Análise estatística do mercado de aloxamento turístico da provincia de Ourense dende unha nova perspectiva: a análise de datos composicionais

A. Pérez-González, A. Padín Romero, T. R. Cotos-Yáñez e M. A. Mosquera
Departamento de Estatística e Investigación Operativa.
Campus de Ourense. Universidade de Vigo
anapg@uvigo.gal; adrianpadin1c@gmail.com; cotos@uvigo.gal; mamrguez@uvigo.gal

Resumo

O obxectivo principal deste traballo é analizar os datos recollidos na enquisa sobre a utilización das vivendas do uso turístico na provincia de Ourense dende o punto de vista composicional. Despois dun proceso de familiarización cos datos de natureza composicional, realizouse unha análise descritiva das variables que indicaban a repartición do gasto dos e das turistas que tiñan esta estrutura composicional. Na segunda etapa aplicouse un modelo de regresión loxística coa covarianza composicional. Finalmente, recolléronse os datos procedentes de fontes secundarias do Instituto Galego de Estatística para realizar unha análise da evolución temporal dunha serie composicional que representaba o número de persoas viaxeiras mensuais na provincia de Ourense. Os resultados obtidos indican que é recomendable ter en conta este tipo de estrutura de datos para acadar información sobre a repartición ou a interacción entre as posibles partes.

Palabras clave:

Biplot, composicional, imputación, regresión loxística e serie de tempo

1. Introdución

O obxectivo do subproxecto «Análise estatística do mercado de aloxamento turístico da provincia de Ourense dende unha nova perspectiva: a análise de datos composicionais» foi analizar os datos extraídos previamente da enquisa sobre as pautas de elección e o consumo de establecementos de aloxamento relacionados co uso das vivendas turísticas. O enfoque desta análise tratábase dende o punto de vista composicional, e fixamos a nosa atención nas variables que poidan seguir unha estrutura deste tipo.

Neste documento trataremos de mostrar as conclusións máis significativas que acadamos na consecución deste proxecto. Comezaremos describindo este tipo de datos e citando algúns ámbitos onde o seu uso é moito máis frecuente, para motivar así a necesidade de traballar con este tipo de datos no ámbito dos estudos turísticos.

1.1. Introdución aos datos composicionais

Unha composición describe as partes dun todo en forma cuantitativa e considérase que a información composicional que contén reside nas razóns entre calquera das partes consideradas [1]. Esta forma de estudar/estimar os datos e a relación entre eles pode aplicarse a diversos supostos como poden ser a repartición de orzamentos, da intención de voto, dos compostos químicos, ou coma no caso que nos ocupa, a repartición do gasto na análise do mercado de aloxamento turístico.

Os exemplos enumerados levan a que se consideren os datos composicionais como un conxunto de números positivos con suma constante para todas as observacións. De xeito formal, defínese o k-símplex de D-partes:

$$\mathbb{S}^D = \{(x_1, x_2, \ldots, x_D) \in \mathbb{R}^D \mid x_i > 0,\ i = 1,2,\ldots,D,\ \sum_{i=1}^{D} x_i = k\},$$

onde k é unha constante que toma un valor determinado en función das características da composición.

Son moitos os ámbitos onde se atopan este tipo de datos. No contexto electoral, Nguyen *et al.* [2] analizan a repartición de votos por partidos nunha subdivisión territorial concreta. Para isto, discuten e ilustran diversos modelos de regresión de datos composicionais aplicados a modelos de economía política.

Outro exemplo da aplicación dos datos composicionais é o mostrado en Gibergans-Baguena *et al.* [3], onde analizan series de datos sobre a calidade do aire a fin de avaliar o impacto de diversas medidas sociais e económicas.

Muller *et al.* [4] achegan unha revisión e unha interpretación da regresión composicional aplicada á análise de orzamentos. Para isto, presentan unha serie de ferramentas útiles para analizar modelos de regresión a través dos *log-ratios* e a súa aplicación a fin de revelar relacións potenciais entre distintos indicadores psicométricos dentro dun conxunto de datos.

Finalmente, e achegándose ao obxecto deste artigo, Coenders e Ferrer-Rosell [5] presentan unha introdución ao uso de datos composicionais ligados ao sector turístico. Para isto, ofrecen unha visión global das principais carencias do tratamento de datos composicionais baseándose na metodoloxía estatística clásica e tamén unha revisión das aplicacións das composicións en diversos campos como a socioloxía, a economía ou a educación.

1.2. Descrición da base de datos

No marco do proxecto deseñouse unha enquisa para os e as turistas co obxectivo de analizar as súas pautas de elección e de consumo de establecementos de aloxamento. En particular, a enquisa permitiu indagar sobre os factores que inflúen sobre as decisións de aloxamento dos e das turistas, as súas pautas de gasto en destino, o uso que fan das Vivendas de Uso Turístico (VUT) e a percepción que teñen sobre esta alternativa de aloxamento. A versión definitiva da enquisa recolleu algunhas suxestións realizadas por persoas expertas en investigación de mercados que realizaron un *pretest* dela.

A enquisa dividiuse en tres bloques:

1. Eleccións de aloxamento (opción habitual e compañía), importancia do aloxamento na elección do destino e factores determinantes da elección do tipo de aloxamento.
2. Gasto en destino e distribución deste (aloxamento fronte a outras opcións).
3. Bloque relativo ás VUT: frecuencia de uso, medios de busca e reserva, vantaxes fronte ao aloxamento hoteleiro e grao de satisfacción.

Ademais do seu perfil (xénero, idade, ocupación, renda, localidade e país de residencia) e dos seus costumes de viaxe, frecuencia antes e durante a pandemia e o mo-

tivo principal (ocio ou negocio), tamén se lles preguntou aos e ás turistas se visitaron recentemente a provincia de Ourense, se fixeron noite e se foi nunha VUT.

Considerouse que a poboación estaba composta por persoas maiores de idade que fixeron algunha viaxe nos últimos tres anos. Para facilitar a obtención dun número aceptable de respostas, optouse por realizar unha mostraxe en liña non probabilística polo mecanismo de bóla de neve. O procedemento para recoller os datos consistiu en enviarlles unha ligazón á enquisa a distintos grupos relacionados co turismo en redes sociais e a contactos mediante un correo electrónico, no cal lles pedían que cubrisen a enquisa e que llela difundisen a outros grupos e contactos. O período de recollida de datos estivo aberto entre o 15 de xuño e o 31 de agosto de 2021. Obtivéronse 159 respostas válidas.

Evidentemente, a mostraxe realizada non é aleatoria e os resultados obtidos da análise da enquisa non se poden extrapolar á poboación. As limitacións por causa da covid-19, o tempo de realización e outras peculiaridades da poboación suxeita a estudo fixo imposible realizar unha mostraxe aleatoria coa que si poderiamos extrapolar a información acadada. De todos os modos, este estudo amosa unha perspectiva distinta das determinadas estruturas de variables que si son significativas á hora de explicar diversos comportamentos.

