Quaderns Agraris (Institució Catalana d'Estudis Agraris), núm. 45 (desembre 2018), p. 91-111
ISSN (ed. impresa): 0213-0319 / ISSN (ed. electrònica): 2013-9780
http://revistes.iec.cat/index.php/QA / DOI: 10.2436/20.1503.01.92

ANÀLISI DE DADES COMPOSICIONALS: CONCEPTES BÀSICS I EXEMPLES EN CIÈNCIES APLICADES I EN EL SECTOR AGROALIMENTARI

Eusebio Jarauta-Bragulat

Departament d'Enginyeria Civil i Ambiental,
Escola Tècnica Superior d'Enginyeria de Camins,
Canals i Ports de Barcelona,
Universitat Politècnica de Catalunya (UPC BarcelonaTech)

Rebut: 16 de maig de 2018 - Acceptat: 18 de maig de 2018

RESUM

Les dades composicionals (*compositional data*, CODA) són dades vectorials que descriuen les diferents parts d'un cert total. Habitualment, les dades composicionals es presenten com vectors de proporcions, percentatges, concentracions o freqüències. L'espai al qual pertanyen les dades composicionals s'anomena *símplex de* n *parts*, que es defineix com el conjunt de vectors de *n* components estrictament positives i tals que la suma d'aquestes components és constant. Atès que les proporcions s'expressen com nombres reals, hi ha la temptació d'interpretar, o fins i tot d'analitzar, les dades composicionals com si fossin dades reals multivariants. Aquesta pràctica pot conduir a paradoxes o a males interpretacions, com ara la correlació espúria i la paradoxa de Simpson.

En ciències aplicades i enginyeria tot sovint s'estudien processos dinàmics en els quals les variables evolucionen amb el temps. Un cas particular d'interès especial és l'estudi i la caracterització de processos en els quals les variables són composicionals i evolucionen amb el temps (o l'espai). Aquests processos són molt habituals en ciències agroalimentàries i biotecnològiques. En aquest tipus de processos, els sistemes estan representats per composicions i es modelitzen mitjançant funcions de valors en el símplex, definides en intervals de la recta real (temps, espai). En aquest treball

Correspondència: Eusebio Jarauta Bragulat. Departament d'Enginyeria Civil i Ambiental, Escola Tècnica Superior d'Enginyeria de Camins, Canals i Ports de Barcelona. C/ Jordi Girona, 1-3. Edifici C2, planta 3, despatx 303-A. 08034 Barcelona. A/e: *eusebi.jarauta@upc.edu*.

es presenten també els models diferencials lineals composicionals i la seva utilitat en la descripció i en l'estimació del comportament futur de les variables del sistema. Finalment, es comenten les conclusions més importants per tal que el treball amb aquest tipus de dades es pugui aplicar en enginyeria agroalimentària amb garanties de fiabilitat i que es garanteixin la formulació i la interpretació correctes dels resultats obtinguts.

Paraules clau: dades composicionals, composicions, símplex, geometria d'Aitchison, equacions diferencials composicionals, corbes de creixement, sistema alimentari.

ANÁLISIS DE DATOS COMPOSICIONALES: CONCEPTOS BÁSICOS Y EJEMPLOS EN CIENCIAS APLICADAS Y EN EL SECTOR AGROALIMENTARIO

RESUMEN

Los datos composicionales (*compositional data*, CODA) son datos vectoriales que describen las diferentes partes de un cierto total. Habitualmente, los datos composicionales se presentan como vectores de proporciones, porcentajes, concentraciones o frecuencias. El espacio al que pertenecen los datos composicionales se denomina *símplex de* n *partes*, que se define como el conjunto de vectores de *n* componentes estrictamente positivas y tales que la suma de estas componentes es constante. Dado que las proporciones se expresan como números reales, existe la tentación de interpretar, o incluso de analizar, los datos composicionales como si se tratara de datos reales multivariantes. Esta práctica puede conducir a paradojas o a malas interpretaciones tales como la correlación espuria y la paradoja de Simpson.

En ciencias aplicadas e ingeniería, se estudian a menudo procesos dinámicos en los que las variables evolucionan con el tiempo. Un caso particular de interés especial es el estudio y la caracterización de procesos en los que las variables son composicionales y evolucionan con el tiempo (o el espacio). Estos procesos son muy habituales en ciencias agroalimentarias y biotecnológicas. En este tipo de procesos, los sistemas están representados por composiciones y se modelizan mediante funciones de valores en el símplex, definidas en intervalos de la recta real (tiempo, espacio). En este trabajo se presentan también los modelos diferenciales lineales composicionales y su utilidad en la descripción y en la estimación del comportamiento futuro de las variables del sistema. Finalmente, se comentan las conclusiones más importantes para que el trabajo con este tipo de datos se pueda aplicar en ingeniería agroalimentaria con garantías de fiabilidad y que se garanticen la formulación y la interpretación correctas de los resultados obtenidos.

PALABRAS CLAVE: datos composicionales, composiciones, símplex, geometría de Aitchison, ecuaciones diferenciales composicionales, curvas de crecimiento, sistema alimentario.

COMPOSITIONAL DATA ANALYSIS: BASIC CONCEPTS AND EXAMPLES IN APPLIED SCIENCES IN THE AGRI-FOOD SECTOR

ABSTRACT

Compositional data (CODA) are vectors that describe the different parts of a certain total. Usually, compositional data are presented as vectors of proportions, percentages, concentrations or frequencies. The space to which compositional data belong is called a "simplex of *n* parts", which is defined as the set of vectors of *n* strictly positive components, such that the sum of these components is constant. Since the proportions are expressed as real numbers, there is a temptation to interpret or even analyse compositional data as if they were real multivariate data. This practice can lead to paradoxes or misinterpretations such as spurious correlation and Simpson's paradox.

In applied sciences and engineering, dynamic processes are often studied in which variables evolve over time. A special case of particular interest is the study and characterization of processes in which the variables are compositional and evolve over time (or space). These processes are very common in agri-food and biotechnological sciences. In this type of processes, the systems are represented by compositions and are modelled by value functions in the simplex, defined in intervals of the real line (time, space). This paper presents the compositional linear differential models and their usefulness in the description and estimation of the future behaviour of system variables. Finally, the most important conclusions are discussed so that work with this type of data can be applied in agri-food engineering with reliability guarantees and so that the correct formulation and interpretation of the results obtained is ensured.

KEYWORDS: compositional data, compositions, simplex, Aitchison geometry, compositional differential equations, growth curves, food system.

1. INTRODUCCIÓ I OBJECTIUS

S'anomenen *dades composicionals* les dades vectorials que contenen informació relativa de les diverses parts en què es considera dividit o classificat un cert total, és a dir, són vectors de components estrictament positives i

de suma constant; com es veurà, aquestes condicions són les que confereixen a les dades composicionals les seves propietats més importants. Les dades composicionals apareixen sempre que es treballa amb magnituds relatives com concentracions, proporcions, freqüències relatives, etc. Les unitats corresponents són, per exemple, grams per centímetre cúbic (g/cm^3), mil·ligrams per litre (mg/L), quilograms per hectàrea (kg/ha), percentatges (%), parts per unitat (ppu) o parts per milió (ppm) de massa o de volum, etc. Tot sovint el total no té cap interès; per exemple, si es vol descriure la composició química d'un sòl, no interessa el pes total de la mostra de sòl extreta, sinó el pes relatiu al total que hi ha de cada un dels elements que componen la mostra, cosa que, a més de la informació sobre la composició química del sòl, pot tenir una utilitat taxonòmica.

