.297052  -1.793  0.07384 .
XSubvenciones        -0.035009   0.286026  -0.122  0.90265
Xlog_empleados       -0.031830   0.205191  -0.155  0.87681
XGenero_bajo          0.292015   0.331375   0.881  0.37880
XGenero_no_divulgado  0.001485   0.296959   0.005  0.99601
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.301 on 350 degrees of freedom
Multiple R-squared:  0.06596,    Adjusted R-squared:  0.02059
F-statistic: 1.454 on 17 and 350 DF,  p-value: 0.1092
```

Teniendo en cuenta la afirmación anterior, solo uno de los resultados puede ser el correcto, y lo es el resultado composicional basado en log(*ANC/AC*), si atendemos a los gráficos de los residuos:

- El diagrama *Residuals vs fitted* exhibe un patrón lineal.
- El diagrama *Scale-Location* exhibe una dispersión mayor a la derecha, con lo que se viola el supuesto de homocedasticidad.
- El diagrama *Q-Q Residuals* exhibe un patrón clarísimamente no lineal, con el correspondiente incumplimiento del supuesto de normalidad.
- El diagrama *Residuals vs Leverage* no muestra distancias de Cook más allá de una frontera de 0,5, pero las empresas 16 y 315 se aproximan bastante. Recordemos que en los diagramas de caja del apartado 4.5 habíamos apreciado observaciones atípicas muy extremas.



La ecuación de previsión es:

$$\begin{aligned} &2{,}901-0{,}895z_1-1{,}174z_2-1{,}240z_3-1{,}145z_4-1{,}013z_5-0{,}570z_6-1{,}587z_7-0{,}880z_8 \\ &-1{,}191z_9+0{,}234z_{10}-0{,}004z_{11}-0{,}254z_{12}-0{,}533z_{13}-0{,}035z_{14}-0{,}032z_{15}+0{,}292z_{16}+0{,}001z_{17} \end{aligned} \quad (50)$$

Por ejemplo, la previsión para la ratio clásica de inmovilización del activo *ANC/AC* para una empresa de Murcia, que lleva cien años en el sector, que es sociedad anónima, exportadora, sin subvenciones con quinientos empleados y con mayor porcentaje de mujeres empleadas que la mediana del sector es:

$$2{,}901-1{,}587\times 1-0{,}004\times 100-0{,}254\times 1-0{,}533\times 1-0{,}032\log(500)=-0{,}072 \quad (51)$$

Dado que la variable dependiente ya es la ratio clásica, no procede calcular la exponencial de la previsión, y constatamos que es un valor imposible (negativo) de la ratio, hecho que por sí solo convierte el modelo de regresión que hemos planteado en inútil.

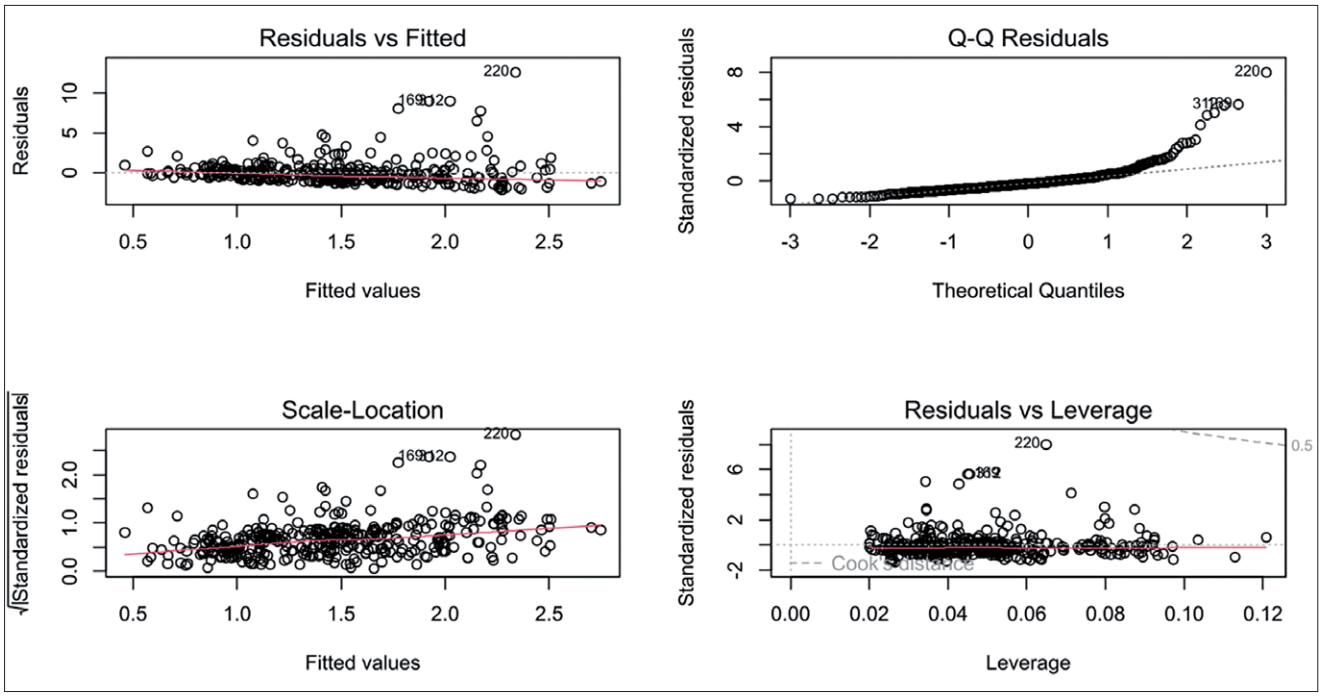
A pesar de ello, y solo para comparar los resultados, intentamos trabajar, a continuación, sobre la ratio inversa de inmovilización del activo tras permutar el numerador y el denominador, es decir, *AC/ANC*. Esta variable ya había sido creada en el apartado 4.5 con el nombre de *Inmovilizacion_inversa*.

Recordemos que con las log-ratios por pares log(*ANC/AC*) y log(*AC/ANC*) eran significativas las variables *CAT*, *CM*, *MUR*, *RIO* y *Exporta*. Con la ratio clásica *ANC/AC* lo eran *PV*, *CAT*, *CL*, *CM*, *MUR* y *RIO*. Con la ratio inversa *AC/ANC* lo son *CAT* y *Subvenciones*, con lo que casi no hay ninguna coincidencia. Debería preocupar mucho a los usuarios de las ratios clásicas que las conclusiones puedan cambiar hasta este extremo por el mero hecho de permutar el denominador y el numerador. Tan legítimo es calcular en qué medida el *ANC* excede el *AC* como calcular en qué medida el *AC* excede el *ANC*, lo que sería

una ratio de capacidad del activo de convertirse en líquido. Contra lo que algunos investigadores dan por supuesto, esta decisión no es nada inocua. Sí lo es con la metodología composicional que ofrece idénticos resultados con las log-ratios por pares log(*ANC/AC*) y log(*AC/ANC*). Los gráficos de los residuos vuelven a mostrar que las ratios clásicas suelen incumplir los supuestos.

- El diagrama *Residuals vs fitted* exhibe un patrón lineal.
- El diagrama *Scale-Location* exhibe una dispersión mayor a la derecha, con lo que se viola el supuesto de homocedasticidad.
- El gráfico *Q-Q Residuals* exhibe un patrón clarísimamente no lineal, con el correspondiente incumplimiento del supuesto de normalidad.
- El diagrama *Residuals vs Leverage* no muestra distancias de Cook más allá de una frontera de 0,05, pero la empresa 220 se aproxima bastante. Recordemos que en los diagramas de caja del apartado 4.5 habíamos apreciado observaciones atípicas muy extremas, que no son las mismas que antes de invertir la ratio.

```
X real, Y real
LINEAR REGRESSION
Dependent variable
Inmovilizacion_inversa
Explanatory variables
AND, PV, CAT, CL, CM, GAL, MUR, NAV, RIO, VAL, Edad, SA, Exporta, Subvenciones, Ln_
empleados, Genero_bajo, Genero_no_divulgado
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.521668   0.517124   2.943  0.00347 **
XAND                 0.583746   0.435661   1.340  0.18114
XPV                  0.286261   0.398972   0.717  0.47355
XCAT                 0.733487   0.323000   2.271  0.02376 *
XCL                  0.203044   0.307924   0.659  0.51008
XCM                  0.602034   0.361156   1.667  0.09642 .
XGAL                 0.184988   0.393481   0.470  0.63855
XMUR                 0.704780   0.491295   1.435  0.15231
XNAV                 0.484893   0.478504   1.013  0.31159
XRIO                 0.767060   0.401099   1.912  0.05664 .
XVAL                -0.132304   0.489480  -0.270  0.78709
XEdad               -0.002875   0.007967  -0.361  0.71838
XSA                  0.157927   0.208824   0.756  0.45000
XExporta             0.357243   0.211502   1.689  0.09210 .
XSubvenciones       -0.606288   0.203652  -2.977  0.00311 **
Xlog_empleados      -0.075970   0.146097  -0.520  0.60339
XGenero_bajo         0.093280   0.235941   0.395  0.69282
XGenero_no_divulgado 0.243226   0.211437   1.150  0.25079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.639 on 350 degrees of freedom
Multiple R-squared:  0.07549,    Adjusted R-squared:  0.03059
F-statistic: 1.681 on 17 and 350 DF,  p-value: 0.04429
```



8.3. Ratios como variables explicativas

Una posibilidad alternativa en el modelo de regresión es que la composición (es decir, el conjunto de ratios financieras composicionales) prediga una variable dependiente numérica no composicional y no financiera *w*. Estadísticamente, el tema ha sido tratado en Aitchison y Bacon-Shone (1984); Coenders y Greenacre (2023); Coenders y Pawlowsky-Glahn (2020) y Hron et al. (2012), pero no existe hasta el momento ninguna aplicación en el campo de los estados financieros.

En este caso, todas las log-ratios por pares se incluyen simultáneamente en el lado derecho de una única ecuación de regresión. También se pueden incluir variables explicativas no financieras adicionales *z* si aumentan el poder explicativo.

$$w = \alpha + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \beta_4 y_4 + \beta_5 y_5 + \beta_6 z_1 + \beta_7 z_2 + \varepsilon \quad (52)$$

La variable *w* es la variable numérica dependiente, generalmente no financiera. Si es financiera, nunca podrá calcularse a partir de los mismos valores de los estados financieros contenidos en las mismas partes x_1 a x_6, con el fin de evitar correlaciones espurias. Las log-ratios por pares y_1 a y_5 son las calculadas según las ecuaciones (19) a (23), z_1 y z_2 son las variables predictoras no financieras como en el apartado 8.1, el parámetro $\boldsymbol{\alpha}$ es el término constante, y los parámetros $\boldsymbol{\beta}$ son los efectos de cada una de las log-ratios por pares y cada uno de los predictores no financieros sobre *w*. Estos efectos se interpretan manteniendo constantes todos los demás predictores (financieros y no financieros). El término ε contiene los *residuos*, que representan la parte de *w* que los predictores financieros y no financieros no explican.