Despois dun proceso de depuración dos datos, relativa á estrutura composicional, da que xa falaremos máis adiante, a base de datos coa que se traballou neste proxecto tiña 145 observacións e 77 variables. A análise descritiva da base de datos completa foi realizada polos membros do subproxecto «Alternativas de aloxamento turístico na provincia de Ourense: análise da evolución recente da oferta e dos novos padróns de conduta da demanda nun contexto de irrupción do fenómeno das vivendas turísticas e de uso turístico». Neste subproxecto só nos centramos na análise dende o punto de vista composicional.

Para realizar este proxecto utilizouse o software libre R Core Team [6].

2. Análise estatística dende o punto de vista composicional

2.1. Transformacións necesarias para traballar con datos composicionais

Unha das principais restricións presentes á hora de analizar os datos composicionais é que ao ser estes de suma constante non é posible aplicar técnicas estatísticas es-

tándar aos procesos de inferencia. Isto débese a un problema de correlacións espurias derivadas da reescalada dos datos para transformalos en composición.

A clave para analizar os datos composicionais é aplicar transformacións e estas poden encadrarse en tres tipos:

A transformación *log-ratio* aditiva ($alr(X)$)

$$alr(X_1, X_2, \ldots, X_D) = \{log\left(\frac{X_1}{X_D}\right), \ldots, log\left(\frac{X_{D-1}}{X_D}\right)\}$$

Esta transformación é asimétrica, polo que unha versión simétrica é obtida a través da seguinte transformación:

A transformación *log-ratio* centrada ($clr(X)$)

$$clr(X_1, X_2, \ldots, X_D) = \{log\left(\frac{X_1}{g(X)}\right), \ldots, log\left(\frac{X_D}{g(X)}\right)\},$$

onde $g(X) = g(X_1, \ldots, X_D) = \sqrt[D]{X_1, \ldots X_D}$ representa a media xeométrica.

O principal inconveniente presente nesta transformación é que a matriz de covarianzas das variables transformadas é singular, o que supón un serio problema á hora de aplicar un amplo espectro de técnicas estatísticas multivariantes. É por isto que se presenta a seguinte transformación:

A transformación *log-ratio* isométrica ($ilr(X)$)

Un dos primeiros traballos en estudar este tipo de transformacións é o de Egozcue [7]. Pero a súa interpretabilidade non resulta moi sinxela nalgúns casos. Por iso, neste traballo pensamos noutro tipo de transformación, proposta por Muller *et al.* [4], que facilita nalgún caso a súa interpretabilidade. A transformación consiste en transformar a variable composicional orixinal X nunha variable Z cuxa matriz de covarianza non sexa singular, do seguinte modo:

$$Z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}}\ log\left[\frac{X_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} X_j^{(l)}}}\right],\ i = 1, \ldots, D-1.$$

Pode verse que a primeira compoñente transformada (Z_1) contén información relativa da primeira compoñente (X_1) con respecto á media xeométrica do resto das compoñentes.

A idea consiste en ir permutando a primeira variable composicional (X_1) coas seguintes (X_1 con $l=2,...D$) e así realizar D transformacións, fixándonos sempre na primeira compoñente transformada (Z_1), que darían lugar a D posibles modelos estimados.

A análise para a primeira parte composicional X_1 obtense observando a primeira compoñente transformada Z_1 con $l = 1$, e a análise para a *–ésima* compoñente obtense realizando unha permutación entre X_1 e X_l e realizando a transformación anterior. Para ver máis detalles desta transformación, podemos citar o traballo de Muller *et al.* [4]. A aplicación destas transformacións permite aplicar técnicas estatísticas estándar, aínda que a interpretación dos coeficientes non é de forma directa.

2.2. Imputación dos ceros composicionais

A análise estatística das composicións co emprego de *log-ratios* das compoñentes presenta problemas cando se constata a presenza de valores anómalos no conxunto de datos como poden ser os ceros, e é precisamente nestes casos onde cómpre distinguir entre distintos tipos. Os ceros máis frecuentes son os que se atopan por debaixo do límite de detección, os cales non implican que estes valores presenten esa proporción respecto do total, senón que o seu valor está por debaixo do detectable no procedemento de mostraxe. Outro caso posible é que estes ceros sexan estruturais, é dicir, que a súa proporción é estritamente nula.

Para o caso que nos ocupa, considerarase que os ceros presentes no conxunto de datos están por debaixo do límite de detección e, polo tanto, empregarase o método multiplicativo. Dado $x \in \mathbb{S}^{\mathbb{D}}$, Martín-Fernándezet *et al.* [8] propoñen substituír a composición orixinal por unha versión modificada desta tal que:

$$X_j = \begin{cases} \delta_j & \text{se } X_j = 0 \\ \left(1 - \dfrac{\sum_{i,X_i=0} \delta_i}{c}\right) X_j & \text{se } X_j > 0 \end{cases}$$

onde δ_j é o valor imputado de x_j (límite de detección), e c é a constante que resulta da suma das compoñentes: $c = 1$ para o caso que nos ocupa.

As variables que imos considerar neste traballo cunha estrutura composicional van ser a repartición do gasto en transporte, aloxamento, manutención e outros gastos.

Despois de realizar unha depuración das observacións anómalas, a distribución de ceros é a que se pode ver na seguinte gráfica (figura 1). As barras vermellas verticais

representan a porcentaxe de ceros de cada unha das partes composicionais. Pola contra, as barras vermellas horizontais representan a porcentaxe de datos que seguen ese esquema de distribución de ceros, por exemplo, un 6,83 % non ten ningún cero, pero un 8,97 % ten ceros só na parte correspondente a gastos varios. Parece obvio que é preciso aplicar un mecanismo de imputación dos ceros. Neste conxunto de datos aplicamos o método de imputación que citabamos anteriormente.