La recerca en dades composicionals té els orígens en la biologia (Pearson, 1897), però s'ha desenvolupat bàsicament en l'àmbit de les geociències i això ha motivat que sigui en aquest àmbit que històricament se'ls hagi prestat una atenció més significativa, que en gran part ha estat motivada pels problemes que sorgeixen en aplicar l'anàlisi estadística tradicional a dades de tipus composicional. En efecte, atès que les proporcions s'expressen com nombres reals, es produeix la temptació d'interpretar-les, o fins i tot d'analitzar-les, com si fossin dades reals multivariants, cosa que pot conduir a paradoxes, o bé a males interpretacions. Les més significatives són la correlació espúria i la paradoxa de Simpson, que es comentaran més endavant.

En aquest treball es pretén donar a conèixer els aspectes bàsics de l'anàlisi de dades composicionals, la problemàtica associada a aquestes dades i les eines metodològiques específiques per a la seva aplicació, sense entrar a fons en desenvolupaments matemàtics o estadístics més enllà dels bàsics per a entendre el plantejament i les metodologies. El que s'exposa aquí té validesa general en qualsevol àmbit científic en el qual es tractin dades composicionals, que són la immensa majoria; tanmateix, es farà esment, sempre que sigui possible, d'aplicacions en l'àmbit agroalimentari. Concretament, en aquest treball es planteja assolir els objectius següents: 1) definir els conceptes bàsics, la naturalesa, les operacions i la problemàtica associada al treball amb les dades composicionals, il·lustrant-ho amb alguns exemples; 2) presentar els principis que fonamenten el treball amb dades composicionals i les transformacions que permeten operar amb aquestes dades com vectors reals multivariants; 3) definir els processos composicionals i la seva caracterització, i 4) presentar una introducció a la formulació de models diferencials lineals composicionals elaborats amb el suport de dades experimentals.

2. CONCEPTES BÀSICS DE L'ANÀLISI DE DADES COMPOSICIONALS

2.1. Exemples introductoris

Exemple 1. Si es vol caracteritzar la situació econòmica del sector agrari pel que fa a la renda de la gent que hi treballa, es poden definir uns llindars de renda anual i establir una certa classificació en grups associats a la renda anual comparada amb aquests llindars. El nombre de grups considerats i els llindars de cada un poden obeir a criteris diferents i donar lloc, per tant, a aproximacions diferents d'aquesta descripció. En qualsevol cas, un cop establert un nombre determinat de grups i si es coneix el nombre de persones en cada grup, es pot calcular aleshores la proporció, expressada normalment en tant per cent, de persones en cada un dels grups. El vector que descriu o caracteritza la distribució de la renda d'aquest sector té tantes components com grups considerats i la suma d'aquestes components és sempre 100 % si les proporcions s'expressen en percentatge. Si es prefereix expressar les proporcions en parts per unitat, aleshores la suma de les components és 1.

Exemple 2. En ciències ambientals, s'estudia la contaminació de l'aire de les ciutats; es consideren unes certes substàncies (pol·luents) i s'expressa la concentració de cada una en un vector que té tantes components com el nombre de pol·luents que es considerin. Cada component conté la concentració del pol·luent respectiu expressada habitualment en $\mu g/m^3$ i, per tant, conté informació d'una part (massa del pol·luent considerat) amb relació a un cert total (volum d'aire). A partir d'aquestes concentracions, es defineix el que s'anomena *índex de qualitat de l'aire* (AQI per les seves sigles en anglès), emprant metodologies diverses. Un cop fixat un cert instant, es poden fer estudis comparatius de diverses zones d'una mateixa ciutat, o bé si es consideren les mesures en temps diversos, es pot estudiar l'evolució en el temps de la concentració de cada un dels pol·luents de l'aire i de l'índex de qualitat de l'aire i, amb la caracterització mitjançant un model d'evolució, fer-ne una estimació futura.

Exemple 3. En l'estudi de l'eficàcia d'un cert tractament per al bestiar condicionada al sexe (mascle o femella) de l'animal al qual s'aplica el tractament, el veterinari que l'aplica i el controla pot donar com a informació el nombre de mascles i de femelles als quals els ha anat bé el tractament i el nombre de mascles i de femelles als quals no els ha anat bé el tractament. Si, en canvi, la informació que dona el veterinari és la proporció de mascles i de femelles als quals els ha anat bé el tractament, l'altra proporció no cal donar-la de manera explícita, ja que resulta immediatament calculant el número complementari. Es veu clarament que, en aquest segon cas, la informació és una dada composicional, mentre que en el primer cas no ho és. Aprofundirem en aquest exemple més endavant.

2.2. El símplex de *n* parts. Operacions amb dades composicionals

Quan s'estudien processos en els quals són vàlids models unidimensionals, el conjunt de treball és el dels nombres reals $\mathbb{R}$. Quan es requereix que els models tinguin en compte més d'una variable, cal anar a l'espai bidimensional $\mathbb{R}^2$, l'espai tridimensional $\mathbb{R}^3$ o, en un marc més general, l'espai multidimensional $\mathbb{R}^n$. Un cas particular d'interès és quan es consideren variables que només poden ser positives (edat, alçada, renda, massa, etc.), fet que requereix treballar en els nombres reals positius $\mathbb{R}_+$ (models univariants) o $\mathbb{R}_+^n$ (models multivariants). El conjunt de vectors que permet expressar matemàticament les dades composicionals és l'anomenat *símplex de* n *parts*, que es defineix mitjançant:

$$\mathcal{S}_K^n = \left\{ \vec{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}_+^n : \; x_j > 0, \; j = 1, ..., n; \; \sum_{j=1}^{n} x_j = K \right\} \qquad (1)$$

El valor de la suma constant *K* de les parts d'una composició està associat a com s'expressen aquestes parts; així, si les parts són percentatges, es compleix $K = 100$; si les parts són parts per unitat, es compleix $K = 1$; si són parts per milió, es compleix $K = 10^6$, i així successivament. Observeu els dos elements clau d'aquesta definició: les parts són nombres *estrictament positius* i la *suma de les parts és constant*. El més habitual és expressar les parts d'una composició en parts per unitat, és a dir, treballar en el símplex S^n.

Per a obtenir una composició a partir d'un vector de components positives, simplement cal dividir per la suma de les components d'aquest vector, operació que es coneix amb el nom de *clausura*. Així, si es considera un vector de components positives

$$\vec{X} = (X_1, X_2, ..., X_n), \; X_j > 0, \; j = 1, 2, ..., n$$

aleshores la clausura d'aquest vector es calcula mitjançant:

$$\mathcal{C}\vec{X} = \mathcal{C}(X_1, X_2, ..., X_n) = \left(\frac{X_1}{\sum X_j}, \frac{X_2}{\sum X_j}, ..., \frac{X_n}{\sum X_j} \right) = (x_1, x_2, ..., x_n) \in S^n \qquad (2)$$

Per tant, es pot expressar una composició de dues maneres: directament si el vector compleix la condició de pertinença al símplex, o bé com la clausura d'un vector de components positives; per exemple, $x_1 = (0{,}2,\ 0{,}3,\ 0{,}5)$ i $x_2 = X(20,\ 30,\ 50)$ serien dues maneres diferents d'expressar la mateixa composició.