Se contrastan las siguientes ocho hipótesis estadísticas correspondientes a los parámetros $\boldsymbol{\beta}$ de la ecuación de regresión (52). La primera de ellas corresponde al contraste global:

$$H_0: \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=0 \text{ (ninguna de las variables explicativas afecta } w)$$
$$H_0: \beta_1=0 \text{ (la rotación del activo corriente no afecta } w)$$
$$H_0: \beta_2=0 \text{ (el margen no afecta } w)$$
$$H_0: \beta_3=0 \text{ (la solvencia a corto plazo no afecta } w)$$
$$H_0: \beta_4=0 \text{ (la inmovilización del activo no afecta } w)$$
$$H_0: \beta_5=0 \text{ (la maduración de la deuda no afecta } w)$$
$$H_0: \beta_6=0 \text{ } (z_1 \text{ no afecta } w)$$
$$H_0: \beta_7=0 \text{ } (z_2, \text{ es decir el tipo de empresa, no afecta } w)$$

El valor *p* asociado a cada contraste estadístico indica el riesgo que implica rechazar la hipótesis nula. Si este es bajo (por ejemplo, inferior a 0,05), la hipótesis puede ser rechazada, lo que lleva a la conclusión de que al menos uno de los predictores (contraste global) o el predictor particular en cuestión (contraste individual) afecta *w*, manteniendo todos los demás predictores constantes. En otras palabras, llegamos a la conclusión de que su efecto es estadísticamente significativo. Por ejemplo, un coeficiente β_1 significativo y positivo indicaría que las empresas con mayor rotación del activo corriente tienden a tener valores de *w* más altos, manteniendo constantes el margen, la solvencia a corto plazo, la inmovilización del activo, la maduración de la deuda, z_1 y el tipo de empresa. Un coeficiente β_6 significativo y negativo indicaría que las empresas con mayor z_1 tienden a tener menor *w*, manteniendo constantes la rotación del activo corriente, el margen, la solvencia a corto plazo, la inmovilización del activo, la estructura de la deuda y el tipo de empresa. Un coeficiente β_7 significativo y negativo indicaría que las empresas de tipo B tienden a tener menor *w* que las de referencia (tipo A), manteniendo constantes la rotación del activo corriente, el margen, la solvencia a corto plazo, la inmovilización del activo, la maduración de la deuda y z_1.

Dado que solo hay una ecuación, solo hay un conjunto de gráficos de los residuos para verificar los supuestos del modelo. Se interpretan como en el apartado 8.1. Además, hay un solo coeficiente de determinación R^2 que indica el porcentaje de varianza en *w* explicado conjuntamente por los predictores financieros y no financieros.

8.4. Manos a la obra con CoDaPack. Predecimos indicadores de sostenibilidad a partir de las ratios

Mostramos un ejemplo de regresión donde una variable no financiera queda explicada por la información financiera. El archivo de datos que hemos usado en los ejemplos anteriores procedente de Arimany-Serrat et al. (2023) no contiene ninguna variable no financiera apta para este propósito, con lo que aquí usamos otro archivo de datos que presentamos a continuación.

Usamos estados financieros del sector cervecero español (CNAE 1105 «Fabricación de cerveza»). Se trata de un sector también de importancia en

España, cuarto país europeo productor en 2021. El sector fue enormemente afectado por la pandemia de COVID-19 y, a continuación, por el alza de los precios de la cebada resultante de la guerra de Ucrania, con ratios medias de margen negativas (Arimany-Serrat y Sgorla, 2024). A pesar de ello existen escasos estudios publicados de la salud financiera del sector (Jantyik et al., 2021; Zanotti et al., 2018).

Los datos fueron obtenidos de la base de datos SABI para el año 2021 sobre una muestra de $n = 51$ empresas mercantiles activas durante ese año, en forma jurídica de sociedad anónima y sociedad limitada y que disponían de página web. El sector se halla sujeto a enormes retos ambientales (Sozen et al., 2022) y a través de un análisis de contenido de dichas páginas identificamos la comunicación web de veinticinco indicadores de ESG incluidos en la Global Reporting Initiative (GRI), en particular, de seis indicadores medioambientales (consumo de energía, consumo de agua, emisiones contaminantes, generación de residuos, residuos gestionados y residuos reutilizados), diez indicadores sociales (número de trabajadores, diversidad de género de los trabajadores, estabilidad laboral, absentismo, rotación de los trabajadores, creación neta de empleo, antigüedad laboral, formación de los trabajadores, plazos de cobro de los clientes y plazos de pago a los proveedores) y nueve de buen gobierno (lista de los consejeros, consejeros independientes, consejeros con funciones de responsabilidad social corporativa, comisión ejecutiva, comité de auditoría, nombramientos del consejo, reuniones del consejo, remuneración total del consejo y diversidad de género del consejo) (véase el apartado 2.3). El análisis de contenido de las webs se hizo en marzo de 2023. Se eliminaron empresas sin ingresos de explotación, al considerarse ceros absolutos. La base de datos fue utilizada en Coenders et al. (2023a). El archivo Excel se llama *cerveceras.xls*, fue compilado por el coautor Andrey Felipe Sgorla, se halla disponible en ResearchGate (https://doi.org/10.13140/RG.2.2.24056.87040) y contiene las variables:

- *Comunidad_autonoma* (Categórica).
- *AT*: x_1 activo total.
- *PT*: x_2 pasivo total.
- *IE*: x_3 ingresos de explotación.
- *GE*: x_4 gastos de explotación.
- *num_indic_total*: número total de indicadores ESG que aparecen en el web de la empresa de entre los veinticinco indicadores examinados.

La elección de los D valores de los estados financieros que hay que estudiar puede cambiarse según los objetivos del investigador o investigadora. Esta agregación en $D = 4$ valores permite calcular algunas log-ratios por pares que no podían calcularse con $D = 6$. Además de la log-ratio de margen $y_2 = \log(IE/GE) = \log(x_3/x_4)$, se pueden calcular la de rotación entendida sobre el activo total $y_1 = \log(IE/AT) = \log(x_3/x_1)$ y la de endeudamiento $y_3 = \log(PT/AT) = \log(x_2/x_1)$. Estas $D - 1 = 4 - 1 = 3$ log-ratios por pares contienen toda la información en $x_1,\ldots,x_4$ al definir un grafo acíclico y conexo (figura 2). Esta misma combinación de cuatro valores contables se ha usado en Saus-Sala et al. (2021; 2023) y son las necesarias en el marco del análisis DuPont de Donaldson Brown (Dale et al., 1980), que descompone el ROE en margen, rotación y apalancamiento (Baležentis et al., 2019; Chen et al., 2014).

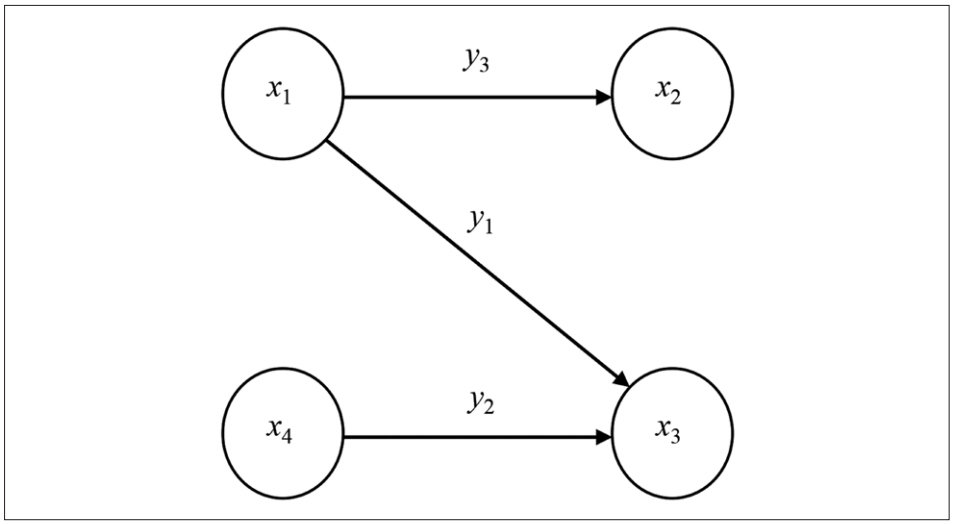
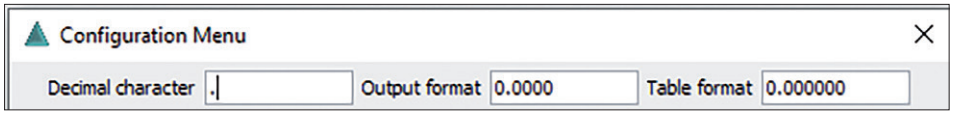


Figura 2. Grafo acíclico y conexo con $D = 4$

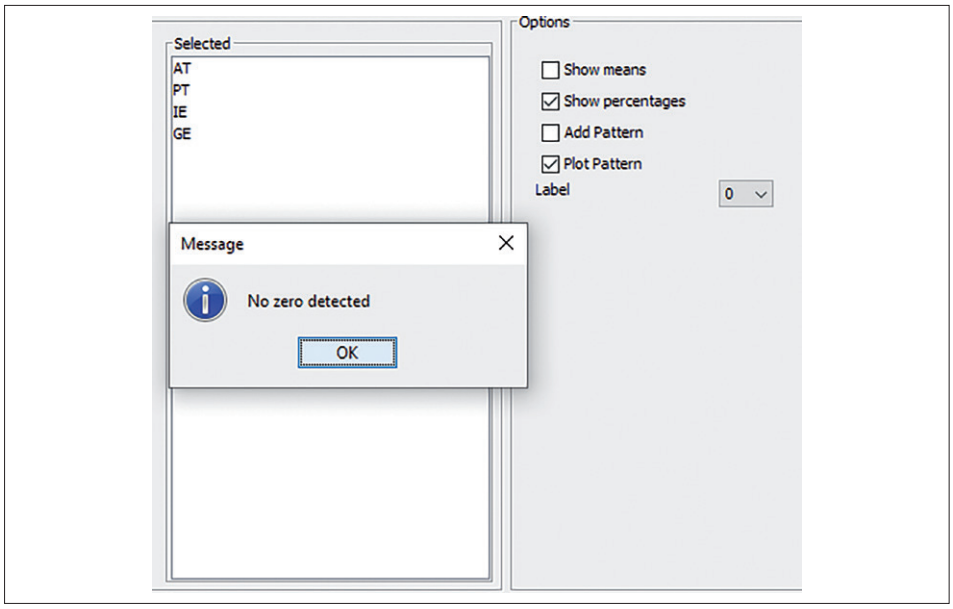
Se trata de ver si esas tres log-ratios por pares (variables independientes, exógenas, explicativas o predictoras) explican significativamente la variable *num_indic_total* (dependiente, endógena, explicada o respuesta). En otras palabras, si la rotación, el margen y el endeudamiento condicionan la comunicación por web realizada. Por ejemplo, ¿podemos afirmar que las empresas con una mayor rotación realizan más comunicación? ¿Y las empresas con mayor margen? ¿Y las empresas menos endeudadas?