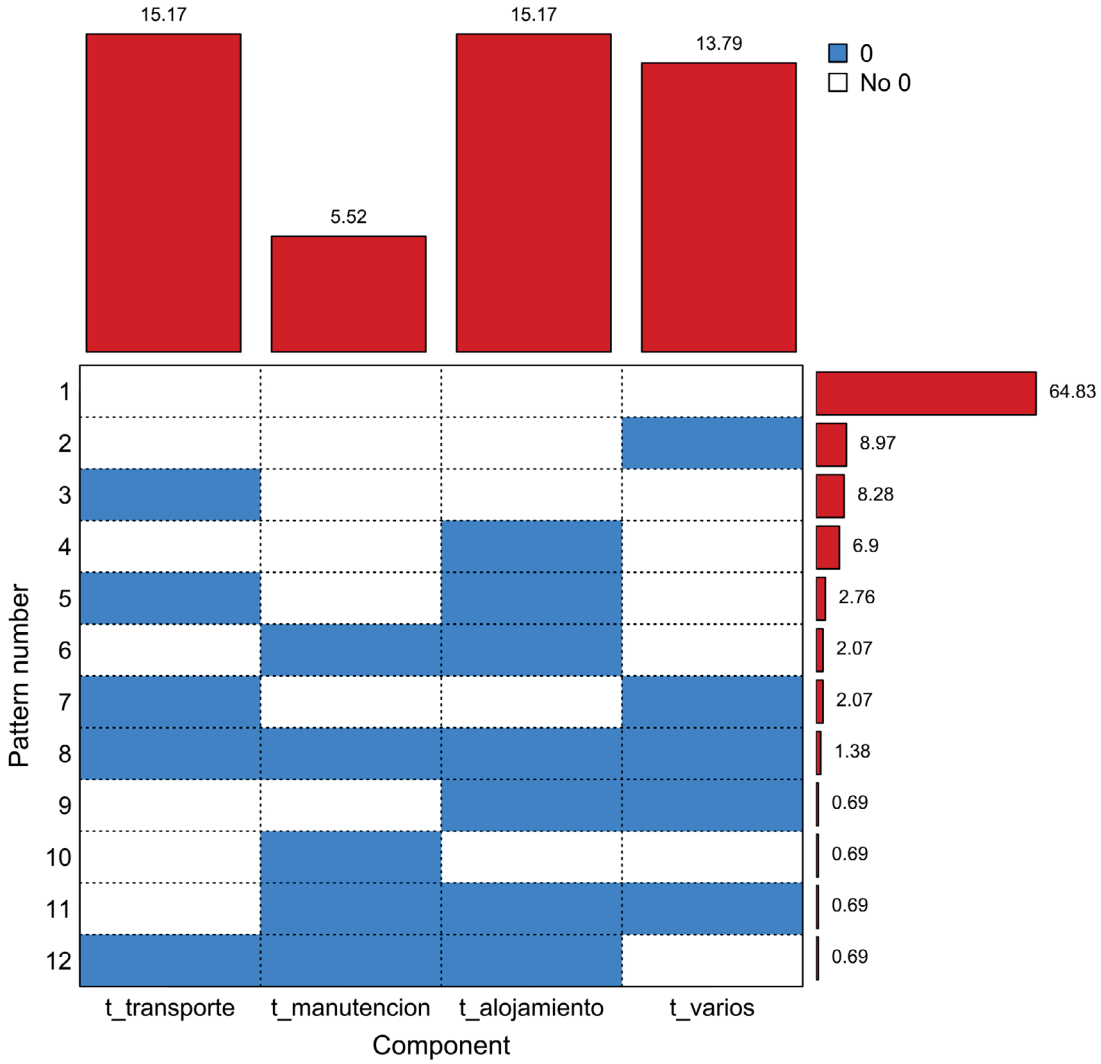

Figura 1. Modelo de datos missing

Unha vez realizada a imputación, realizamos a análise composicional.

2.3. Análise descritiva composicional

Nesta epígrafe realizaremos unha análise descritiva da variable composicional.

Comezaremos co valor medio composicional (táboa 1) que, como vemos, o peso do gasto recae na compoñente da manutención seguida do aloxamento.

Gasto en	Transporte	Manutención	Aloxamento	Varios
Media composicional %	8,9 %	54,6 %	25,3 %	11,2 %

Táboa 1. Vector de medias

Tamén podemos realizar o gráfico da repartición de gasto por sexo (figura 2) ou por ocupación (figura 3). Obsérvense as diferenzas na repartición do gasto que hai entre homes e mulleres, especialmente a parte que corresponde ao aloxamento, á manutención e aos gastos varios. Con respecto á ocupación das persoas entrevistadas, destaca a repartición do gasto do sector de xente desocupada, que prioriza o gasto en manutención seguido de transporte, e redúcese a porcentaxe dos gastos varios e do aloxamento.