La peculiaritat de les composicions fa que no se'ls puguin aplicar les operacions ordinàries amb magnituds vectorials, com, per exemple, la suma i

el producte per un escalar. En efecte, si es consideren les composicions de S^3 x_1 = (0,2, 0,3, 0,5) i x_2 = (0,4, 0,3, 0,3), la suma ordinària d'aquestes composicions dona com a resultat (0,6, 0,6, 0,8), que no és un element del símplex S^3, és a dir, la suma vectorial ordinària de dues composicions no és una composició i, per tant, la suma vectorial ordinària no és una operació interna en el símplex. Passa una situació similar amb el producte d'una composició per un escalar real, ja que aquest producte no pertany al símplex i no es pot aplicar en composicions. Aquesta situació obliga a definir unes operacions adequades a la naturalesa de les composicions i que permetin dotar el símplex d'estructura d'espai vectorial; aquestes operacions s'anomenen *pertorbació*, designada per $\oplus$, i *potenciació*, designada per $\odot$, i es defineixen així, respectivamente:

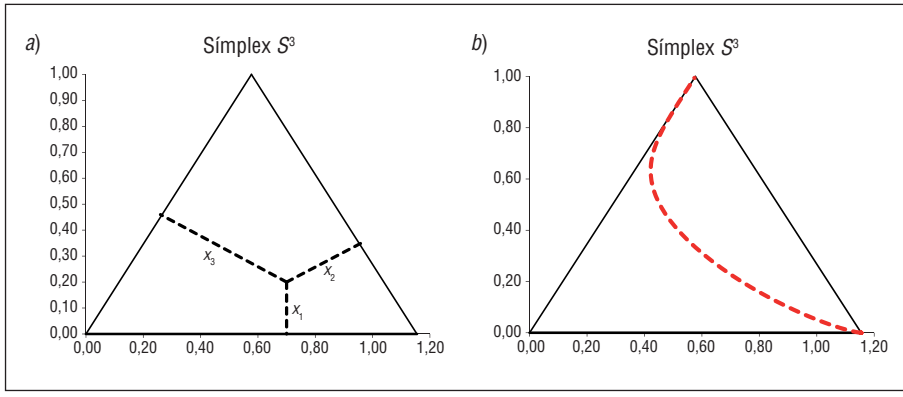
$$\begin{aligned} \vec{x} \oplus \vec{y} &= \mathcal{C}\left(x_1 y_1, \ldots, x_n y_n\right) = \left(\frac{x_1 y_1}{\sum x_j y_j}, \ldots, \frac{x_n y_n}{\sum x_j y_j}\right) = \mathcal{C}\exp\left(\log(\vec{x}) + \log(\vec{y})\right) \\ \lambda \odot \vec{x} &= \mathcal{C}\left(x_1^{\lambda}, \ldots, x_n^{\lambda}\right) = \left(\frac{x_1^{\lambda}}{\sum x_j^{\lambda}}, \ldots, \frac{x_n^{\lambda}}{\sum x_j^{\lambda}}\right) = \mathcal{C}\exp\left(\lambda \log(\vec{x})\right) \\ \vec{x}, \vec{y} &\in S^n \ , \ \lambda \in \mathbb{R} \end{aligned} \tag{3}$$

En (3) les funcions exponencial (exp) i logaritme neperià (log) s'apliquen a un vector fent-ho sobre cada una de les components d'aquest vector. Observeu també que s'ha fet servir la notació moderna log per al logaritme neperià en lloc de la més clàssica ln; la raó està en el fet que, sota la perspectiva funcional actual, de logaritme només n'hi ha un, el neperià o natural, i tots els altres en són simples homotècies (Jarauta-Bragulat, 2000). Fins i tot diversos programes de càlcul simbòlic fan servir aquesta notació més actualitzada, que també és l'adoptada per diverses editorials de textos de l'àmbit de la matemàtica i l'estadística.

El símplex de tres parts S^3 admet una representació gràfica en el pla que té certa utilitat i que, en part, ha estat aplicada en alguns casos (per exemple, en edafologia, el diagrama de textures del sòl). Aquesta representació gràfica es basa en el teorema de Viviani, el qual estableix que en un triangle equilàter d'altura h (i per tant de longitud del costat $L = h/\sin(60°)$), en cada punt interior del triangle la suma de les distàncies (ortogonals) del punt a cada un dels tres costats del triangle és constant i l'altura té valor h. En la figura 1*a* hi ha una representació d'aquesta propietat. En la figura 1*b* hi ha la representació gràfica d'una recta en el símplex; en aquesta gràfica les coordenades (u, v) dels punts del pla per a representar una composició (x_1, x_2, x_3), es calculen com una combinació lineal convexa de les coordenades dels vèrtexs del triangle, tal com es mostra en l'equació (4):

$$A\left(\frac{1}{\sqrt{3}},1\right),\ B(0,0),\ C=\left(\frac{2}{\sqrt{3}},0\right)$$
$$(u,v)=Ax_1+Bx_2+Cx_3=\left(\frac{1}{\sqrt{3}}x_1+\frac{2}{\sqrt{3}}x_3,x_1\right) \quad (4)$$

Figura 1. a) *Representació gràfica del símplex de tres parts i significat geomètric de cada una de les parts.* b) *Representació gràfica d'una recta en el símplex de tres parts*



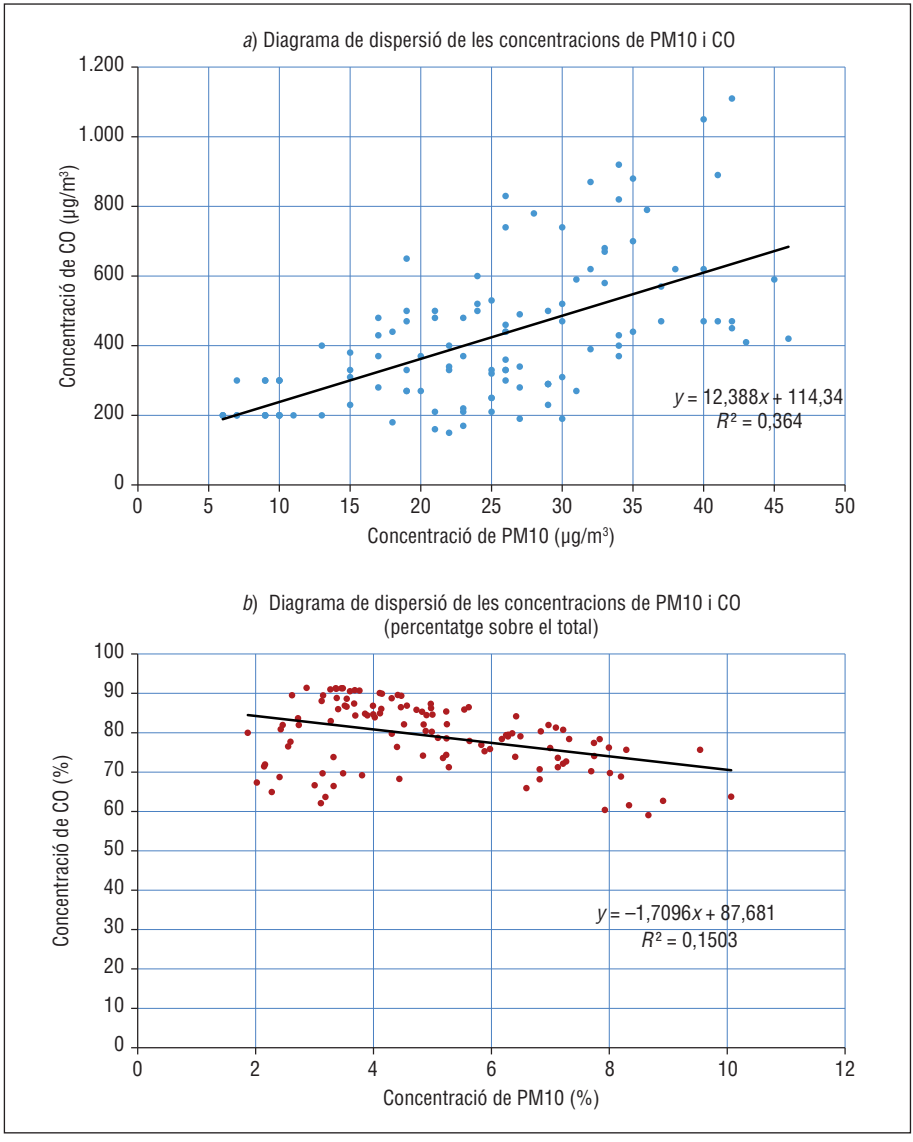
Font: Elaboració pròpia.