Para leer el archivo de datos Excel (por ahora, CoDaPack es compatible solo con archivos en formato *.xls*), seleccionamos el menú *File* > *Import* > *Import XLS Data* con las opciones por defecto. El archivo de Excel solo debe contener una hoja con los nombres de las variables en la primera fila y los datos de la segunda fila en adelante. Los datos pueden ser texto o números, no fórmulas. Los nombres de las variables solo pueden contener letras del alfabeto inglés, números, puntos y guiones bajos «_», y no pueden incluir espacios ni tildes. Los ceros en los datos contables deben introducirse como tales. Los datos faltantes en las variables no contables (variables desconocidas en alguna de las empresas), como «NA». La variable *Comunidad_autonoma* está resaltada en color naranja, que indica su carácter categórico.

En el menú *File* > *Configuration*, seleccionamos el número de decimales que queremos visualizar en la tabla de datos y en los resultados.



En primer lugar, hay que mirar si los datos contienen ceros. El menú *Irregular Data* > *Zero Patterns* calcula los porcentajes de ceros por parte y en general, y los porcentajes de coocurrencia de ceros, después de introducir simultáneamente las partes *AT*, *PT*, *IE* y *GE* en el cuadro *Selected* con las opciones *Show percentages* y *Plot Pattern*. Tras clicar *Accept*, un mensaje nos advierte de que no hay ceros, con lo que ya no tenemos que preocuparnos ni por sus límites de detección ni por su reemplazamiento.

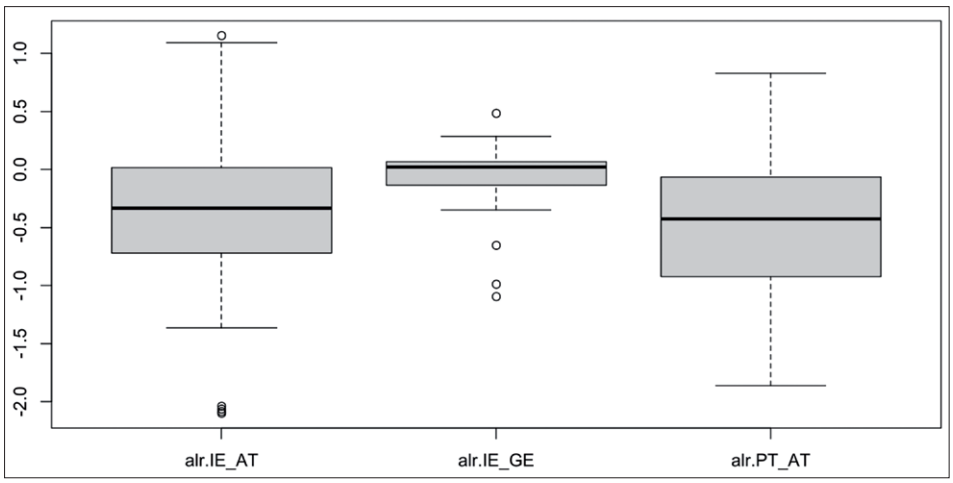


Construimos las D – 1 log-ratios por pares que deseamos. El menú *Data* > *Transformation* > *ALR* almacena las log-ratios por pares como variables adicionales al final del archivo de datos, después de introducir las dos partes involucradas en el cuadro *Selected,* la parte del numerador primero, la parte del denominador después, y con la opción *Raw-ALR*. Por ejemplo, para la ratio de rotación del activo total, introducimos en este orden *IE* y *AT* y la variable creada se llama *alr.IE_AT*. Hacemos lo propio con las otras dos log-ratios. Clicando dos veces sobre los nombres de las variables en el encabezamiento de la tabla de datos se podrían cambiar por otros a gusto del usuario o usuaria (siempre que contengan solo números, puntos, guiones bajos y letras del alfabeto inglés sin tildes). Hemos optado por dejarlos como están, al entender que expresan con claridad cómo se han construido cada una. Para las primeras filas del archivo de datos obtenemos:

	Comunidad_autonoma	AT	PT	IE	GE	num_indic_total	alr.IE_AT	alr.IE_GE	alr.PT_AT
1	Madrid	1967180....	471981.0...	1225160....	1057983....	17.000000	-0.473530	0.146707	-1.427418
2	Catalunya	1427910....	661260.0...	786112.0...	738335.0...	13.000000	-0.596868	0.062702	-0.769820
3	Galicia	567933.0...	273214.0...	573556.0...	443895.0...	10.000000	0.009852	0.256268	-0.731748
4	Catalunya	278696.0...	109641.0...	229439.0...	225205.0...	13.000000	-0.194484	0.018626	-0.932910
5	Comunidad_Valenciana	260133.0...	78121.00...	370802.0...	331884.0...	6.000000	0.354475	0.110883	-1.202934
6	Canarias	134539.0...	41975.00...	121880.0...	113410.0...	10.000000	-0.098817	0.072027	-1.164780
7	Andalucia	73296.00...	22807.00...	68719.00...	63177.00...	17.000000	-0.064480	0.084085	-1.167439
8	Madrid	101026.0...	36933.00...	79639.00...	73324.00...	18.000000	-0.237874	0.082616	-1.006272
9	Murcia	46557.00...	29972.00...	74248.00...	71962.00...	3.000000	0.466733	0.031273	-0.440414
10	Andalucia	16971.00...	8695.000...	12632.00...	12038.00...	0	-0.295273	0.048165	-0.668758

Es un buen momento para guardar todas las variables que hemos ido creando por medio del menú *File* > *Save as*. El nuevo archivo tendrá el formato nativo de CoDaPack con la terminación *.cdp* . En sesiones futuras podrá abrirse ejecutando el menú *File* > *Open Workspace*.

El menú *Graphs* > *Boxplot* representa los diagramas de caja después de introducir simultáneamente las tres log-ratios por pares en el cuadro *Selected*. Observamos algunas observaciones atípicas por el lado inferior de las log-ratios de rotación (*alr.IE_AT*) y margen (*alr.IE_GE*) que podría dudarse de si son o no extremas. En el ejemplo que sigue consideramos que no lo son y no las eliminamos.



El centro composicional se obtiene por medio del menú *Statistics* > *Compositional Statistics Summary*. Introducimos las partes *AT*, *PT*, *IE* y *GE* en el cuadro *Selected* solo con la opción *Center*. Estos serían los valores que hay que usar para calcular cualesquiera ratios clásicas representativas de la empresa media del sector cervecero.

Compositional statistics summary:
NA's:
0

Sample size:
51

Statistics

	Center
AT	0.3300
PT	0.2050
IE	0.2264
GE	0.2386

De estos resultados surge un margen medio negativo igual a (0,2264 – 0,2386)/0,2264 = –0,0539, una rotación media igual a 0,2264/0,3300 = 0,6861 y un apalancamiento medio igual a 0,3300/(0,3300 –0,2050) = 2,6400. A partir de estos, surge un ROA igual a –0,0539 × 0,6861 = –0,0370 y un ROE igual a –0,0370 × 2,6400 = –0,0977, que completan la descomposición del ROE según el análisis DuPont. En este sector con margen negativo, el alto apalancamiento empeora las cosas y resulta un ROE todavía más negativo que se aproxima a una rentabilidad financiera de –10 %. El alto apalancamiento representa una

frágil situación financiera a largo plazo pues indica que los activos representan casi tres veces los fondos propios o, lo que es lo mismo, que los fondos propios solo financian algo más de un tercio de los activos. En efecto, la ratio media de endeudamiento es 0,2050/0,3300 = 0,6212. Dicho esto, la mayor amenaza para una empresa media del sector son los márgenes negativos en más de –5 %.

Como la variable *num_indic_total* es no composicional, procede su descripción con estadísticos clásicos por medio del menú *Statistics* > *Classical Statistics Summary*. Si marcamos las opciones *Mean* (media) y *Standard Deviation* (desviación típica), obtenemos una media de casi tres indicadores por empresa:

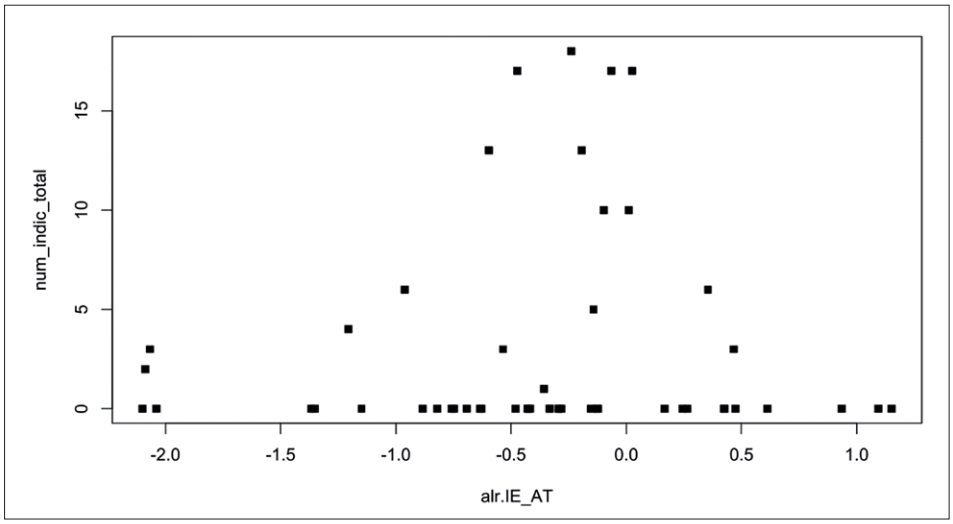
Clasical statistics summary:
NA's:
0

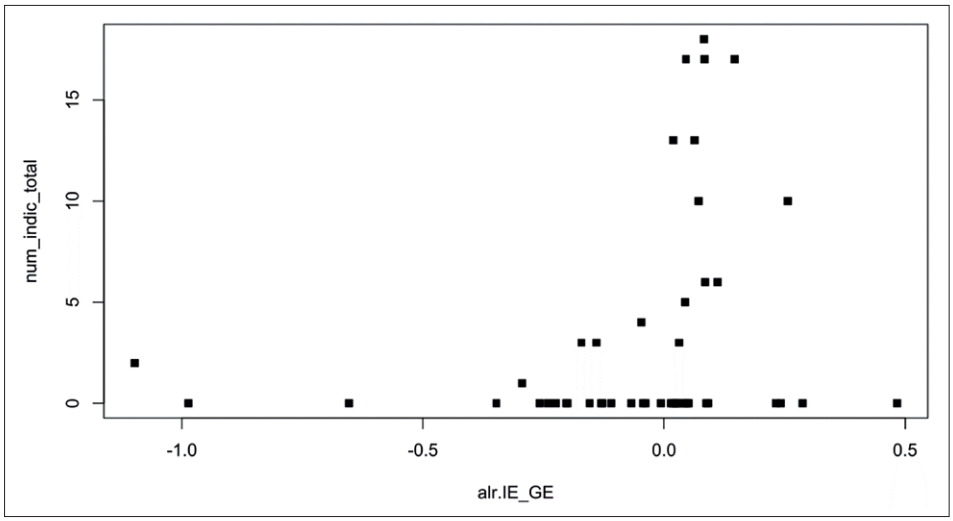
Sample size:
51

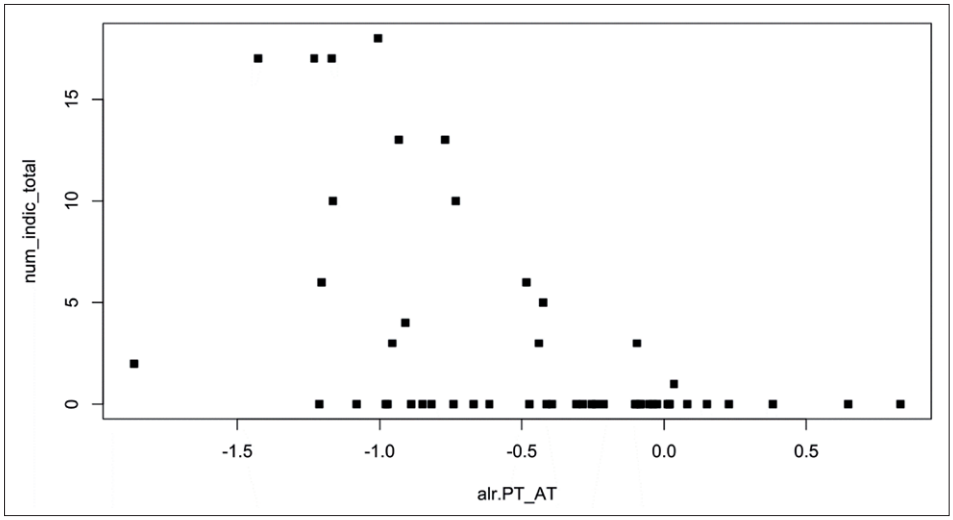
Statistics

	Mean	Std.Dev
num_indic_total	2.9020	5.3220

Pasemos al análisis exploratorio de los datos previo al análisis de regresión. El menú *Graphs* > *Scatterplot* produce diagramas de dispersión mediante la introducción de dos variables numéricas en el cuadro *Selected*. La variable introducida en primer lugar aparece en el eje horizontal y corresponde con una variable explicativa en la regresión, en nuestro caso una log-ratio por pares. La variable introducida en el segundo lugar es la dependiente en la regresión, en nuestro caso *num_indic_total*. Son necesarios, pues, tres gráficos.