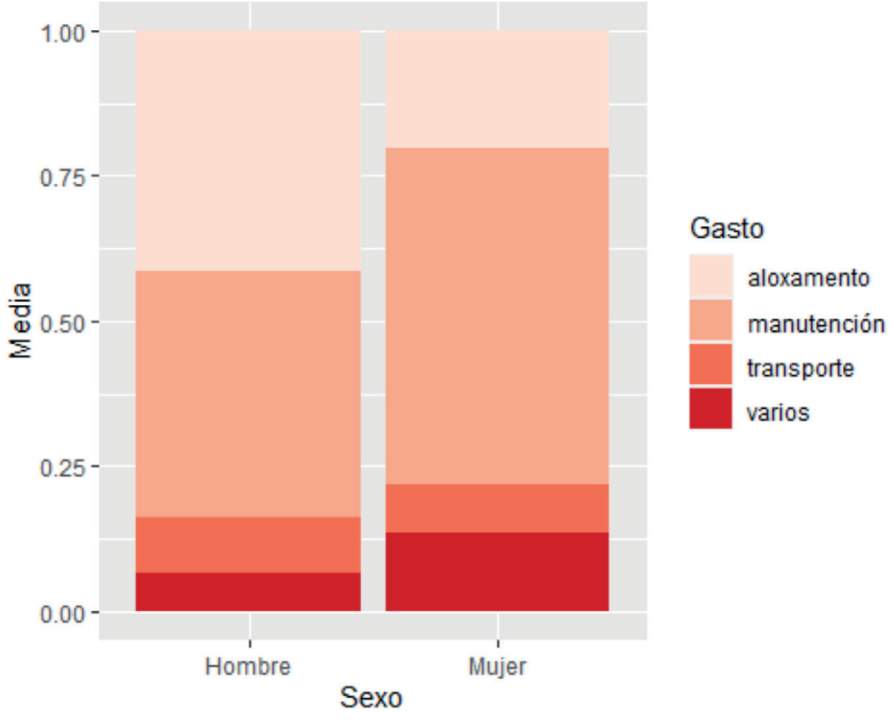

Figura 2. Repartición do gasto por sexo

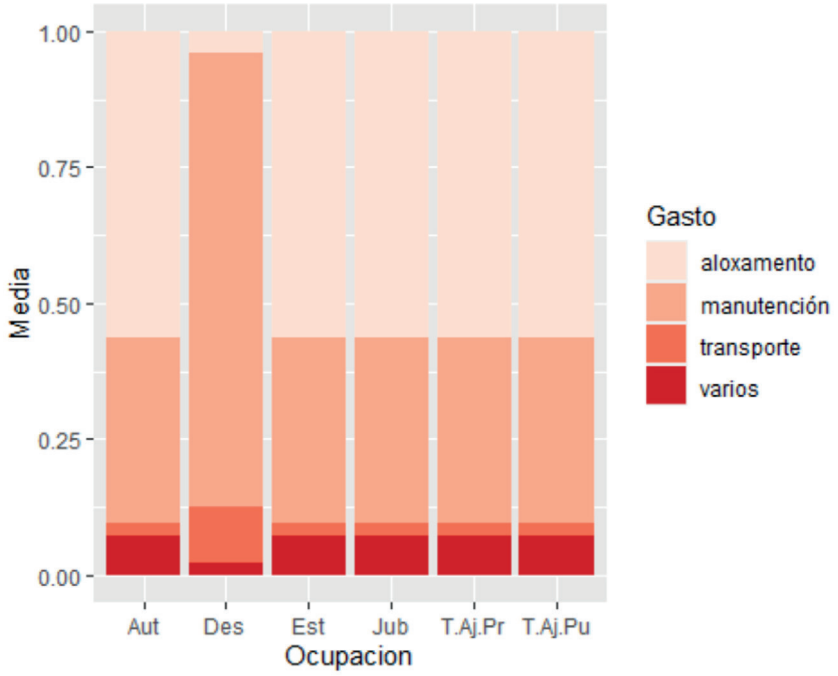

Figura 3. Repartición do gasto por ocupación

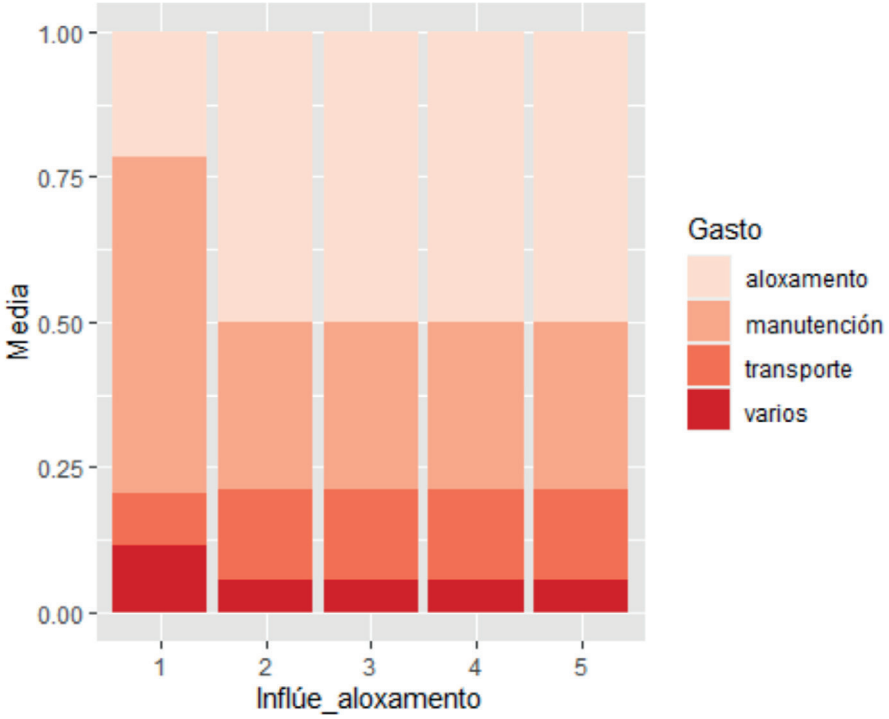

Figura 4. Repartición do gasto por grao de influencia no aloxamento

Outra análise interesante consistiría en ver se inflúe ou non o aloxamento na elección do destino turístico. Parece ser, segundo o gráfico (figura 4), que as e os visitantes aos que menos lles inflúe o destino turístico (valor 1) teñen unha repartición de gasto significativamente distinto ao resto, cunha porcentaxe dedicada á manutención moi superior ao resto das categorías.

Se queremos analizar a dispersión da variable composicional é preciso calcular a matriz de variación (táboa 2).

	Transporte	**Manutención**	**Aloxamento**	**Varios**
Transporte		12,92	21,86	24,98
Manutención	12,92		12,61	11,45
Aloxamento	21,86	12,61		25,83
Varios	24,98	11,45	25,83	

Táboa 2. Matriz de variación

Esta é unha matriz simétrica. Cada elemento representa a varianza do logaritmo do par de compoñentes. Os valores próximos a cero indican unha gran proporcionalidade entre esas dúas compoñentes. Como vemos neste caso, non se verifica esta condición en absoluto. Destaca a gran variabilidade do *log-ratio* entre os gastos varios e o aloxamento ou o transporte.

Un dos gráficos máis significativos na análise composicional é o gráfico ternario. Sobre un triángulo equilátero represéntanse as observacións. Cada vértice representa unha parte da composición; no noso caso, dado que temos catro compoñentes composicionais, o gráfico resultante sería unha matriz de todas as combinacións de dúas compoñentes mais a media xeométrica das outras dúas. Outra opción é representar a subcomposición de tres elementos con maior variabilidade. Para iso, consideramos como medida global da dispersión a varianza métrica (véxase Pawlowsky *et al.* [9]), tamén coñecida como varianza xeneralizada. No noso caso, as tres compoñentes con maior variabilidade son o gasto en transporte, aloxamento e gastos varios. O gráfico ternario podemos velo a continuación.

Para a interpretación é necesario recordar que a suma das tres compoñentes de cada observación é 1. Canto máis se aproxime un punto a un vértice, maior peso terá esa compoñente na partición da citada observación. Por outra parte, se un punto se

atopa nunha das liñas entre dous vértices, indicaría que a súa parte con respecto á terceira é nula. Por exemplo, os puntos situados na liña entre transporte e aloxamento teñen unha coordenada nula con respecto aos gastos varios. Como só consideramos a subcomposición de transporte, aloxamento e gastos varios, a media composicional (táboa 3) cambia e o aloxamento é o compoñente con maior peso. Lembremos que na partición orixinal era a manutención a que presentaba un maior peso na repartición.

Gasto en	Transporte	Aloxamento	Varios
Media	0,197	0,557	0,246

Táboa 3. Vector de media da subcomposición

Como vemos no gráfico (figura 5), a maior parte dos e das visitantes teñen un gasto en aloxamento importante, con puntos próximos a ese vértice. O valor medio composicional tamén aparece reflectido no punto negro do gráfico.