2.3. La correlació espúria

Una característica essencial de les composicions és que la informació que contenen no rau en el valor de les seves parts, sinó en el valor relatiu d'unes parts amb relació a les altres, és a dir, en els quocients entre les parts (Aitchison, 1986). Aquesta característica essencial afegida a la propietat de suma constant fa que els mètodes clàssics d'anàlisi estadística multivariant no es puguin aplicar a les dades composicionals de manera directa i sense tenir en compte les especificitats de les dades composicionals, tal com mostra l'efecte conegut com *correlació espúria* (o *falsa*).

Aquest efecte es produeix quan en una matriu de dades composicionals es modifica la forma d'expressió de les parts, cosa que pot donar lloc a canvis en la correlació d'algunes d'aquestes parts. Aquest fet va ser detectat i publicat per primera vegada per Karl Pearson, un dels pares de l'estadística moderna, l'any 1897. Per exemple, en la figura 2 es mostra el diagrama de dispersió i el coeficient de correlació entre dos contaminants de l'aire, en què

Figura 2. *Diagrames de dispersió dels pol·luents de l'aire PM10 i CO corresponents a una matriu de dades de contaminació de l'aire de Barcelona (2003-2013). En* a) *la concentració s'expressa en µg/m³ i el coeficient de correlació és 0,6033; en* b) *la concentració s'expressa en tant per cent sobre el pes total de la matriu i el coeficient de correlació és –0,3877*



PM10: partícules sòlides o líquides en suspensió de diàmetre aerodinàmic inferior a 10 µm.

Font: Elaboració pròpia a partir de Jarauta-Bragulat *et al.* (2016).

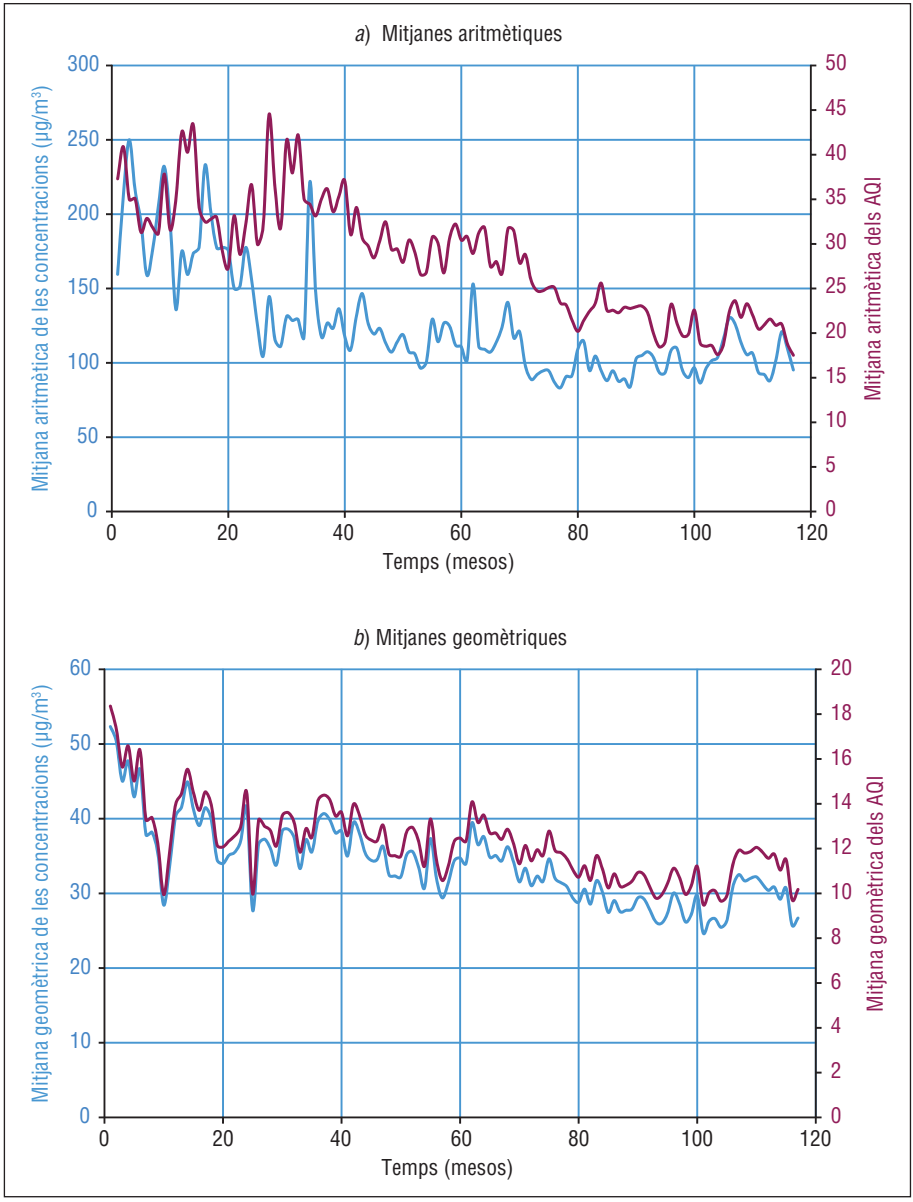
en *a* la concentració s'expressa en valors absoluts i en *b* s'expressa en valors relatius.

Per a completar el comentari sobre aquest efecte, volem destacar la seva presència en l'expressió de la qualitat de l'aire atmosfèric; per a més detalls, vegeu Jarauta-Bragulat *et al.* (2016). Un dels índexs de qualitat de l'aire, l'AQI (de l'anglès *air quality index*), és el definit per la United States Environmental Protection Agency (EPA), el qual, en resum, es defineix per a cada pol·luent amb una funció lineal a trossos (és a dir, una poligonal) de la concentració del pol·luent i després es calcula la mitjana aritmètica dels AQI de cada pol·luent considerat per a obtenir l'AQI global. Si aquesta metodologia fos correcta, hi hauria d'haver una correlació estreta entre la mitjana aritmètica de les concentracions de cada pol·luent i la mitjana aritmètica dels AQI, és a dir, l'AQI global. En la figura 3*a* hi ha la representació de les dues corbes en una sèrie de dades de Barcelona; observeu el que succeeix quan en lloc de la mitjana aritmètica es considera la mitjana geomètrica, tal com mostra la figura 3*b*.