Las relaciones que se observan son débiles en la mayoría de las log-ratios. Las más visibles son una relación positiva con la log-ratio de margen (*alr.IE_GE*) i una negativa con la log-ratio de endeudamiento (*alr.PT_AT*). Hay indicios de curvatura en el gráfico con el margen, que quedarían pendientes de comprobar en los gráficos de residuos.

Pasemos a estimar el modelo de regresión. Como partimos de las log-ratios previamente calculadas, todas las variables tienen la consideración de reales y el menú que se usa es *Statistics* > *Multivariate Analysis* > *Regression* > *X real Y real*. Introducimos las tres log-ratios *alr.IE_AT*, *alr.IE_GE* y *alr.PT_AT* simultáneamente en el cuadro *Explanatory variables*. En este caso, no disponemos de variables explicativas no financieras; de haberlas, se incluirían simultáneamente en el mismo cuadro junto con las log-ratios. La variable dependiente *num_indic_total* se introduce en el cuadro *Response variable*.

```
X real, Y real
LINEAR REGRESSION
Dependent variable
num_indic_total
Explanatory variables
alr.IE_AT, alr.IE_GE, alr.PT_AT
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4305     0.8319   0.518  0.60723
Xalr.IE_AT   -0.1704     0.9453  -0.180  0.85775
Xalr.IE_GE    7.5313     2.6664   2.824  0.00693 **
Xalr.PT_AT   -5.8879     1.1313  -5.205  4.2e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.293 on 47 degrees of freedom
Multiple R-squared:  0.4002,     Adjusted R-squared:  0.362
F-statistic: 10.45 on 3 and 47 DF,  p-value: 2.194e-05
```

Al contrario que cuando la composición es dependiente, el modelo tiene una sola ecuación. Es:

$$num\_indic\_total = \alpha + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \varepsilon \quad (53)$$

Donde y_1 se refiere a la log-ratio por pares de rotación del activo total (*alr.IE_AT*), y_2 al margen (*alr.IE_GE*) e y_3 al endeudamiento (*alr.PT_AT*).

La ecuación estimada de previsión (`Estimate`) es:

$$0{,}4305 - 0{,}1704 y_1 + 7{,}5313 y_2 - 5{,}8879 y_3 \quad (54)$$

Por ejemplo, el número de indicadores comunicados previsto para una empresa con 100 000 euros de activo, 50 000 euros de pasivo, 100 000 euros de ingresos de explotación y 90 000 euros de gastos de explotación es:

$$0{,}4305 - 0{,}1704 \log\left(\frac{100\,000}{100\,000}\right) + 7{,}5313 \log\left(\frac{100\,000}{90\,000}\right) - 5{,}8879 \log\left(\frac{50\,000}{100\,000}\right) = 5{,}305 \quad (55)$$

El porcentaje de varianza explicado es 40,0 % (`Multiple R-squared`). El contraste global de la hipótesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

permite su rechazo con un valor p de 0,00002194 < 0,05 (`p-value`).

Los valores p de las variables *alr.IE_GE* y *alr.PT_AT* (`Pr(>|t|)`) son inferiores a 0,05 (0,00693 y 0,0000042, respectivamente), lo que permite rechazar las siguientes hipótesis con riesgo α = 5 %:

$$H_0: \beta_2 = 0$$
$$H_0: \beta_3 = 0$$

Según los signos de los coeficientes para las hipótesis rechazadas:

- Si aumenta el margen, manteniendo constantes la rotación y el endeudamiento, tiende a aumentar el número de indicadores medioambientales, sociales y de buen gobierno que la empresa comunica en su web.
- Si aumenta el endeudamiento, manteniendo constantes la rotación y el margen, tiende a disminuir el número de indicadores medioambientales, sociales y de buen gobierno que la empresa comunica en su web (signo negativo del coeficiente en `Estimate`).

No podemos afirmar que exista relación entre la rotación y el número de indicadores medioambientales, sociales y de buen gobierno que la empresa comunica en su web.

8.5. Para saber más. Modelado estadístico avanzado

Un ejemplo sencillo de aplicación con ratios financieras composicionales dependientes se encuentra en Mulet-Forteza et al. (2024). La situación con ratios financieras composicionales explicativas puede extenderse a una variable dependiente *w* no numérica, pero esto no se puede hacer con un modelo de regresión lineal. En su lugar, se debe utilizar un *modelo lineal generalizado* (Coenders et al., 2017; Coenders y Greenacre, 2023). Un caso particular útil de modelo lineal generalizado es el modelo de *regresión logística* o *modelo logit*, para una variable *w* binaria (por ejemplo, un indicador de mora o de quiebra codificado como «1» cuando hay impago o quiebra y como «0» en caso contrario). Los modelos lineales generalizados están fuera del alcance de este libro. Hasta donde sabemos, no hay trabajos publicados que utilicen como variables explicativas las ratios financieras composicionales, ya sea con modelos lineales o lineales generalizados. Esto constituye una vía potencial de futuras aplicaciones.

Además de los métodos estadísticos aquí descritos, el análisis composicional de los estados financieros ha utilizado *modelos de ecuaciones estructurales* estimados por *mínimos cuadrados parciales* (Creixans-Tenas et al., 2019), *modelos de vectores autoregresivos* (Carreras-Simó y Coenders, 2021), *modelos de datos de panel* (Arimany-Serrat et al., 2023; Carreras-Simó y Coenders, 2021; Escaramís y Arbussà, 2025), y tiene potencial para expandirse a cualquier otro método o modelo estadístico o econométrico utilizado en contabilidad y finanzas. La incorporación de varios años de datos en el análisis debe hacerse, precisamente, con modelos de datos de panel y queda fuera del alcance de este libro.

La metodología composicional también tiene potencial para cualquier proyecto de investigación empresarial que utilice modelos estadísticos o econométricos e incluya ratios financieras entre el conjunto de variables de estudio. Carreras-Simó y Coenders (2021) relacionan las estructuras de activos y las estructuras de capital; Escaramís y Arbussà (2025) comparan las estructuras de capital de las empresas familiares y las que no lo son; Creixans-Tenas et al. (2019) estudian el impacto de la responsabilidad social corporativa en la rentabilidad y la solvencia; Mulet-Forteza et al. (2024), el impacto de las estrategias de expansión, y Arimany-Serrat et al. (2023) el impacto de la pandemia de COVID-19.

Por lo general, una vez calculadas las log-ratios por pares, el uso de los modelos estadísticos es el estándar, cosa que es uno de los principales atractivos de la metodología CoDa: ofrecer una solución unificada y adaptable al análisis estadístico de ratios financieras. De hecho, ninguno de los métodos estadísticos citados en este apartado está disponible en CoDaPack. El usuario o usuaria, tras el reemplazamiento de los ceros y el cálculo de las log-ratios, puede exportar los datos de CoDaPack a Excel, y de Excel importarlos en su software de análisis de datos favorito para seguir trabajando del modo habitual. Muchos usuarios de CoDa optan por el software de libre distribución R (R Core Team, 2022), que dispone de muchas librerías para análisis CoDa, como zCompositions (Palarea-Albaladejo y Martín-Fernández, 2015), compositions (Van den Boogaart y Tolosana-Delgado, 2013), robCompositions (Filzmoser et al., 2018), easyCODA (Greenacre, 2018) y coda4microbiome (Calle et al., 2023).

9. Decálogo final

1. Las ratios se deben calcular sobre valores positivos.

2. Las ratios representativas de un sector deben utilizar medias geométricas y no aritméticas.

3. Los logaritmos son complementos necesarios de las ratios.

4. Distintos análisis pueden requerir distintas log-ratios. Las log-ratios centradas se adaptan al *biplot* y al análisis clúster; las log-ratios por pares, a la regresión.

5. La imagen fiel del análisis de los estados financieros en el ámbito contable y en el ámbito estadístico se consigue con datos composicionales.

6. Las observaciones atípicas, la asimetría y las curvaturas no deben afectar los resultados.

7. Los sectores no son homogéneos en el ámbito financiero y las medias por clústeres suelen ser más útiles que las globales.

8. La metodología CoDa utiliza un tratamiento riguroso de los valores iguales a cero en los estados financieros.

9. Es posible visualizar todas las empresas y todas las ratios en un único gráfico bidimensional.

10. La metodología CoDa mejora la toma de decisiones económicas derivadas del análisis de los estados financieros.

Bibliografía

Adcock, C., Eling, M. y Loperfido, N. (2015). Skewed distributions in finance and actuarial science: A review. *The European Journal of Finance*, *21*(13–14), 1253–1281. https://doi.org/10.1080/1351847X.2012.720269

Ahmed, A. H., Elmaghrabi, M., Burton, B. y Dunne, T. (2023). Corporate internet reporting in Egypt: A pre-and peri-uprising analysis. *International Journal of Organizational Analysis*, *31*(6), 2409–2440. https://doi.org/10.1108/IJOA-09-2021-2970

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. https://www.jstor.org/stable/2345821

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, *70*(1), 57–65. https://doi.org/10.1093/biomet/70.1.57

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

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. En V. Pawlowsky-Glahn (Ed.), *Proceedings of the IAMG'97—The 3rd annual conference of the international association for mathematical geology* (pp. 3–35). International Center for Numerical Methods in Engineering (CIMNE).