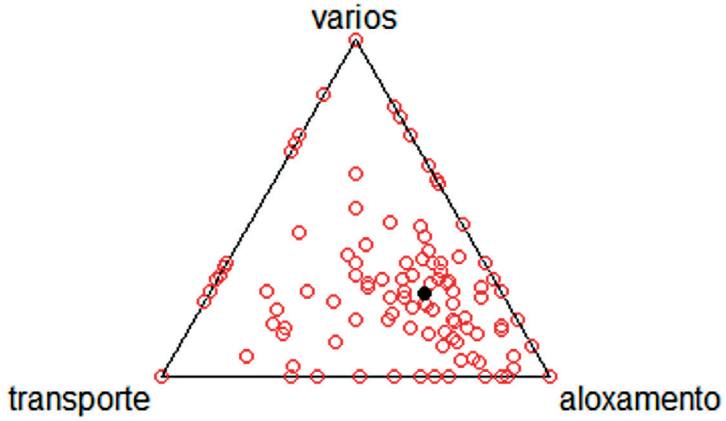

Figura 5. Gráfico ternario

Outro punto importante era ver a posible relación que existe entre as partes da composición. Con este obxectivo, realizouse unha análise de compoñentes principais e debuxamos o *biplot* asociado. O *biplot* é un gráfico que se utiliza para representar tanto as observacións coma as variables involucradas na análise de compoñentes principais. Hai que recordar que dado que a nosa variable composicional orixinalmente ten as partes linealmente dependentes, partimos de catro partes, pero o resultado son tres compoñentes principais.

	Importancia das compoñentes		
	Comp. 1	Comp. 2	Comp. 3
Standard deviation	3,659	3,304	1,763
Proportion of Variance	0,488	0,398	0,113
Cumulative Proportion	0,488	0,887	1

Táboa 4. Resumo da variabilidade da análise das compoñentes principais

Na táboa 4, podemos ver como as dúas primeiras compoñentes principais xa explican case un 90 % da variabilidade. Isto indica que a representación gráfica do *biplot* (por defecto representa as dúas primeiras compoñentes principais) vai explicar razoablemente o que ocorre tanto coas variables coma coas observacións.

Outra análise importante son as cargas derivadas da análise composicional das compoñentes principais.

	Cargas das compoñentes principais		
	Comp. 1	Comp. 2	Comp. 3
Transporte	-0,364	0,750	0,233
Manutención	0,073	-0,029	-0,862
Aloxamento	-0,495	-0,657	0,270
Varios	0,786	-0,064	0,359

Táboa 5. Resumo das cargas da análise das compoñentes principais

A suma das cargas de cada compoñente (suma por columnas) é nula. A súa interpretación non é a mesma que se tivésemos variables non composicionais. Aquí analízase a diferenza entre as cargas factoriais das variables (véxase a táboa 5). Por exemplo, na primeira compoñente principal o cambio relativo á ratio de transporte/aloxamento é moi pequeno porque as cargas factoriais son moi parecidas. No entanto, para a segunda compoñente principal a diferenza das cargas para estas dúas variables é moi significativa, o que indica que este *log-ratio* se incrementa ao longo desta compoñente principal cunha pendente de 1,4 resultante do seguinte cálculo: 0,657 + 0,750.

Unha representación gráfica destas pendentes pódese ver no seguinte gráfico (figura 6), que representa o factor multiplicativo que cada compoñente principal exerce sobre cada ratio de pares da composición. As barras que superen o 1, véxase por exemplo na segunda compoñente principal, indican que esa compoñente afecta positivamente ao *log-ratio* correspondente.

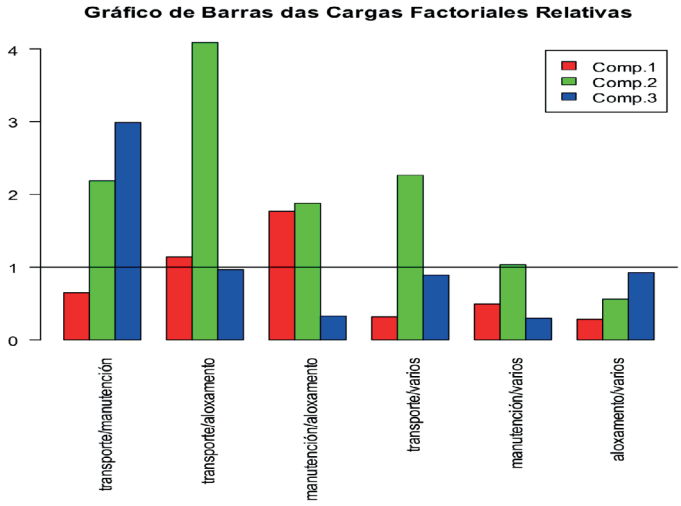

Figura 6. Cargas factoriais relativas

O seguinte gráfico (figura 7) representa as cargas de cada compoñente composicional como se fosen vectores composicionais. Destacan a parte de gastos varios e de transporte no peso da primeira e da segunda compoñente respectivamente.

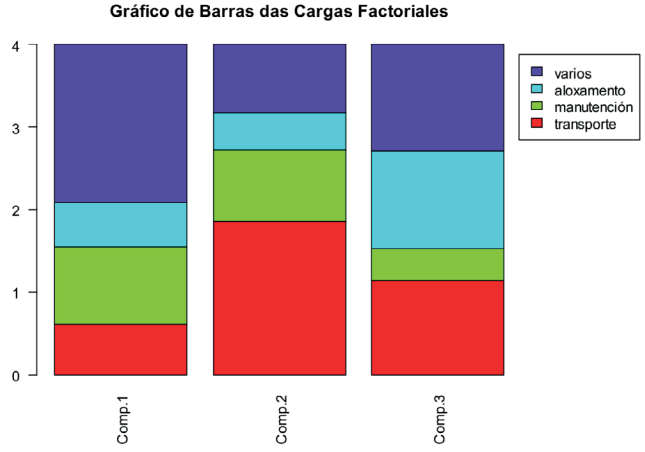

Figura 7. Cargas factoriais representadas como vectores composicionais

Outro gráfico interesante é o *biplot*. É un gráfico, normalmente bidimensional, que representa as proxeccións das variables e das observacións sobre un par de compoñentes principais. Na análise composicional pódense distinguir dous tipos de *biplots*: *covariance biplot* e *form biplot*. O tipo *covariance biplot* utilízase principalmente para observar as variables, mentres que o *form biplot* representa mellor o comportamento dos individuos. Neste traballo imos representar o *covariance biplot* para analizar a relación entre as variables.

A interpretación dos *biplots* para variables composicionais non é a mesma ca para variables non composicionais. No caso do *covariance biplot*, débense observar os enlaces (ligazóns) que se forman ao unir as puntas das frechas. Cómpre sinalar que as frechas en si non representan a variable composicional, senón a súa transformación *clr*, é dicir, representan a varianza do logaritmo desa parte dividida pola media xeométrica do resto das partes composicionais. Podemos ver que no noso *biplot* (figura 8) os enlaces entre transporte, aloxamento e gastos varios son todos bastante longos, o que implica bastante variabilidade nos *log-ratios* deses pares de variables. Tamén se observa que na primeira compoñente a parte composicional que ten máis peso é a relativa aos gastos varios, algo que xa viamos no gráfico anterior. Porén, na segunda compoñente son os gastos en transporte e en aloxamento os que presentan maior relevancia. Tamén parece que as observacións se agrupan en catro clústeres atendendo ás dúas primeiras compoñentes principais.