2.4. La paradoxa de Simpson

La paradoxa de Simpson (Simpson, 1951) es produeix quan s'estudia un procés que depèn de dues variables dicotòmiques i la font d'informació pot donar els resultats en proporcions o en nombres absoluts. Per exemple: es vol fer un assaig sobre l'eficàcia d'un tractament per al bestiar i comparar el resultat pel sexe del bestiar; se seleccionen aleatòriament deu granges i aleshores les dues variables dicotòmiques en aquest cas són el sexe de l'animal (mascle, femella) i el resultat del tractament (èxit, fracàs). En la taula 1*a* es mostren els resultats de cada granja en percentatge dels mascles i femelles amb relació a l'èxit o al fracàs del tractament; al final, s'inclou una operació molt habitual: la mitjana aritmètica dels percentatges com a mesura global per a cada cas (mascles i femelles). En la taula 1*b* es mostren els nombres absoluts de mascles i de femelles que s'han enregistrat a cada granja amb relació a l'èxit o al fracàs del tractament i el percentatge global corresponent. Com s'observa, el resultat en els dos casos és contradictori (paradoxa de Simpson), ja que en el cas *a* es calcula la mitjana aritmètica dels percentatges i es conclou erròniament que el tractament funciona millor en les femelles. En canvi, en *b* calculant la proporció com el quocient dels valors en nombres absoluts amb relació als totals respectius, es conclou correctament que el tractament funciona millor en els mascles. Es pot trobar una explicació més detallada de les particularitats de les dades composicionals a Egozcue *et al.* (2011*a* i 2011*c*) i Mateu-Figueras *et al.* (2003).

Figura 3. a) *Sèrie de mitjanes aritmètiques de concentracions i mitjanes aritmètiques dels AQI.* b) *Sèrie de mitjanes geomètriques de concentracions i mitjanes geomètriques dels AQI. (Dades corresponents a Barcelona 2001-2010)*



Font: Elaboració pròpia a partir de Jarauta-Bragulat *et al.* (2016).

Taula I. a*) Proporcions en mascles i femelles de l'èxit o fracàs del tractament i mitjana aritmètica dels percentatges donats a cada granja.* b*) Nombres absoluts de mascles i de femelles amb relació a l'èxit o al fracàs del tractament; calculant la proporció com el quocient d'aquests valors, es conclou correctament que el tractament funciona millor en mascles que en femelles*

a)

Granja	Percentatge de mascles ÈXIT (%)	Percentatge de femelles ÈXIT (%)	Percentatge de mascles FRACÀS (%)	Percentatge de femelles FRACÀS (%)
A	85,48	90,91	14,52	9,09
B	66,67	73,53	33,33	26,47
C	86,67	95,00	13,33	5,00
D	66,67	73,91	33,33	26,09
E	84,62	91,30	15,38	8,70
F	84,38	86,36	15,63	13,64
G	87,72	90,48	12,28	9,52
H	65,22	75,38	34,78	24,62
I	68,42	76,19	31,58	23,81
J	86,44	90,48	13,56	9,52
Mitjana aritmètica	**78,23 %**	**84,35 %**	**21,77 %**	**15,65 %**

b)

Granja	Nombre de mascles ÈXIT	Nombre de femelles ÈXIT	Nombre de mascles FRACÀS	Nombre de femelles FRACÀS
A	53	20	9	2
B	12	50	6	18
C	52	19	8	1
D	14	51	7	18
E	55	21	10	2
F	54	19	10	3
G	50	19	7	2
H	15	49	8	16
I	13	48	6	15
J	51	19	8	2
Nombre total	**369**	**315**	**79**	**79**
Percentatge sobre el mateix sexe	**82,37 %**	**79,95 %**	**17,63 %**	**20,05 %**

Font: Elaboració pròpia.

3. ELS PRINCIPIS DE L'ANÀLISI DE DADES COMPOSICIONALS. TRANSFORMACIONS COMPOSICIONALS

3.1. Els principis de l'anàlisi de dades composicionals

Com en moltes altres branques de la ciència, cal enunciar uns principis o axiomes sota els quals es desenvolupa una teoria determinada. En el cas de l'anàlisi de dades composicionals, els principis sota els quals es desenvolupa la teoria corresponent són els que s'enuncien a continuació.

Primer principi: principi de naturalesa o invariància d'escala. En un problema de naturalesa composicional, el valor en termes absoluts de les parts és irrellevant i només és significatiu el valor relatiu d'aquestes parts, és a dir, els quocients entre elles. Com a conseqüència d'aquest principi, es dedueix que qualsevol funció aplicada sobre dades composicionals ha de poder expressar-se en termes de quocients entre les seves parts o components.

Segon principi: principi de coherència subcomposicional. La magnitud relativa entre les parts d'una subcomposició no pot canviar en relació amb la magnitud relativa entre les parts de la composició original.

Tercer principi: principi del treball en coordenades. Si es treballa en coordenades respecte d'una base ortonormal tots els càlculs es redueixen als que s'apliquen en l'espai euclidià ordinari (és a dir, si es pren com a referència una base ortonormal, que és la formada per vectors unitaris i mútuament ortogonals, els càlculs es poden fer com si es treballés amb l'espai multivariant dels nombres reals, cosa que simplifica notablement els càlculs). Com a conseqüència d'aquest principi, s'apliquen transformacions del símplex en l'espai euclidià ordinari en el qual es treballa habitualment amb comoditat; finalitzat el treball en aquest espai euclidià ordinari, es retorna al símplex per a conèixer i interpretar els resultats.

Per a més detalls sobre aquests principis, vegeu Mateu-Figueras *et al*. (2003 i 2011).

3.2. Transformacions de dades composicionals

L'aplicació del tercer principi permet fer assumible i operatiu el treball analític amb dades composicionals, que altrament seria poc menys que impossible. Això s'aconsegueix amb transformacions dels vectors del símplex (composicions) en vectors de l'espai euclidià ordinari, mitjançant funcions vectorials:

Transformació log-quocient additiva, ALR (de l'anglès, *additive logratio transformation*). Es defineix com una aplicació del símplex de n parts S^n en l'espai euclidià ordinari $\mathbb{R}^{n-1}$ mitjançant:

$$\vec{u} = \mathrm{ALR}(\vec{x}) = \left(\log \frac{x_1}{x_n}, \log \frac{x_2}{x_n}, \ldots, \log \frac{x_{n-1}}{x_n} \right) \tag{5}$$

Les components de la imatge d'una composició per la transformació ALR són independents. La inversa d'aquesta transformació es calcula mitjançant:

$$\vec{x} = \mathrm{ALR}^{-1}(\vec{u}) = \mathcal{C}\left(\exp(u_1), \exp(u_2), \ldots, \exp(u_{n-1}), 1\right) \tag{6}$$

Transformació log-quocient centrada, CLR (de l'anglès, *centered logratio transformation*). Es defineix com una aplicació del símplex de *n* parts S^n en l'espai euclidià ordinari $\mathbb{R}^n$ mitjançant:

$$\begin{aligned} \vec{v} &= \mathrm{CLR}(\vec{x}) = \left(\log \frac{x_1}{g(\vec{x})}, \log \frac{x_2}{g(\vec{x})}, ..., \log \frac{x_n}{g(\vec{x})} \right) \\ g(\vec{x}) &= \sqrt[n]{x_1 x_2 \cdots x_n} = \left(x_1 x_2 \cdots x_n\right)^{1/n} \end{aligned} \tag{7}$$

Les components de la imatge d'una composició per la transformació CLR sumen zero i, per tant, no són independents. La inversa d'aquesta transformació es calcula mitjançant:

$$\vec{x} = \mathrm{CLR}^{-1}(\vec{v}) = \mathcal{C}\left(\exp(v_1), \exp(v_2), ..., \exp(v_n)\right) \tag{8}$$