Aitchison, J. y Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*, *71*, 323–330. https://doi.org/10.1093/biomet/71.2.323

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. y Pawlowsky-Glahn V. (2000). Logratio analysis and compositional distances. *Mathematical Geology*, *32*(3), 271–275. https://doi.org/10.1023/A:1007529726302

Aitchison, J. y Greenacre, M. (2002). *Biplots* of compositional data. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *51*(4), 375–392. https://doi.org/10.1111/1467-9876.00275

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.2307/2978933

Amat Salas, O. (2020). Caso práctico de utilización de ratios para la detección de fraudes contables. *Técnica Contable y Financiera*, *33*, 98–105.

Arimany-Serrat, N. y Coenders, G. (2025). Biodiversidad y contabilidad: Metodología composicional en el análisis contable del sector apícola. *Economía Agraria y Recursos Naturales*, *25*(1), 7–36. https://doi.org/10.7201/earn.2025.01.01

Arimany-Serrat, N. y Farreras-Noguer, À. (2020). A comparison of the wine sectors in Catalonia, La Rioja, Languedoc-Roussillon and Emilia-Romagna. *Journal of Intelligent & Fuzzy Systems*, *38*(5), 5553–5563. https://doi.org/10.3233/JIFS-179646

Arimany-Serrat, N., Farreras-Noguer, M. À. y Coenders, G. (2022). New developments in financial statement analysis. Liquidity in the winery sector. *Accounting*, *8*, 355–366. https://doi.org/10.5267/j.ac.2021.10.002

Arimany-Serrat, N., Farreras-Noguer, M. À. y Coenders, G. (2023). Financial resilience of Spanish wineries during the COVID-19 lockdown. *International Journal of Wine Business Research*, *35*(2), 346–364. https://doi.org/10.1108/IJWBR-03-2022-0012

Arimany-Serrat, N., Farreras-Noguer, À. y Rabaseda i Tarrés, J. (2016). Análisis económico financiero del sector vinícola de La Rioja en un entorno de crisis. *Intangible Capital*, *12*(1), 268–294. https://doi.org/10.3926/ic.686

Arimany-Serrat, N. y Sgorla, A. F. (2024). Financial and ESG analysis of the beer sector pre and post COVID-19 in Italy and Spain. *Sustainability*, *16*(17), 7412. https://doi.org/10.3390/su16177412

Balcaen, S. y Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, *38*(1), 63–93. https://doi.org/10.1016/j.bar.2005.09.001

Baležentis, T., Galnaitytė, A., Krišciukaitienė, I., Namiotko, V., Novickytė, L., Streimikiene, D. y Melnikiene, R. (2019). Decomposing dynamics in the farm profitability: An application of index decomposition analysis to Lithuanian FADN sample. *Sustainability*, *11*(10), 2861. https://doi.org/10.3390/su11102861

Barnes, P. (1987). The analysis and use of financial ratios: A review article. *Journal of Business Finance & Accounting*, *14*(4), 449–461. https://doi.org/10.1111/j.1468-5957.1987.tb00106.x

Bastida Vialcanet, R. y Subirats Alcoverro, X. (2023). Las normas europeas de información sobre sostenibilidad (NEIS). *Técnica Contable y Financiera*, *67*, 8.

Belles-Sampera, J., Guillen, M. y Santolino, M. (2016). Compositional methods applied to capital allocation problems. *Journal of Risk*, *19*(2), 15–30. https://doi.org/10.21314/JOR.2016.345

Bhimani, A. (2008). The role of a crisis in reshaping the role of accounting. *Journal of Accounting and Public Policy*, *27*(6), 444–454. https://doi.org/10.1016/j.jaccpubpol.2008.09.002Get rights and content

Van den Boogaart, K. G. y Tolosana-Delgado, R. (2013). *Analyzing compositional data with R*. Springer.

Boonen, T., Guillén, M. y Santolino, M. (2019). Forecasting compositional risk allocations. *Insurance Mathematics and Economics*, *84*, 79–86. https://doi.org/10.1016/j.insmatheco.2018.10.002

Bresciani, S., Ferraris, A., Santoro, G. y Nilsen, H. R. (2016). Wine sector: companies' performance and green economy as a means of societal marketing. *Journal of Promotion Management*, *22*(2), 251–267. https://doi.org/10.1080/10496491.2016.1121753

Brown Sister, I. (1955). *The historical development of the use of ratios in financial statement analysis to 1933*. The Catholic University of America Press.

Buccianti, A., Mateu-Figueras, G. y Pawlowsky-Glahn, V. (2006). *Compositional data analysis in the geosciences: From theory to practice*. Geological Society of London.

Buchetti, B., Parbonetti, A. y Pugliese, A. (2022). Covid-19, corporate survival and public policy: The role of accounting information and regulation in the wake of a systemic crisis. *Journal of Accounting and Public Policy*, *41*(1), 106919. https://doi.org/10.1016/j.jaccpubpol.2021.106919

Buijink, W. y Jegers, M. (1986). Cross-sectional distributional properties of financial ratios in Belgian manufacturing industries: Aggregation effects and persistence over time. *Journal of Business Finance & Accounting*, *13*(3), 337–362. https://doi.org/10.1111/j.1468-5957.1986.tb00501.x

Caliński, T. y Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Calle, M. L., Pujolassos, M. y Susin, A. (2023). coda4microbiome: Compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinformatics*, *24*(1), 82. https://doi.org/10.1186/s12859-023-05205-3

Carreras-Simó, M. y Coenders, G. (2020). Principal component analysis of financial statements. A compositional approach. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, *29*, 18–37. https://doi.org/10.46661/revmetodoscuanteconempresa.3580

Carreras Simó, M. y Coenders, G. (2021). The relationship between asset and capital structure: A compositional approach with panel vector autoregressive models. *Quantitative Finance and Economics*, *5*(4), 571–590. https://doi.org/10.3934/QFE.2021025

Castillo Valero, J. S. y García Cortijo, M. D. C. (2013). Analysis of explanatory factors of profitability for wine firms in Castilla-La Mancha. *Revista de la Facultad de Ciencias Agrarias, UNCuyo*, *45*(2), 141–154. https://bdigital.uncu.edu.ar/6102

Chen, K. H. y Shimerda, T. A. (1981). An empirical analysis of useful financial ratios. *Financial Management*, *10*(1), 51–60. https://doi.org/10.2307/3665113

Chen, L., Wang, S. y Qiao, Z. (2014). DuPont model and product profitability analysis based on activity-based costing and economic value added. *European Journal of Business and Management*, *6*(30), 25–35. https://citeseerx.ist.psu.edu/document?doi=5ef35b903c5e9038b80ceaeee8a42ebfc433b6a4

Ciompa, P. (1910). *Grundrisse einer Oeconometrie und die auf Nationalökonomie aufgebaute natürliche Theorie der Buchhaltung. Ein auf Grund neuer ökonometrischer Gleichungen erbrachter Beweis, dass alle heutigen Bilanzen falsch dargestellt warden*. Verlag des Handelschulvereines in Lemberg.

Coenders, G. (2025). Application aux ratios financiers. En F. Bertrand, A. Gégout-Petit y C. Thomas-Agnan (Eds.), *Données de composition* (pp. 227-244). Éditions TECHNIP.

Coenders, G. y Arimany-Serrat, N. (2023). Accounting statement analysis at industry level. A gentle introduction to the compositional approach. *arXiv*, 2305.16842. https://arxiv.org/abs/2305.16842

Coenders, G., Egozcue, J. J., Fačevicová, K., Navarro-López, C., Palarea-Albaladejo, J., Pawlowsky-Glahn, V. y Tolosana-Delgado, R. (2023b). 40 years after Aitchison's article «The statistical analysis of compositional data». Where we are and where we are heading. *SORT. Statistics and Operations Research Transactions*, *47*(2), 207–228. https://doi.org/10.57645/20.8080.02.6

Coenders, G. y Ferrer-Rosell, B. (2020). Compositional data analysis in tourism. Review and future directions. *Tourism Analysis*, *25*(1), 153–168. https://doi.org/10.3727/108354220X15758301241594

Coenders, G. y Greenacre, M. (2023). Three approaches to supervised learning for compositional data with pairwise logratios. *Journal of Applied Statistics*, *50*(16), 3272–3293. https://doi.org/10.1080/02664763.2022.2108007

Coenders, G., Martín-Fernández, J. A. y Ferrer-Rosell, B. (2017). When relative and absolute information matter. Compositional predictor with a total in generalized linear models. *Statistical Modelling*, *17*(6), 494–512. https://doi.org/10.1177/1471082X17710398

Coenders, G. y Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT. Statistics and Operations Research Transactions*, *44*(1), 201–220. https://doi.org/10.2436/20.8080.02.100

Coenders, G., Sgorla, A. F., Arimany-Serrat, N., Linares-Mustarós, S. y Farreras-Noguer, M. À. (2023a). Nuevos métodos estadísticos composicionales para el análisis de ratios contables. *Revista de Comptabilitat i Direcció*, *35*, 135–148. https://accid.org/wp-content/uploads/2024/08/Nous-metodes-estadistics-composicionals_watermark.pdf

Comas-Cufí, M. y Thió-Henestrosa, S. (2011). CoDaPack 2.0: A stand-alone, multi-platform compositional software. En J. J. Egozcue, R. Tolosana-Delgado y M. I. Ortego (Eds.), *CoDaWork'11: 4th international workshop on compositional data analysis. Sant Feliu de Guíxols* (pp. 1–10). Universitat de Girona.

Cowen, S. S. y Hoffer, J. A. (1982). Usefulness of financial ratios in a single industry. *Journal of Business Research*, *10*(1), 103–118. https://doi.org/10.1016/0148-2963(82)90020-0

Creixans-Tenas, J., Coenders, G. y Arimany-Serrat, N. (2019). Corporate social responsibility and financial profile of Spanish private hospitals. *Heliyon*, *5*(10), e02623. https://doi.org/10.1016/j.heliyon.2019.e02623

Dale, E., Greenwood, R. S. y Greenwood, R. G. (1980). Donaldson Brown: GM's pioneer management theorist and practioner. *Academy of Management Proceedings*, *1980*(1), 119–123. https://doi.org/10.5465/ambpp.1980.4976162

Dao, B. T. T., Coenders, G., Lai, P. H., Dam, T. T. y Trinh, H. T. (2024). An empirical examination of financial performance and distress profiles during Covid-19: The case of fishery and food production firms in Vietnam. *Journal of Financial Reporting and Accounting*. https://doi.org/10.1108/JFRA-09-2023-0509

Davis, B. C., Hmieleski, K. M., Webb, J. W. y Coombs, J. E. (2017). Funders' positive affective reactions to entrepreneurs' crowdfunding pitches: The influence of perceived product creativity and entrepreneurial passion. *Journal of Business Venturing*, *32*(1), 90–106. https://doi.org/10.1016/j.jbusvent.2016.10.0060883-9026

Deakin, E. B. (1976). Distributions of financial accounting ratios: Some empirical evidence. *The Accounting Review*, *51*(1), 90–96. https://www.jstor.org/stable/245375

Demiraj, R., Labadze, L., Dsouza, S., Demiraj, E. y Grigolia, M. (2024). The quest for an optimal capital structure: An empirical analysis of European firms using GMM regression analysis. *EuroMed Journal of Business*. https://doi.org/10.1108/EMJB-07-2023-0206

Deshpande, A. (2023). The effect of financial leverage on firm profitability and working capital management in the Asia-Pacific Region. *Central European Review of Economics and Management (CEREM)*, *7*(4), 43–71. https://doi.org/10.29015/cerem.977

Dolnicar, S., Grün, B. y Leisch, F. (2018). *Market segmentation analysis: Understanding it, doing it, and making it useful*. Springer Nature.