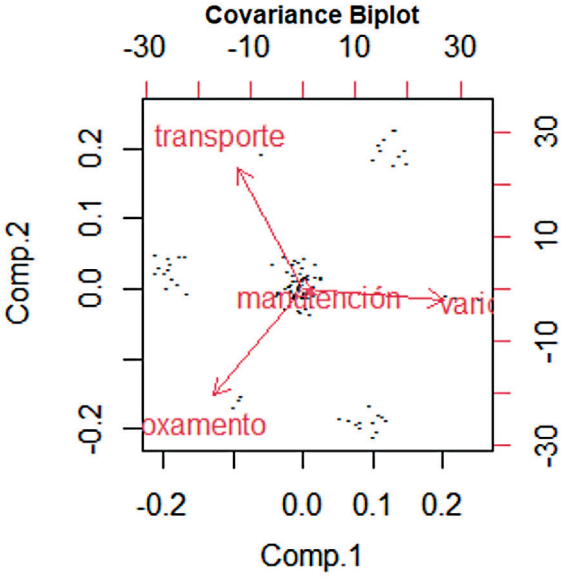

Figura 8. *Biplot*

2.4. Análise de regresión loxística

Outro dos obxectivos do estudo era ver se a repartición do gasto do e da turista influía ou non na elección de vivendas de uso turístico. Por esta razón, en primeiro lugar recodificamos a variable recollida no estudo como 1 ou 0 en función de se usaran ou non algunha vez este tipo de aloxamento. O modelo que utilizamos para analizar este comportamento foi un modelo de regresión loxística coa covarianza composicional. Á parte da variable composicional tamén usamos outras variables recollidas no estudo que podían ter relación. O modelo final foi o seguinte:

$$P(I_{UsoVivenda} = 1/\mathbf{X}) = logit\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right)$$

onde $logit(u) = e^u/(1+e^u)$, e $\mathbf{X} = (X_1, X_2, \ldots, X_p)^t$ é o vector de variables preditoras.

No noso caso, este conxunto está formado polas catro variables composicionais mais a variable de idade agrupada en tres categorías, a variable da pregunta 6 do cuestionario que indica se inflúe o aloxamento na elección do destino nas viaxes de ocio e, finalmente, a variable que informa se o usuario ou usuaria viaxa só ou con acompañante. Temos que dicir que foron varios os modelos que se probaron previamente e o máis significativo foi o que mostramos a continuación.

A dificultade deste modelo radica na inclusión do vector composicional como unha covarianza. Este vector trátase de forma conxunta, é dicir, non podemos eliminar unha parte composicional do modelo e deixar outras.

Dado que as partes da variable composicional son linealmente dependentes, é preciso aplicar unha transformación para que sexan *incorreladas* e poder tratar con elas no modelo de regresión. Se usamos a transformación *ilr* que mencionabamos anteriormente, só a primeira coordenada desa transformación é facilmente interpretable. Por esa razón, imos repetir o proceso un número $D = 4$ veces, permutando cada compoñente pola primeira posición. Desta forma, a primeira coordenada do primeiro modelo daranos información sobre a parte relativa da primeira compoñente con respecto ás restantes; a primeira coordenada do segundo modelo (permutando a primeira compoñente pola segunda) daranos información sobre a segunda compoñente con respecto ás restantes, e así sucesivamente. Pódese ver o traballo de Muller *et al.* [4] onde aparecen máis detalles desta metodoloxía.

A táboa 6 mostra a estimación dos coeficientes e os *p-valores* para cada unha das partes composicionais.

Coeficientes	Estimación	Std. Erro	z *value*	P valor
(Intercept)	3,68597	0,74777	4,929	8,25e-07***
Transporte	-0,05261	0,07554	-0,696	0,48619
Manutención	-0,06477	0,12470	-0,519	0,60350
Aloxamento	-0,06994	0,08396	-0,833	0,40480
Varios	0,18732	0,06877	2,724	0,00645**
Inflúe o aloxamento na elección	-0,54477	0,17730	-3,073	0,00212**
Idade 30-35	-1,29311	0,51468	-2,512	0,01199*
Idade >50	-0,50635	0,94873	-0,534	0,59354
Acompañante	-1,60917	0,60992	-2,638	0,00833**

Táboa 6. Coeficientes da análise de regresión loxística

Parece que só é significativo o *log-ratio* de gastos varios fronte á media xeométrica das outras compoñentes. Lembremos que as variables composicionais van nun lote; por tanto, non podemos extraer do modelo as partes que non parecen significativas porque están todas vinculadas.

Do modelo anterior podemos dicir que a probabilidade estimada de utilizar vivendas de uso turístico para unha persoa que pertenza ás categorías de referencia (acompañado e do grupo de idade 20-35) e cunha importancia media (3) á hora de elixir aloxamento, vén dada por 0,89, resultante de $1/(1+exp(3{,}685965 - 0{,}5447671 * 3))$. Con todo, se viaxa só, esta probabilidade estimada baixa ata un 0,61.

No tramo de idade de 35 a 50, estas probabilidades baixan a 0,68 e 0,30 se viaxan acompañados ou sós respectivamente. No grupo de maior idade, máis de 50, volve aumentar estas probabilidades a un 0,82 e 0,48. Isto dá lugar a pensar que a elección deste tipo de vivendas está moi determinada polo rango de idade do ou da turista.

Por outra banda, o coeficiente da variable que indica se inflúe o aloxamento na elección do destino mostra que o logaritmo do *odds ratio* diminúe en 0,54 unidades ao aumentar a valoración desta pregunta nunha unidade.

Interpretación das variables composicionais

Podemos ver como a variable composicional é significativa á hora de explicar a utilización das vivendas de uso turístico (polo menos unha das compoñentes é significativa). O valor do coeficiente da compoñente de gastos varios é significativo e positivo, o que indica que se o predominio relativo dos gastos varios na repartición se duplica (con respecto á contribución media das outras partes), mantendo fixas as outras covarianzas, a probabilidade de utilizar este tipo de vivendas $(P(I_{UsoVivenda} = 1))$ verase incrementada nun 20 % aproximadamente (exp(0,18732) = 1,206013).