Transformació log-quocient isomètrica, ILR (de l'anglès, *isometric logratio transformation*). Es defineix com una aplicació del símplex de *n* parts S^n en l'espai euclidià ordinari $\mathbb{R}^{n-1}$ mitjançant:

$$\vec{w} = \mathrm{ILR}(\vec{x}) = \left(\sqrt{\frac{1}{2}} \log \frac{x_1}{x_2}, \sqrt{\frac{2}{3}} \log \frac{(x_1 x_2)^{1/2}}{x_3}, ..., \sqrt{\frac{n-1}{n}} \log \frac{(x_1 \cdots x_{n-1})^{1/(n-1)}}{x_n} \right) \tag{9}$$

Les components de la imatge d'una composició per la transformació ILR són independents. És molt útil disposar de la versió matricial d'aquesta transformació:

$$\vec{w} = \mathrm{ILR}(\vec{x}) = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & 0 & 0 & \cdots & 0 \\ \cdots & & & & & & \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & & \cdots & & \frac{1}{\sqrt{n(n-1)}} & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix} \begin{pmatrix} \log x_1 \\ \log x_2 \\ \vdots \\ \log x_{n-1} \end{pmatrix} \tag{10}$$

La inversa d'aquesta transformació es calcula mitjançant:

$$\begin{aligned}\vec{w} &= \mathrm{ILR}(\vec{x}) = V^{\top}\log(\vec{x})\\ \vec{x} &= \mathrm{ILR}^{-1}(\vec{w}) = \mathcal{C}\exp\left(V\vec{w}\right)\end{aligned} \tag{11}$$

Les dues primeres transformacions de dades composicionals van ser definides per Aitchison (1986) i la tercera per Egozcue *et al.* (2003).

4. PROCESSOS COMPOSICIONALS. MODELS DIFERENCIALS COMPOSICIONALS

4.1. Processos composicionals

Amb molta freqüència, l'estudi d'un problema pràctic comporta l'anàlisi d'un procés en el qual intervenen una funció o més que depenen d'una variable o més. Quan es consideren funcions d'una variable, aquesta acostuma a ser el temps, i el procés es pot definir amb una funció de valors vectorials positius definida en un interval de la recta real:

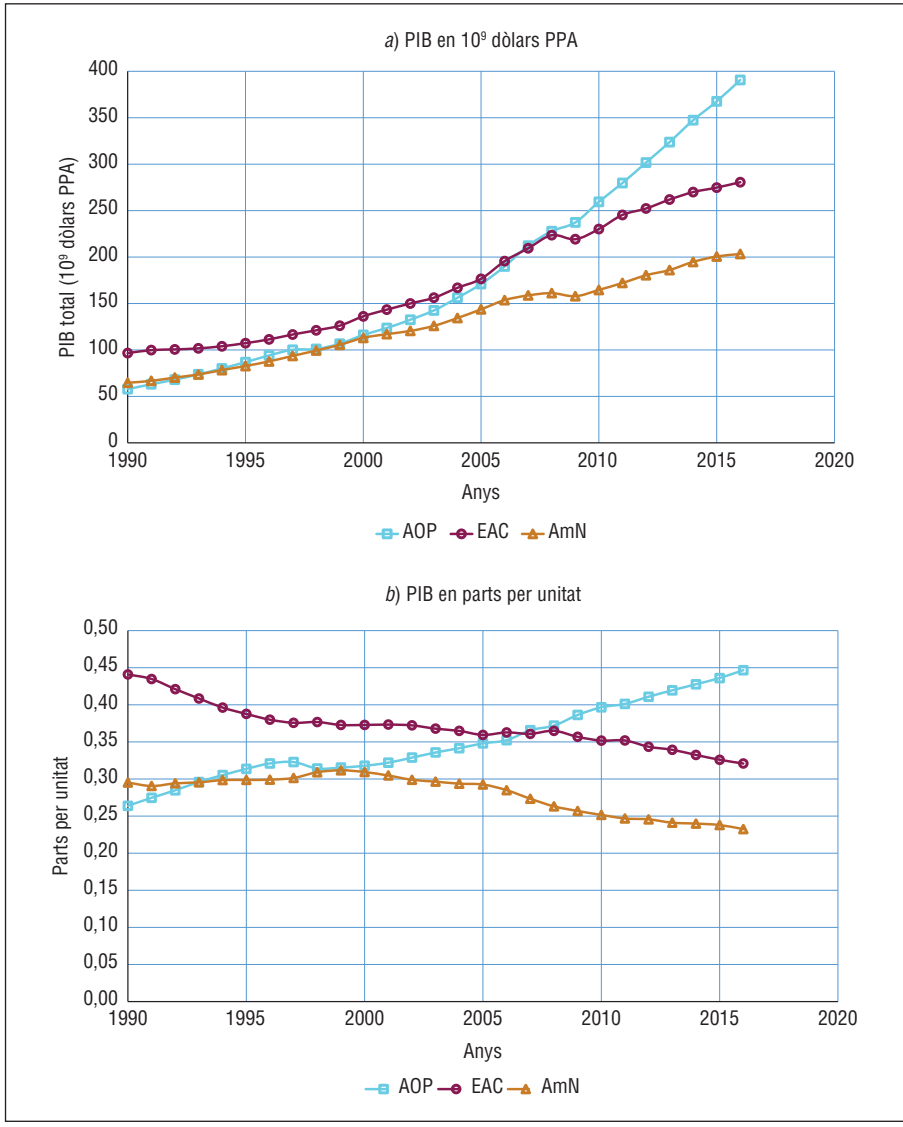
$$\vec{F}(t):[0,T]\subseteq\mathbb{R}\rightarrow\mathbb{R}_{+}^{n};\quad \vec{F}(t)=\left(F_1(t),F_2(t),...,F_n(t)\right) \tag{12}$$

Els valors corresponents a aquesta funció s'anomenen *valors en massa* perquè les unitats en les quals s'expressen són absolutes; per exemple: euros, nombre de persones, quilos, tones, etc. La suma de les components d'una funció vectorial com la que expressa (12) s'anomena *massa total* i es designa *M*(t), la qual pot ser constant o variable. Si es considera la clausura de la funció definida en (12), és a dir, una funció vectorial en la qual cada component és la corresponent de (12) dividida per la massa total; aleshores s'obté l'expressió analítica d'un *procés composicional*, tal com s'expressa en l'equació següent:

$$\begin{aligned}&\vec{f}(t):[0,T]\subseteq\mathbb{R}\rightarrow S^{n}\\ &\vec{f}(t)=\mathcal{C}\,\vec{F}(t)=\left(\frac{F_1(t)}{M(t)},\ ...,\frac{F_n(t)}{M(t)}\right)=\left(f_1(t),...,f_n(t)\right)\end{aligned} \tag{13}$$

Un element molt important en la caracterització del sistema alimentari d'un país és el seu producte interior brut (PIB) tant en termes absoluts (unitats monetàries) com relatiu a cada habitant (unitats monetàries *per capita*), que apareix en l'eix de l'economia definit per Clotet Ballús *et al.* (2013) i

Figura 4. a) *PIB total de tres àrees geogràfiques del món en milers de milions de dòlars PPA.* b) *Proporció de cadascuna de les tres àrees geogràfiques sobre el PIB conjunt. Anys 1990-2016*



Dòlars PPA: dòlars de paritat de poder adquisitiu, o unitats monetàries locals que es necessiten per a adquirir, dins del país en qüestió, la mateixa quantitat de béns que als Estats Units es comprarien amb un dòlar nord-americà.