Durana, P., Kovalova, E., Blazek, R. y Bicanovska, K. (2025). Business efficiency: Insights from Visegrad four before, during, and after the COVID-19 pandemic. *Economies*, *13*(2), 26. https://doi.org/10.3390/economies13020026

Egozcue, J. J. y Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, *37*, 795–828. https://doi.org/10.1007/s11004-005-7381-9

Egozcue, J. J. y Pawlowsky-Glahn, V. (2016). What are compositional data and how should they be analyzed. *Boletín de Estadística e Investigación Operativa*, *32*(1), 5–29. https://www.seio.es/beio/BEIOVol32Num1.pdf

Egozcue, J. J. y Pawlowsky-Glahn V. (2019). Compositional data: The sample space and its structure. *TEST*, 28(3), 599–638. https://doi.org/10.1007/s11749-019-00670-6

Egozcue, J. J., Pawlowsky-Glahn, V., Daunis-i-Estadella, J., Hron, K. y Filzmoser, P. (2012). Simplicial regression. The normal model. *Journal of Applied Probability and Statistics*, *6*(1–2), 87–108.

Egozcue, J. J., Pawlowsky-Glahn, V. Mateu-Figueras, G. y Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, *35*(3), 279–300. https://doi.org/10.1023/A:1023818214614

Escaramís, G. y Arbussà, A. (2025). Considerations on the use of financial ratios in the study of family businesses. *arXiv*, 2501.16793. https://arxiv.org/abs/2501.16793

Ezzamel, M. y Mar-Molinero, C. (1990). The distributional properties of financial ratios in UK manufacturing companies. *Journal of Business Finance & Accounting*, *17*(1), 1–29. https://doi.org/10.1111/j.1468-5957.1990.tb00547.x

Faello, J. (2015). Understanding the limitations of financial ratios. *Academy of Accounting and Financial Studies Journal*, *19*(3), 75–85.

Feranecová, A. y Krigovská, A. (2016). Measuring the performance of universities through cluster analysis and the use of financial ratio indexes. *Economics & Sociology*, *9*(4), 259–271. https://doi.org/10.14254/2071-789X.2016/9-4/16

Ferrer-Rosell, B. y Coenders, G. (2018). Destinations and crisis. Profiling tourists' budget share from 2006 to 2012. *Journal of Destination Marketing & Management*, *7*, 26–35. https://doi.org/10.1016/j.jdmm.2016.07.002

Ferrer-Rosell, B., Coenders, G. y Martín-Fuentes, E. (2022). Compositional data analysis in e-tourism research. En X. Zheng, M. Fuchs, U. Gretzel y W. Höpken (Eds.), *Handbook of e-tourism* (pp. 893–917). Springer.

Ferrer-Rosell, B., Martín-Fuentes, E., Vives-Mestres, M. y Coenders, G. (2021). When size does not matter: Compositional data analysis in marketing

research. En R. Nunkoo, V. Teeroovengadum y C. M. Ringle (Eds.), *Handbook of research methods for marketing management* (pp. 73–90). Edward Elgar.

Filzmoser, P., Hron, K. y Templ, M. (2018). *Applied compositional data analysis with worked examples in R*. Springer.

Fiori, A. M. y Coenders, G. (2025). Turning points in core-periphery displacement of systemic risk in the Eurozone: Constrained weighted compositional clustering. *Risks*, *13*(1), 21. https://doi.org/10.3390/risks13020021

Fiori, A. M. y Porro, F. (2023). A compositional analysis of systemic risk in European financial institutions. *Annals of Finance*, *19*, 325–354. https://doi.org/10.1007/s10436-023-00427-0

Frecka, T. J. y Hopwood, W. S. (1983). The effects of outliers on the cross-sectional distributional properties of financial ratios. *Accounting Review*, *58*(1), 115–128. https://doi.org/10.1016/j.adiac.2017.10.003

Fry, T. (2011). Applications in economics. En V. Pawlowsky-Glahn y A. Buccianti (Eds.), *Compositional data analysis. Theory and applications* (pp. 318–326). Wiley.

Fry, J. M., Fry, T. R. L. y McLaren, K. R. (1996). The stochastic specification of demand share equations: Restricting budget shares to the unit simplex. *Journal of Econometrics*, *73*(2), 377–385. https://doi.org/10.1016/S0304-4076(95)01727-5

Fry, J. M., Fry, T. R. L. y McLaren, K. R. (2000). Compositional data analysis and zeros in micro data. *Applied Economics*, *32*(8), 953–959. https://doi.org/10.1080/000368400322002

Fry, J. M., Fry, T. R. L., McLaren, K. R. y Smith, T. N. (2001). Modelling zeroes in microdata. *Applied Economics*, *33*(3), 383–392. https://doi.org/10.1080/00036840122916

Gabriel, K. R. (1971). The biplot-graphic display of matrices with application to principal component analysis. *Biometrika*, *58*, 453–467. https://doi.org/10.1093/biomet/58.3.453

Gámez-Velázquez, D. y Coenders, G. (2020). Identification of exchange rate shocks with compositional data and written press. *Finance, Markets and Valuation*, *6*(1), 99–113. https://doi.org/10.46503/LDAW9307

Gan, G. y Valdez, E. A. (2021). Compositional data regression in insurance with exponential family PCA. *arXiv*, 2112.14865. https://arxiv.org/abs/2112.14865

Giacomelli, S., Mocetti, S. y Rodano, G. (2021). *Fallimenti d'impresa in epoca Covid. Note Covid-19.* Banca d'Italia.

Glassman, D. A. y Riddick, L. A. (1996). Why empirical international portfolio models fail: Evidence that model misspecification creates home asset bias. *Journal of International Money and Finance*, *15*(2), 275–312. https://doi.org/10.1016/0261-5606(95)00046-1

Gokhale, S., Blomquist, J., Lindegren, M., Richter, A. y Waldo, S. (2024). The role of non-fishing and partner incomes in managing fishers' economic risk. *Marine Resource Economics*, *39*(4). https://doi.org/10.1086/731762

Gourinchas, P. O., Kalemli-Özcan, Ş., Penciakova, V. y Sander, N. (2020). Covid-19 and SME failures. *NBER Working Paper Series*, 27877. https://www.nber.org/system/files/working_papers/w27877

Greenacre, M. (2018). *Compositional data analysis in practice*. Chapman and Hall / CRC Press.

Greenacre, M. (2019). Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences*, *51*(5), 649–682. https://doi.org/10.1007/s11004-018-9754-x

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A. y Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I. y Quinn, T. (2023). Aitchison's compositional data analysis 40 years on: A reappraisal. *Statistical Science*, *38*(3), 386–410. https://doi.org/10.1214/22-STS880

Gruszczyński, M. (2022). Accounting and econometrics: From Paweł Ciompa to contemporary research. *Journal of Risk and Financial Management*, *15*(11), 510. https://doi.org/10.3390/jrfm15110510

Gupta, V. (2024). Evaluating the impact of geopolitical risk on the financial distress of Indian hospitality firms. *Journal of Risk and Financial Management*, *17*(12), 535. https://doi.org/10.3390/jrfm17120535

Hazami-Ammar, S. (2024). Related party transactions and financial distress: Role of governance and audit attributes. *Journal of Accounting & Organizational Change*. https://doi.org/10.1108/JAOC-03-2024-0105

Horrigan, J. O. (1968). A short history of financial ratio analysis. *The Accounting Review*, *43*(2), 284–294. https://www.jstor.org/stable/243765

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417-441. https://doi.org/10.1037/h0071325

Hron, K., Coenders, G., Filzmoser, P., Palarea-Albaladejo, J., Faměra, M. y Matys-Grygar, T. (2021). Analysing pairwise logratios revisited. *Mathematical Geosciences*, *53*(7), 1643–1666. https://doi.org/10.1007/s11004-021-09938-w

Hron, K., Filzmoser, P. y Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, *39*(5), 1115–1128. https://doi.org/10.1080/02664763.2011.644268

Iotti, M., Ferri, G. y Bonazzi, F. (2024a). Financial ratios, credit risk and business strategy: Application to the PDO Parma ham sector in single production and non-single production firms. *Journal of Agriculture and Food Research*, *16*, 101122. https://doi.org/10.1016/j.jafr.2024.101122

Iotti, M., Ferri, G., Manghi, E., Calugi, A. y Bonazzi, G. (2024b). Sustainability assessment of the performance of parmigiano reggiano PDO firms: A comparative analysis of firms' legal form and altitude range. *Sustainability*, *16*(20), 9093. https://doi.org/10.3390/su16209093

Iotti, M., Manghi, E. y Bonazzi, G. (2023). Financial performance of companies associated with the PDO Parma ham consortium: Analysis by quartile of firms. *Journal of Agriculture and Food Research*, *13*, 100598. https://doi.org/10.1016/j.jafr.2023.100598

Iotti, M., Manghi, E. y Bonazzi, G. (2024c). Debt sustainability assessment in the biogas sector: Application of interest coverage ratios in a sample of agricultural firms in Italy. *Energies*, *17*(6), 1404. https://doi.org/10.3390/en17061404

Isles, P. D. F. (2020). The misuse of ratios in ecological stoichiometry. *Ecology*, *101*, e03153. https://doi.org/10.1002/ecy.3153

Jantyik, L., Balogh, J. M. y Török, Á. (2021). What are the reasons behind the economic performance of the Hungarian beer industry? The case of the Hungarian microbreweries. *Sustainability*, *13*(5), 2829. https://doi.org/10.3390/su13052829

Jofre-Campuzano, P. y Coenders, G. (2022). Compositional classification of financial statement profiles. The weighted case. *Journal of Risk and Financial Management*, *15*(12), 546. https://doi.org/10.3390/jrfm15120546

Joueid, A. y Coenders, G. (2018). Marketing innovation and new product portfolios. A compositional approach. *Journal of Open Innovation: Technology, Market and Complexity*, *4*, 19. https://doi.org/10.3390/joitmc4020019

Kane, G. D., Richardson, F. M. y Meade, N. L. (1998). Rank transformations and the prediction of corporate failure. *Contemporary Accounting Research*, *15*(2), 145–166. https://doi.org/10.1111/j.1911-3846.1998.tb00553.x

Kaufman, L. y Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Keasey, K. y Watson, R. (1991). Financial distress models: a review of their usefulness. *British Journal of Management*, *2*(2), 89–102. https://doi.org/10.1111/j.1467-8551.1991.tb00019.x

Kokoszka, P., Miao, H., Petersen, A. y Shang, H. L. (2019). Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting*, *35*(4), 1304–1317. https://doi.org/10.1016/j.ijforecast.2019.05.007

Latief, G. E. y Suhendah, R. (2023). Financial performance comparative analysis between consumer goods and real-estate companies before and during the covid-19 pandemic. *International Journal of Application on Economics and Business*, *1*(2), 509–520. https://journal.untar.ac.id/index.php/ijaeb/article/view/25634/15392

Lev, B. y Sunder, S. (1979). Methodological issues in the use of financial ratios. *Journal of Accounting and Economics*, *1*(3), 187–210. https://doi.org/10.1016/0165-4101(79)90007-7

Li, Y., Han, H. y Li, Y. (2019). A new HHT-based denoising algorithm for financial time series data mining. En *2019 IEEE 8th joint international information technology and artificial intelligence conference (ITAIC) Chongqing, China* (pp. 397–401). Institute of Electrical and Electronic Engineers. https://doi.org/10.1109/ITAIC.2019.8785616

Linares-Mustarós, S., Coenders, G. y Vives-Mestres, M. (2018). Financial performance and distress profiles. From classification according to financial ratios to compositional classification. *Advances in Accounting*, *40*, 1–10. https://doi.org/10.1016/j.adiac.2017.10.003

Linares-Mustarós, S., Farreras-Noguer, M. A., Arimany-Serrat, N. y Coenders, G. (2022). New financial ratios based on the compositional data methodology. *Axioms*, *11*(12), 694. https://doi.org/10.3390/axioms11120694

Liu, X., Yang, X., Cao, J. y Huang, C. (2025). Extreme temperature shocks and firms' financial distress. *International Review of Economics & Finance*, *103946*. https://doi.org/10.1016/j.iref.2025.103946

Lueg, R., Punda, P. y Burkert, M. (2014). Does transition to IFRS substantially affect key financial ratios in shareholder-oriented common law regimes? Evidence from the UK. *Advances in Accounting*, *30*(1), 241–250. https://doi.org/10.1016/j.adiac.2014.03.002

Lukason, O. y Laitinen, E. K. (2019). Firm failure processes and components of failure risk: An analysis of European bankrupt firms. *Journal of Business Research*, *98*, 380–390. https://doi.org/10.1016/j.jbusres.2018.06.025

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En L. Lecam y J. Neyman (Eds), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1* (pp. 281–297). University of California Press.