3. Análise temporal

Outro dos obxectivos do estudo era ver a aplicación desta metodoloxía baseada en datos composicionais na análise de series de tempo no traballo sobre o número de persoas viaxeiras mensuais en Ourense (fonte IGE). Imos considerar as seguintes series de viaxeiros/as en hoteis e residentes *(hotres)*, en hoteis e estranxeiros *(hotextr)*, en turismo rural e residente *(rurresi)* e, finalmente, en turismo rural e estranxeiros *(rurextr)*. Consideramos esas catro categorías porque só temos datos suficientes para a provincia de Ourense para a modelización composicional deste tipo de persoas viaxeiras. Os datos recollidos son viaxeiros e viaxeiras mensuais na provincia de Ourense dende xaneiro de 2005 ata xuño de 2021.

Comezamos por debuxar as series (figura 9). Podemos ver o comportamento periódico de cada unha das series, aínda que a variabilidade e, por suposto, a escala é diferente en cada unha delas.

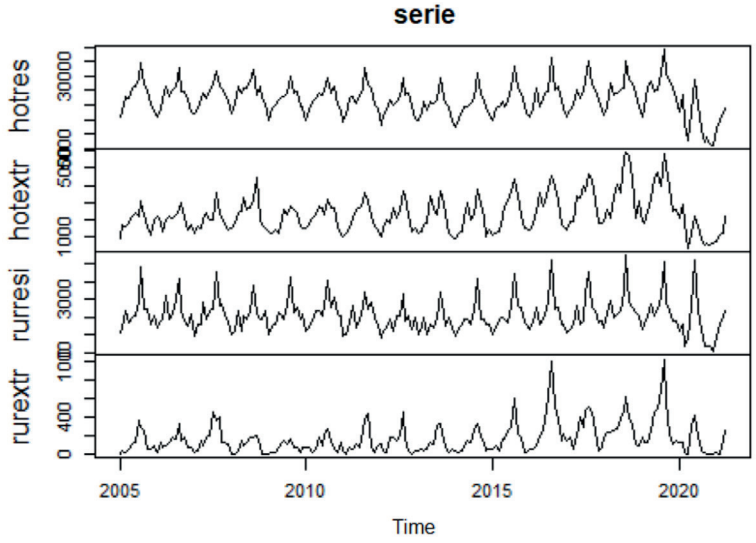

Figura 9. Series de datos orixinais

Un dos modelos máis utilizados para o tratamento de series multivariantes é o Vector Auto Regression (VAR). Sexa $\mathrm{x}_t = (x_{1,t}, \ldots x_{d,t})$ e sexa $\mathrm{z}_t = (z_{1,t}, \ldots z_{d-1,t})$ a súa representación establécese en coordenadas *ilr*. O modelo VAR(p) vén dado por:

$$\mathrm{z}_t = c + A^1 \mathrm{z}_{t-1} + A^2 \mathrm{z}_{t-2} + \ldots + A^p \mathrm{z}_{t-p} + \epsilon,$$

onde c é un vector real e A^i son matrices de parámetros. O termo de erro vén representado por ϵ.

A idea consiste en facer a transformación dos datos orixinais e logo aplicarlle o método multivariante á serie con termos *incorrelados*. Unha vez feita a estimación, sería necesario aplicar a transformación inversa para poder facer predicións. No traballo de Kynčlová, Filzmoser e Hron [10] podemos ver unha aplicación destes modelos a unha serie composicional.

Por tanto, comezaremos por calcular as coordenadas *ilr* que non están relacionadas e así poder aplicar o modelo VAR sobre elas. Lémbrese que neste proceso perdemos unha dimensión; polo tanto, pasamos de catro variables orixinais a tres variables transformadas.

Imos engadirlle un elemento á serie que consiste no número total de persoas viaxeiras resultante de sumar as catro series, ademais das catro categorías que vimos anteriormente. Nalgúns estudos da análise composicional reflíctese que á parte das composicións que representan a ratio ou a proporción de cada parte, é interesante ter o valor global, é dicir, non só interesa saber a repartición en proporción, senón o total, xa que o valor total vai cambiando de mes a mes. Isto non acontece no noso caso, non é o mesmo o número total de visitantes en xaneiro ca en agosto, polo que nos parece axeitado considerar esta variable tamén na serie. Por tanto, a nosa serie consta das tres compoñentes resultantes da transformación *ilr*, mais a variable *tot* que indica o número total de persoas viaxeiras mensuais.

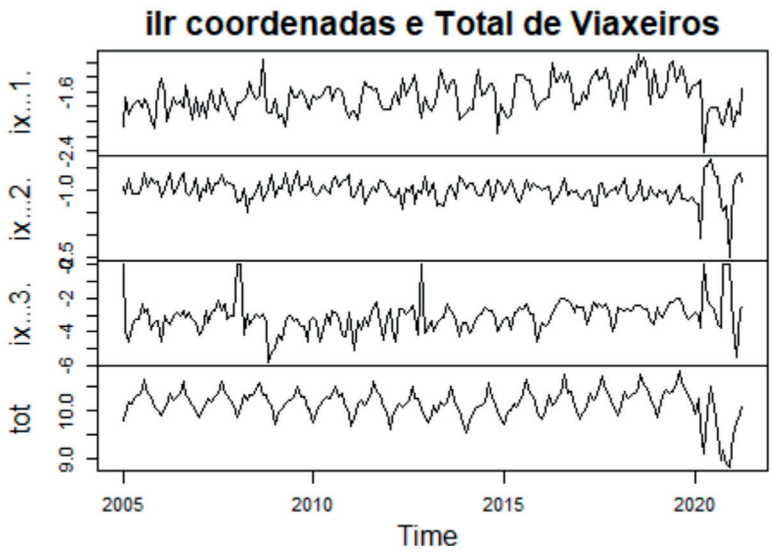

Figura 10. Series de datos transformadas mais a serie total

A continuación, imos comprobar se as series resultantes son estacionarias mediante o *Augmented Dickey-Fuller Test*. Unha vez aplicado un proceso de diferenciación da serie, resultou un *p*-valor do test menor ca 0,05 para todas as compoñentes, polo que se pode asumir a estacionariedade en cada termo da serie.

No seguinte paso, deberiamos saber o valor de p para coñecer ata que retardo é significativo o modelo para facer predicións. Os criterios que utilizamos para facer esta selección foron *Akaike information criterion* (AIC), *Hannan-Quin criterion* (HQ), *Schwarz criterion* (SC) e *Final Prediction error* (FPE). Todos eles están baseados no estimador máximo verosímil da matriz residual de covarianza, da dimensión da serie e do número de observacións. Despois de levar a cabo un proceso de diferenciación da serie para que fose estacionaria, estimamos o número de retardos necesarios para o modelo VAR. Na táboa 7 podemos ver os resultados.

AIC(n)	HQ(n)	SC(n)	FPE(n)
17	8	2	8

Táboa 7. Número de retardos (p) para estimar do modelo VAR (p) coa serie completa

Dous destes criterios coinciden cun número de retardos de 8. É dicir, o que aconteza no mes de agosto dun ano vén dado en función do que aconteza nos oito meses

anteriores. Tendo en conta que os datos considerados teñen incluído o período máis cru da pandemia, no que a inmensa parte da poboación estaba confinada e o número de persoas viaxeiras cambiou drasticamente, quixemos repetir a análise coas observacións ata marzo de 2020 e repetir o proceso.