Àrees geogràfiques: AOP = Àsia oriental i Pacífic, EAC = Europa i Àsia central, AmN = Amèrica del Nord.

Font: Elaboració pròpia a partir de dades de *Google public data.*

Jarauta-Bragulat *et al.* (2018). En la figura 4*a* hi ha la representació gràfica de la renda total (en milers de milions de dòlars) al llarg d'una sèrie d'anys, de tres àrees geogràfiques del món que agrupen diversos països cada una. En la figura 4*b* hi ha la representació gràfica de les proporcions corresponents (parts per unitat sobre el PIB conjunt de les tres àrees). Cal observar que la tendència de creixement o decreixement que posa de manifest la corba de *a* pot ser força diferent de la corresponent a *b*, ja que el creixement o decreixement relatiu no depèn només de l'àrea, sinó del comportament respecte del conjunt de les àrees.

4.2. La derivada composicional

Així com en funcions reals de variable real s'aplica el concepte de *derivada* per a caracteritzar la variació de la funció i s'estudien les equacions diferencials (ordinàries) per a establir-ne aquesta caracterització, per a composicions s'ha establert el concepte de *derivada composicional* (Egozcue *et al.*, 2011*b*), que es defineix com:

$$\begin{aligned} &\vec{f}:[0,T]\subset\mathbb{R}^{+}\rightarrow S^{n}\ ;\ \ \vec{f}(t)=\left(f_1(t),f_2(t),...,f_n(t)\right)\\ &D^{\oplus}\vec{f}(t)=\lim_{h\rightarrow 0}\left(\frac{1}{h}\odot\left(\vec{f}(t+h)\ominus\vec{f}(t)\right)\right)\end{aligned} \tag{14}$$

La derivada composicional es calcula mitjançant:

$$D^{\oplus}\vec{f}(t)=\mathcal{C}\exp\left(D\log\vec{f}(t)\right)=\mathcal{C}\exp\left(\frac{Df_1(t)}{f_1(t)},\frac{Df_2(t)}{f_2(t)},...,\frac{Df_n(t)}{f_n(t)}\right) \tag{15}$$

Per a més detalls sobre aquesta definició, els càlculs associats i la seva interpretació, podeu consultar Egozcue *et al.* (2011*b*).

4.3. Equacions diferencials ordinàries composicionals

La derivada composicional s'aplica de manera molt útil a la descripció i a la modelització de l'evolució en el temps de sistemes dinàmics composicionals, tal com s'estableix a Egozcue i Jarauta-Bragulat (2014). Tenen un interès especial els models lineals, que en coordenades ILR es defineixen mitjançant el sistema d'equacions diferencials ordinàries que s'expressa amb l'equació matricial:

$$\begin{pmatrix} Du_1(t) \\ \vdots \\ Du_{n-1}(t) \end{pmatrix} = \begin{pmatrix} a_1^1 & a_2^1 & \cdots & a_{n-1}^1 \\ a_1^2 & a_2^2 & \cdots & a_{n-1}^2 \\ \cdots & & & \\ a_1^{n-1} & a_2^{n-1} & \cdots & a_{n-1}^{n-1} \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_{n-1}(t) \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{pmatrix} \quad (16)$$

A l'equació (16) la notació *D* significa derivació respecte de la variable *t*. L'equació (16) es pot escriure de manera simplificada mitjançant:

$$D\vec{u}(t) = A\vec{u}(t) + \vec{b} \quad (17)$$

A partir de dades experimentals i mitjançant tècniques desenvolupades per Egozcue i Jarauta-Bragulat (2014), es calculen els coeficients de la matriu i dels termes independents, cosa que permet l'ajust d'un model lineal i la possibilitat de dur a terme inferència. Per exemple, en el cas del PIB de les tres àrees geogràfiques considerades en l'exemple anterior, el millor ajust per a l'estimació dels coeficients del model és per a un model amb matriu no nul·la i termes independents nuls. La figura 5 mostra el resultat de l'ajust d'aquest model diferencial lineal composicional a les dades de PIB de les tres àrees geogràfiques. Per a més informació sobre l'aplicació en l'estudi del sistema agroalimentari, es pot consultar Clotet *et al.* (2013), Colomer-Xena i Jarauta-Bragulat (2016) i Jarauta-Bragulat *et al.* (2018).

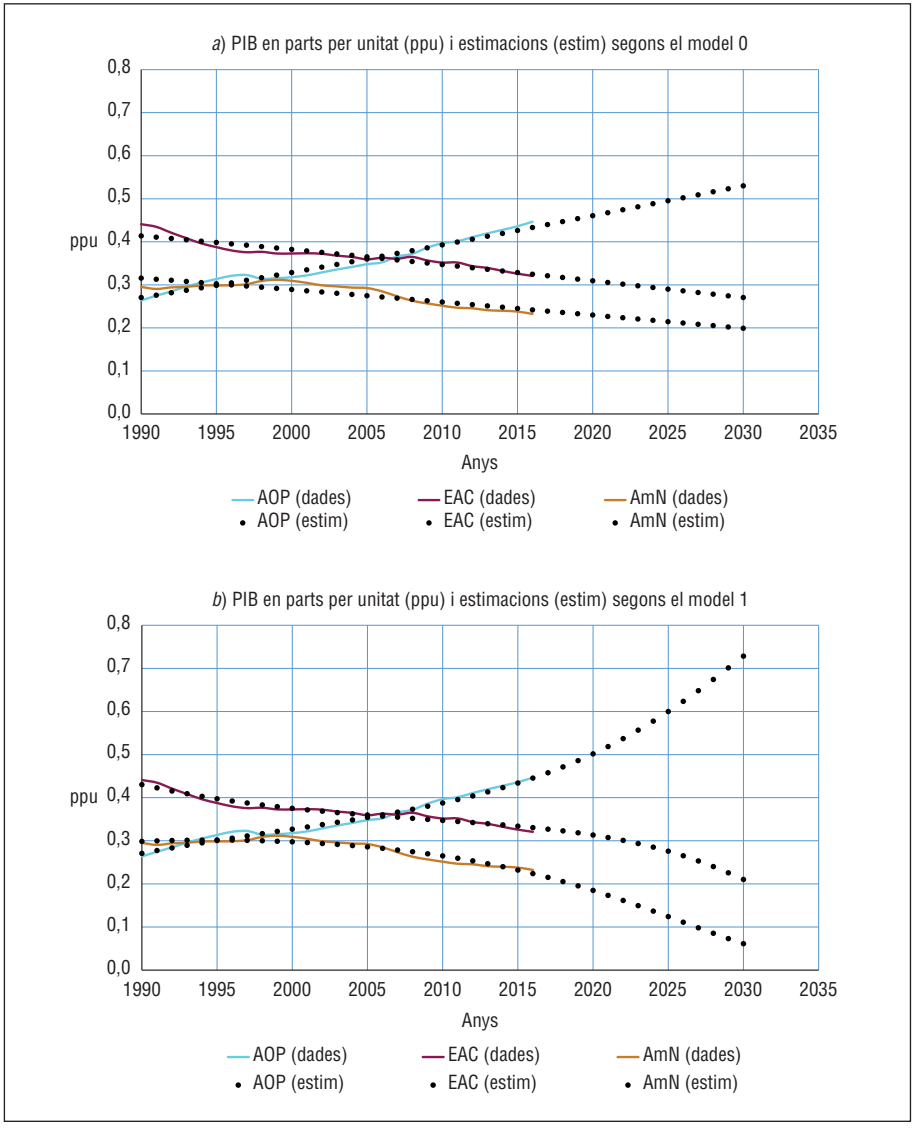
5. CONCLUSIONS

La recerca que involucra models matemàtics i informació quantitativa s'ha de dur a terme amb una selecció acurada dels conceptes i de les eines matemàtiques i estadístiques, i amb un tractament adequat de les dades numèriques. Les dades composicionals han estat ignorades inexplicablement en el passat i és clau tenir-les en compte en el futur. A continuació, es detallen les conclusions més rellevants del que s'ha exposat en aquest treball.