Maldonado, W. L., Egozcue, J. J. y Pawlowsky-Glahn, V. (2021a). No-arbitrage matrices of exchange rates: Some characterizations. *International Journal of Economic Theory*, *17*, 375–389. https://doi.org/10.1111/ijet.12249

Maldonado, W. L., Egozcue, J. J. y Pawlowsky-Glahn, V. (2021b). Compositional analysis of exchange rates. En A. Daouia y A. Ruiz-Gazen (Eds.), *Advances in contemporary statistics and econometrics. Festschrift in honor of Christine Thomas-Agnan* (pp. 489–507). Springer.

Mariadassou, M. y Coenders, G. (2025). Traitement des zéros. En F. Bertrand, A. Gégout-Petit y C. Thomas-Agnan (Eds.), *Données de composition*. (pp. 69-84). Éditions TECHNIP.

Martikainen, T., Perttunen, J., Yli-Olli, P. y Gunasekaran, A. (1995). Financial ratio distribution irregularities: Implications for ratio classification. *European Journal of Operational Research*, *80*(1), 34–44. https://doi.org/10.1016/0377-2217(93)E0134-J

Martín-Fernández, J. A. (2019). Comments on: Compositional data: The sample space and its structure. *TEST*, *28*(3), 653–657. https://doi.org/10.1007/s11749-019-00672-4

Martín-Fernández, J. A., Barceló-Vidal, C. y Pawlowsky-Glahn, V. (1998). A critical approach to non-parametric classification of compositional data. En A. Rizzi, M. Vichi y H. H. Bock (Eds.), *Advances in data science and classification* (pp. 49–56). Springer.

Martín-Fernández, J. A., Barceló-Vidal, C. y Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, *35*, 253–278. https://doi.org/10.1023/A:1023866030544

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P. y Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics & Data Analysis*, *56*(9), 2688–2704. https://doi.org/10.1016/j.csda.2012.02.012

Martín-Fernández, J. A., Palarea-Albaladejo, J. y Olea, R. A. (2011). Dealing with zeros. En V. Pawlowsky-Glahn y A. Buccianti (Eds.), *Compositional data analysis. Theory and applications* (pp. 47–62). Wiley.

Martínez-García, A., Horrach-Rosselló, P. y Mulet-Forteza, C. (2023). Mapping the intellectual and conceptual structure of research on CoDa in the «Social Sciences» scientific domain. A bibliometric overview. *Journal of Geochemical Exploration*, *252*, 107273. https://doi.org/10.1016/j.gexplo.2023.107273

McLaren, K. R., Fry, J. M. y Fry, T. R. L. (1995). A simple nested test of the almost ideal demand system. *Empirical Economics*, *20*(1), 149–161. https://doi.org/10.1007/BF01235162

McLeay, S. (1986). Student's t and the distribution of financial ratios. *Journal of Business Finance and Accounting*, *13*(2), 209–222. https://doi.org/10.1111/j.1468-5957.1986.tb00091.x

McLeay, S. y Omar, A. (2000). The sensitivity of prediction models to the non-normality of bounded and unbounded financial ratios. *The British Accounting Review*, *32*(2), 213–230. https://doi.org/10.1006/bare.1999.0120

Minnis, M. y Shroff, N. (2017). Why regulate private firm disclosure and auditing? *Accounting and Business Research*, *47*(5), 473–502. https://doi.org/10.1080/00014788.2017.1303962

Molas-Colomer, X., Linares-Mustarós, S., Farreras-Noguer, M. À. y Ferrer-Comalat, J. C. (2024). A new methodological proposal for classifying firms according to the similarity of their financial structures based on combining compositional data with fuzzy clustering. *Journal of Multiple-Valued Logic and Soft Computing*, *43*(1–2), 73–100. https://www.oldcitypublishing.com/wp-content/uploads/2024/05/MVLSCv43n1-2p73-100Molas-Colomer.pdf

Mulet-Forteza, C., Ferrer-Rosell, B., Martorell-Cunill, O. y Linares-Mustarós, S. (2024). The role of expansion strategies and operational attributes on hotel performance: A compositional approach. *arXiv*, 2411.04640. https://arxiv.org/abs/2411.04640

Naz, A., Bhutta, A. I., Sheikh, M. F. y Sultan, J. (2023). Corporate real estate investment and firm performance: Empirical evidence from listed non financial firms of Pakistan. *Journal of Corporate Real Estate*, *25*(3), 246–262. https://doi.org/10.1108/JCRE-05-2022-0013

Nyitrai, T. y Virág, M. (2019). The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences*, *67*, 34–42. https://doi.org/10.1016/j.seps.2018.08.004

Oktaviano, B., Wulandari, D. S. y Rasidi, A. B. (2024). Does firm size buffer tax aggressiveness? Examining financial distress and capital intensity. *International Journal of Scientific Multidisciplinary Research*, *2*(10), 1591–1608. https://doi.org/10.55927/ijsmr.v2i10.12082

Ortells, R., Egozcue, J. J., Ortego, M. I. y Garola, A. (2016). Relationship between popularity of key words in the Google browser and the evolution of worldwide financial indices. En J. A. Martín-Fernández y S. Thió-Henestrosa (Eds.), *Compositional data analysis. Springer proceedings in mathematics & statistics. Vol. 187* (pp. 145–166). Springer.

Palarea-Albaladejo, J. y Martín-Fernández, J. A. (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, *34*(8), 902–917. https://doi.org/10.1016/j.cageo.2007.09.015

Palarea-Albaladejo, J. y Martín-Fernández, J. A. (2015). zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, *143*, 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019

Pawlowsky-Glahn, V. y Egozcue, J. J. (2011). Exploring compositional data with the CoDa-dendrogram. *Austrian Journal of Statistics*, *40*(1–2), 103–113. https://doi.org/10.17713/ajs.v40i1&2.202

Pawlowsky-Glahn, V., Egozcue, J. J. y Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Wiley.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, *60*, 489–498. https://doi.org/10.1098/rspl.1896.0076

Pohlman, R. A. y Hollinger, R. D. (1981). Information redundancy in sets of financial ratios. *Journal of Business Finance & Accounting*, *8*(4), 511–528. https://doi.org/10.1111/j.1468-5957.1981.tb00832.x

Porro, F. (2022). A geographical analysis of the systemic risk by a compositional data (CoDa) approach. En M. Corazza, C. Perna, C. Pizzi y M. Sibillo (Eds.), *Mathematical and statistical methods for actuarial sciences and finance* (pp. 383–389). Springer.

Qin, Z., Hassan, A. y Adhikariparajuli, M. (2022). Direct and indirect implications of the COVID-19 pandemic on Amazon's financial situation. *Journal of Risk and Financial Management*, *15*(9), 414. https://doi.org/10.3390/jrfm15090414

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org

Rondós-Casas, E., Linares-Mustarós, S. y Farreras-Noguer, M. À. (2018). Expansion of the current methodology for the study of the short-term liquidity problems in a sector. *Intangible Capital*, *14*(1), 25–34. https://doi.org/10.3926/ic.1085

Ross, S. A., Westfield, R. W. y Jordan, B. D. (2003). *Fundamentals of corporate finance. (Vol. 1)* (6.ª ed.). McGraw-Hill.

Saus-Sala, E., Farreras-Noguer, M. À., Arimany-Serrat N. y Coenders, G. (2021). Compositional DuPont analysis. A visual tool for strategic financial performance assessment. En P. Filzmoser, K. Hron, J. A. Martín-Fernández y J. Palarea-Albaladejo (Eds.), *Advances in compositional data analysis. Festschrift in honour of Vera Pawlowsky-Glahn* (pp. 189–206). Springer.