AIC(n)	HQ(n)	SC(n)	FPE(n)
15	4	4	14

Táboa 8. Número de retardos (p) para estimar do modelo VAR (p) coa recortada

Como vemos, neste caso (táboa 8), o retardo aumentou o número de meses para estes todos criterios agás o AIC que diminuíu. Vemos aquí que o efecto da covid-19 fai que a análise dunha simple serie temporal teña que facerse con moita cautela, pois eses datos distorsionan por completo a estimación do modelo e, por tanto, as predicións. Sería máis aconsellable utilizar un modelo non paramétrico que sexa máis flexible e así poder axustar mellor este tipo de datos. Se realizamos a estimación supoñendo $p=15$, podemos ver como serían as predicións. Para coñecer os valores estimados da repartición por tipoloxía de turista, só habería que facer as transformacións inversas.

Se comparamos os valores reais cos estimados (táboa 9) para o total das persoas viaxeiras, podemos ver claramente as diferenzas existentes para os meses de marzo, xuño e xullo.

Serie total de viaxeiros/as	Marzo	Xuño	Xullo
Valor real	15777	9088	21975
Valor estimado	34237	39822	35127

Táboa 9. Estimación e valor real do numero de visitantes totais dos meses de marzo, xuño e xullo de 2020

Queda para futuras investigacións explorar outros métodos máis avanzados que poidan mellorar este tipo de estimacións aplicados ao caso de datos con estrutura composicional.

4. Conclusións

En resumo, a análise da repartición do gasto é significativo á hora de explicar a utilización das vivendas de uso turístico e, en xeral, para analizar calquera tipo de repartición. Cabe recordar que dadas as limitacións no proceso de selección da mostra, que xa comentamos previamente, non é posible a extrapolación dos resultados para toda a poboación.

Con este estudo, acadamos o obxectivo de estudar un sector do turismo pouco coñecido, como son as vivendas de uso turístico, dende unha nova perspectiva. Utilizamos unha metodoloxía pouco frecuente da análise de datos composicionais e vimos, con toda a cautela sobre a mostra, que este tipo de datos podían ser de interese para realizar este tipo de análise.

Agradecementos

Esta investigación foi financiada pola Deputación Provincial de Ourense mediante as axudas concedidas na convocatoria de axudas a grupos de investigación do campus de Ourense (INOU 2021-01B) ao proxecto con título «Análise estatística do mercado de aloxamento turístico da provincia de Ourense dende unha nova perspectiva: a análise de datos composicionais».

Bibliografía

[1] Egozcue, J. J, and V Pawlowsky-Glahn. 2016. "What are compositional data and how should they be analyzed?" *Boletín de Estadística E Investigación Operativa* 32 (1): 5–29. https://doi.org/10.1023/A:1023818214614.

[2] Nguyen, T. H. A., T. Laurent, C. Thomas-Agnan, and A. Ruiz-Gazen. 2020. "Analyzing the impacts of socio-economic factors on French departmental elections with CoDa methods." *Journal of Applied Statistics*, 1–17. https://doi.org/10.1080/02664763.2020.1858274.

[3] Gibergans-Baguena, Jose, Carme Hervada-Sala, and Eusebi Jarauta-Bragulat. 2020. "The Quality of Urban Air in Barcelona: A New Approach Applying Compositional Data Analysis Methods." *Emerging Science Journal* 4 (2): 113–21. https://doi.org/10.28991/esj-2020-01215.

[4] Muller, Ivo, Karel Hron, Eva Fiserova, Jan Smahaj, Panajotis Cakirpaloglu, and Jana Vancakova. 2018. "Interpretation of Compositional Regression with Application to Time Budget Analysis." *Austrian Journal of Statistics* 47 (2): 3–19. https://doi.org/10.17713/ajs.v47i2.652.

[5] Coenders, Germà, and Berta Ferrer-Rosell. 2020. "Compositional Data Analysis in Tourism: Review and Future Directions." *Tourism Analysis* 25 (1): 153–68. https://doi.org/10.3727/108354220x15758301241594.

[6] R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

[7 Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. 2003. "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology* 35 (3): 279–300. https://doi.org/10.1023/a:1023818214614.

[8] Martín-Fernández, J. A., C Barceló-Vidal, and V Pawlowsky-Glahn. 2003. "Dealin and Missing Values in Compositional Data Sets Using Non-parametric Imputation." *Mathematical Geology* 35 (3): 253–78. https://doi.org/10.1023/a:1023866030544.

[9] Pawlowsky-Glahn, V., and J. J. Egozcue. 2001. "Geometric Approach to Statistical Analysis on the Simplex" 15: 384–98. https://doi.org/10.1007/s004770100077.

[10] Kynčlová, Petra, Peter Filzmoser, and Karel Hron. 2015. "Modeling Compositional Time Series with Vector Autoregressive Models." *Journal of Forecasting* 34 (4): 303–14. https://doi.org/10.1002/for.2336.

Proxectos INOU 2021

Investigación aplicada na provincia de Ourense

Coordinación:

Vicerreitoría do
Campus de Ourense-Campus Auga

DEPUTACION OURENSE

Vicerreitoría do
Campus de Ourense
Universidade Vigo

Proxectos INOU 2021.

Investigación aplicada na provincia de Ourense

Coordinación:

Vicerreitoría do
Campus de Ourense-Campus Auga

Ourense, 2022

Universidade de Vigo • Campus de Ourense

Proxectos INOU 2021. Investigación aplicada na provincia de Ourense

Autores/as:
De Carlos Villamarín, Pablo
Pérez González, Ana
Fernández González, María
Laza Fidalgo, Rosalía
Isorna Folgar, Manuel
Pérez Rodríguez, Francisco Javier
Raposo Rivas, Manuela
Alfonso Gil, Sonia
Rodríguez Teijeiro, Domingo
Casado Neira, David
García Pérez-Schofield, José Baltasar

Coordinación:
Vicerreitoría do Campus de Ourense-Campus Auga

Comisión de Avaliación:
de Blas Varela, Esther
Cid Fernández, Xosé Manuel
Fernández Gil, César Manuel
García Queijeiro, José Manuel
Gómez Rodríguez, Alma
Reboreda Morillo, Susana
Rodeiro Iglesias, Javier

Nº de páxinas: 224

ISBN: 978-84-8158-949-8

Edición
Vicerreitoría do Campus de Ourense - Campus Auga
www.uvigo.gal/campus/ourense-campus-auga

Maquetación
Rodi Artes Gráficas, S. L.

Índice