1. Les dades composicionals apareixen molt sovint i en gairebé totes les disciplines. La naturalesa de les dades composicionals i les seves propietats són diferents de les propietats de les dades numèriques estàndard (vectors de components reals). Per tant, cal disposar d'eines adequades per a la seva anàlisi estadística.

2. La restricció que suposa el fet que la suma de les parts d'una composició sigui constant té com a conseqüència que els mètodes estadístics clàssics siguin inadequats. Per tant, cal desenvolupar noves metodologies compatibles amb el caràcter composicional de les dades.

Figura 5. *PIB de tres àrees geogràfiques del món.* a*) Ajust d'un model diferencial lineal amb matriu nul·la (model tipus 0).* b*) Ajust d'un model diferencial lineal amb matriu no nul·la i termes independents nuls (model tipus 1)*



Nota: Les proporcions (AOP, EAC, AmN) són inicialment 26,4 %, 44,1 % i 29,5 % l'any 1990; l'any 2016 són 44,7 %, 32,1 % i 23,2 %; les previsions per a l'any 2030 són 53,0 %, 27,1 % i 19,9 %, segons el model tipus 0, i 72,8 %, 21,0 % i 6,1 %, segons el model tipus 1.

Àrees geogràfiques: AOP = Àsia oriental i Pacífic, EAC = Europa i Àsia central, AmN = Amèrica del Nord.

Font: Elaboració pròpia.

3. Les metodologies per a l'anàlisi i per al tractament de dades composicionals tenen com a fonament els treballs iniciats per John Aitchison (1986) i estan basades en la transformació de les dades composicionals mitjançant logaritmes de quocients (log-quocients); així s'obtenen vectors de l'espai euclidià real ordinari multidimensional.

4. Per a l'estudi de fenòmens aleatoris, en general, i composicionals, en particular, és essencial determinar l'espai suport de les observacions i optar per una mètrica adequada al problema abans d'iniciar l'estudi. Si el suport i la mètrica corresponen a una estructura d'espai euclidià, en general és més fàcil treballar en coordenades respecte a una base ortonormal.

5. La geometria d'Aitchison en el símplex i les coordenades log-quocient centrada (CLR) i log-quocient isomètrica (ILR) permeten aplicar sense problemes tècniques d'anàlisi de dades i inferència estadística a conjunts de dades composicionals. En cada cas s'han de trobar les expressions que facilitin millor la interpretació dels resultats.

6. Els processos en els quals intervenen composicions evolutives univariants es poden modelitzar de manera molt adequada mitjançant equacions diferencials ordinàries lineals composicionals. Aquests models permeten una descripció adequada del comportament evolutiu de les parts del sistema i formular una prospectiva basada en la interacció entre les diverses parts que el componen.

BIBLIOGRAFIA

Aitchison, J. (1986). *The statistical analysis of compositional data*: *Monographs on statistics and applied probability*. Londres: Chapman & Hall Ltd., p. 416. [Reimprès el 2003 amb material addicional per The Blackburn Press.]

Clotet Ballús, R.; Colomer-Xena, Y.; Jarauta-Bragulat, E.; Mayor Zaragoza, F. (2013). «El sistema alimentario global: I – Definición de un espacio». *Revista Española de Estudios Agrosociales y Pesqueros*, vol. 235 (2), p. 13-32. ISSN 1575-1198.

Colomer-Xena, Y.; Jarauta-Bragulat, E. (2016). «La modelización del sistema alimentario: un desafío». A: *El sistema alimentario*: *Globalización, sostenibilidad, seguridad y cultura alimentaria*. Navarra: Thomson Reuters Aranzadi. Cap. 3, p. 91-109. ISBN 978-84-9135-265-5.

Egozcue, J. J.; Barceló-Vidal, C.; Martín-Fernández, J. A.; Jarauta-Bragulat, E.; Díaz-Barrero, J. L.; Mateu-Figueras, G. (2011*a*). «Elements of simplicial linear algebra and geometry». A: Pawlowsky-Glahn, V.; Buccianti, A. (ed.). *Compositional data analysis*: *Theory and applications*. Chichester: Wiley.

Egozcue, J. J.; Jarauta-Bragulat, E. (2014). «Differential models for evolutionary compositions». *Mathematical Geosciences*, vol. 46 (4), p. 381-410. ISSN 1874-8961 en paper; 1874-8953 electrònic.

Egozcue, J. J.; Jarauta-Bragulat, E.; Díaz-Barrero, J. L. (2011*b*). «Calculus of simplex-valued functions». A: Pawlowsky-Glahn, V.; Buccianti, A. (ed.). *Compositional data analysis*: *Theory and applications*. Chichester: Wiley.

Egozcue, J. J.; Pawlowsky-Glahn, V. (2011*c*). «Basic concepts and procedures». A: Pawlowsky-Glahn, V.; Buccianti, A. (ed.). *Compositional data analysis*: *Theory and applications*. Chichester: Wiley.

Egozcue, J. J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C. (2003). «Isometric logratio transformations for compositional data analysis». *Mathematical Geology*, vol. 35 (3), p. 279-300.

Jarauta-Bragulat, E. (2000). *Análisis matemático de una variable*: *Fundamentos y aplicaciones*. Barcelona: UPC, 455 p.

Jarauta-Bragulat, E.; Colomer-Xena, Y.; Clotet-Ballús, R. (2018). «El sistema alimentario global: II - Aproximación cuantitativa al espacio agroalimentario de la Europa mediterránea». *Revista Española de Estudios Agrosociales y Pesqueros*, vol. 249, p. 15-38. ISSN 1575-1198.

Jarauta-Bragulat, E.; Hervada-Sala, C.; Egozcue, J. J. (2016). «Air quality index revisited from a compositional point of view». *Mathematical Geosciences*, vol. 48 (5), p. 581-593. ISSN 1874-8961 en paper; 1874-8953 electrònic.

Mateu-Figueras, G.; Martín-Fernández, J. A.; Pawlowsky-Glahn, V.; Barceló-Vidal, C. (2003). «El problema del análisis estadístico de datos composicionales». A: *27 Congreso Nacional de Estadística e Investigación Operativa*. Lleida: Universitat de Lleida.

Mateu-Figueras, G.; Pawlowsky-Glahn, V.; Egozcue, J. J. (2011). «The principle of working on coordinates». A: Pawlowsky-Glahn, V.; Buccianti, A. (ed.). *Compositional data analysis*: *Theory and applications*. Chichester: 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». A: *Proceedings of the Royal Society of London*. Londres: The Royal Society. Vol. LX, p. 489-502.

Simpson, E. H. (1951). «The interpretation of interaction in contingency tables». *Journal of the Royal Statistical Society*, sèrie B, vol. 13, p. 238-241.