Saus-Sala, E., Farreras-Noguer, M. À., Arimany-Serrat, N. y Coenders, G. (2023). Análisis de las empresas de turismo rural en Cataluña y Galicia: Rentabilidad económica y solvencia 2014–2018. *Cuadernos del CIMBAGE*, *25*(1), 33–54. https://doi.org/10.56503/cimbage/vol.1/nro.25(2023)p.33-54

Saus-Sala, E., Farreras-Noguer, M. À., Arimany-Serrat, N. y Coenders, G. (2024). Financial analysis of rural tourism in Catalonia and Galicia pre- and post COVID-19. *International Journal of Tourism Research*, *26*(4), e2698. https://doi.org/10.1002/jtr.2698

Sharma, S., Shebalkov, M. y Yukhanaev, A. (2016). Evaluating banks performance using key financial indicators–a quantitative modeling of Russian banks. *The Journal of Developing Areas*, *50*(1), 425–453. https://www.jstor.org/stable/24737357

So, J. C. (1987). Some empirical evidence on the outliers and the non-normal distribution of financial ratios. *Journal of Business Finance & Accounting*, *14*(4), 483–496. https://doi.org/10.1111/j.1468-5957.1987.tb00108.x

Sozen, E., Rahman, I. y O'Neill, M. (2022). Craft breweries' environmental proactivity: An upper echelons perspective. *International Journal of Wine Business Research*, *34*(2), 237–256. https://doi.org/10.1108/IJWBR-02-2021-0013

Soukal, I., Mačí, J., Trnková, G., Svobodova, L., Hedvičáková, M., Hamplova, E., Maresova, P. y Lefley, F. (2024). A state-of-the-art appraisal of bankruptcy prediction models focussing on the field's core authors: 2010–2022. *Central European Management Journal*, *32*(1), 3–30. https://doi.org/10.1108/CEMJ-08-2022-0095

Staňková, M. y Hampel, D. (2023). Optimal threshold of data envelopment analysis in bankruptcy prediction. *SORT. Statistics and Operations Research Transactions*, *47*(1), 129–150. https://doi.org/10.57645/20.8080.02.3

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680. https://doi.org/10.1126/science.103.2684.677

Sunder, S. (2016). Better financial reporting: Meanings and means. *Journal of Accounting and Public Policy*, *35*(3), 211–223. https://doi.org/10.1016/j.jaccpubpol.2016.03.002

Tallapally, P. (2009). *The association between data intermediaries and bond rating classification model prediction accuracy*. Unpublished Doctoral Dissertation. Louisiana Tech University.

Tascón, M. T., Castaño, F. J. y Castro, P. (2018). A new tool for failure analysis in small firms: Frontiers of financial ratios based on percentile differences (PDFR). *Spanish Journal of Finance and Accounting*, *47*(4), 433–463. https://doi.org/10.1080/02102412.2018.1468058

Thió-Henestrosa, S. y Martín-Fernández, J. A. (2005). Dealing with compositional data: The freeware CoDaPack. *Mathematical Geology*, *37*(7), 773–793. https://doi.org/10.1007/s11004-005-7379-3

Tian, Y., Ali, M. K. M. y Wu, L. (2024). The application of adaptive group LASSO imputation method with missing values in personal income compositional data. *Journal of Big Data*, *11*, 166. https://doi.org/10.1186/s40537-024-01009-1

Todorov, V. y Simonacci, V. (2020). Three-way compositional analysis of energy intensity in manufacturing. En *Book of short papers. 50th scientific meeting*

of the Italian Statistical Society, Pisa, 22-24 June 2020 (pp. 111–116). Pearson Education Italia.

Tolosana-Delgado, R. y Van den Boogaart, K. G. (2011). Linear models with compositions in R. En V. Pawlowsky-Glahn y A. Buccianti (Eds.), *Compositional data analysis: Theory and applications* (pp. 356–371). Wiley.

Trejo-Pech, C. J., DeLong, K. L. y Johansson, R. (2023). How does the financial performance of sugar-using firms compare to other agribusinesses? An accounting and economic profit rates analysis. *Agricultural Finance Review*, *83*(3), 453–477. https://doi.org/10.1108/AFR-08-2022-0103

Valaskova, K., Gajdosikova, D. y Lazaroiu, G. (2023). Has the COVID-19 pandemic affected the corporate financial performance? A case study of Slovak enterprises. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, *18*(4), 1133–1178. https://doi.org/10.24136/eq.2023.036

Vega-Baquero, J. D. y Santolino, M. (2022a). Capital flows in integrated capital markets: MILA case. *Quantitative Finance and Economics*, *6*(4), 622–639. https://doi.org/10.3934/QFE.2022027

Vega-Baquero, J. D. y Santolino, M. (2022b). Too big to fail? An analysis of the Colombian banking system through compositional data. *Latin American Journal of Central Banking*, *3*(2), 100060. https://doi.org/10.1016/j.latcb.2022.100060

Vega-Gámez, F. y Alonso-González, P. J. (2024). How likely is it to beat the target at different investment horizons: An approach using compositional data in strategic portfolios. *Financial Innovation*, *10*(1), 125. https://doi.org/10.1186/s40854-023-00601-3

Veganzones, D. y Severin, E. (2021). Corporate failure prediction models in the twenty-first century: A review. *European Business Review*, *33*(2), 204–226. https://doi.org/10.1108/EBR-12-2018-0209

Verbelen, R., Antonio, K. y Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *67*(5), 1275–1304. https://doi.org/10.1111/rssc.12283

De Vito, A. y Gómez, J. P. (2020). Estimating the COVID-19 cash crunch: Global evidence and policy. *Journal of Accounting and Public Policy*, *39*(2), 106741. https://doi.org/10.1016/j.jaccpubpol.2020.106741

Vizcaino, D., Domínguez, A. y Sosa, J. (2020). *Importancia económica y social del sector vitivinícola en España*. Organización Interprofesional del vino en España. https://www.agro-alimentarias.coop/ficheros/doc/06306.pdf

Voltes-Dorta, A., Jiménez, J. L. y Suárez-Alemán, A. (2014). An initial investigation into the impact of tourism on local budgets: A comparative analysis of Spanish municipalities. *Tourism Management*, *45*, 124–133. https://doi.org/10.1016/j.tourman.2014.02.016

Vu, N. T., Nguyen, N. H., Tran, T., Le, B. T. y Vo, D. H. (2023). A LASSO-based model for financial distress of the Vietnamese listed firms: Does the covid-19 pandemic matter? *Cogent Economics & Finance*, *11*(1), 2210361. https://doi.org/10.1080/23322039.2023.2210361

Wang, H., Lu, S. y Zhao, J. (2019). Aggregating multiple types of complex data in stock market prediction: A model-independent framework. *Knowledge-Based Systems*, *164*(15), 193–204. https://doi.org/10.1016/j.knosys.2018.10.035

Ward jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Watson, C. J. (1990). Multivariate distributional properties, outliers, and transformation of financial ratios. *The Accounting Review*, *65*(3), 682–695. https://www.jstor.org/stable/247957

Willer do Prado, J., De Castro Alcântara, V., De Melo Carvalho, F., Carvalho Vieira, K., Cruz Machado, L. K. y Flávio Tonelli, D. (2016). Multivariate analysis of credit risk and bankruptcy research data: A bibliometric study involving different knowledge fields (1968–2014). *Scientometrics*, *106*(3), 1007–1029. https://doi.org/10.1007/s11192-015-1829-6

Zanotti, C., Reyes, F. y Fernández, B. (2018). Relationship between competitiveness and operational and financial performance of firms: An exploratory study on the European brewing industry. *Intangible Capital*, *14*(1), 1–17. https://doi.org/10.3926/ic.1104

Sobre los autores

Germà Coenders Gallart es doctor en ciencias de la gestión por ESADE (Universidad Ramon Llull, 1996) y catedrático en métodos cuantitativos para la economía y la empresa en la Universidad de Girona desde 2016. Sus principales intereses de investigación han sido los modelos de ecuaciones estructurales aplicados a errores de medida en encuestas (miembro fundador de la European Survey Research Association –ESRA–, editor asociado fundador de la revista *Survey Research Methods*) y el análisis de datos composicionales (miembro fundador de la Association for Compositional Data -CoDa-Association, de la que actúa como secretario general). Ha publicado más de ochenta artículos en revistas indexadas sobre los dos temas mencionados anteriormente, y específicamente su aplicación en gestión, economía, turismo y campos relacionados. Más precisamente, ha sido pionero en la aplicación del análisis composicional de datos en economía y empresa, con más de cuarenta publicaciones sobre este tema, metodológicas y también aplicadas, que cubren aspectos como la comunicación, la innovación en marketing, el análisis de la cartera de productos, la economía del turismo, el comportamiento de los turistas, la gestión del transporte, la segmentación de mercados, el *e-marketing*, los tipos de cambio, la educación gerencial y las finanzas corporativas. Desde 2018, centra su investigación en el análisis composicional de estados financieros como responsable de sendas líneas de investigación sobre el tema en los proyectos del Ministerio de Ciencia, Innovación y Universidades / FEDER-a way of making Europe RTI2018-095518-B-C21 y PID2021-123833OB-I00.

Núria Arimany Serrat es doctora en contabilidad y auditoría (Universidad de Barcelona, 2005) y profesora titular de contabilidad y finanzas en la Universidad de Vic-Universidad Central de Catalunya, desde 1990. Sus principales intereses de investigación son la contabilidad y las finanzas, y la responsabilidad social corporativa, junto al análisis integral (financiero y de sostenibilidad) en diferentes sectores de actividad. Es miembro de la Asociación Catalana de Contabilidad y Dirección (ACCID) y de la Agrupación de Profesorado de Contabilidad y Control (APC). Ha publicado más de 75 artículos en revistas indexadas sobre los temas mencionados anteriormente, y específicamente su aplicación en gestión integral, análisis de los estados financieros y análisis según criterios ESG, seguimiento de los ODS de la Agenda 2030 y el Pacto Verde Europeo de 2019. Ha participado en diversos proyectos competitivos y de transferencia de conocimiento y ha dirigido diversas tesis defendidas y en curso. Ha intervenido en diferentes estudios sobre análisis de los estados financieros sectoriales utilizando el análisis composicional, con más de diez publicaciones del ámbito que tratan aspectos como la comunicación; el turismo rural; la salud del sector apícola, de las empresas hospitalarias, del sector del vino, del sector de la cerveza, etc. Desde 2018, como miembro de los proyectos del Ministerio de Ciencia, Innovación y Universidades / FEDER-a way of making Europe RTI2018-095518-B-C21 y PID2021-123833OB-I00, trabaja con la metodología CoDa en la presentación de la información financiera y no financiera de forma integrada.

Documenta
Universitaria
@DocUniv
documentauniversitaria.com

El análisis de los estados financieros por medio de ratios es una herramienta poderosa de diagnóstico empresarial que se ha venido utilizando desde el último tercio del siglo XIX. En el momento que se desarrolló, el análisis estadístico estaba en su infancia, y la teoría de las escalas de medición que determina las operaciones matemáticas y estadísticas válidas sobre una ratio ni siquiera había nacido. Las ratios financieras, pues, nunca fueron concebidas para usarse como variables en análisis estadísticos, sino para analizar empresas individualmente.

A pesar de esto, a partir del último tercio del siglo XX las ratios empezaron a usarse profusamente en análisis estadísticos, aunque existían problemas graves de asimetría, no linealidad y observaciones atípicas, entre otros, de los que solo se era parcialmente consciente y a los que solo se propusieron soluciones parciales y *ad hoc*.

El presente libro presenta un enfoque integral al análisis estadístico de las ratios financieras llamado metodología de datos composicionales, que resuelve los mencionados problemas y se adapta a cualquier técnica estadística que analice ratios financieras, aisladamente o en compañía de otras variables no financieras.