CUERPO ACADÉMICO DE PROBABILIDAD Y ESTADÍSTICA BUAP

Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

CPO-GRAMA: UNA HERRAMIENTA DE VISUALIZACIÓN A TRAVÉS DEL ANÁLISIS DE DATOS COMPOSICIONALES PARA LA ESTRUCTURA MULTIVARIANTE DE LOS COMPONENTES DEL CPO

RAMÓN ÁLVAREZ-VAZ[a], FERNANDO MASSA[a], MATÍAS MUÑOZ WOLF[b]

[a]*Instituto de Estadística,Departamento de Métodos Cuantitativos, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo, Uruguay*
[a]ramon@iesta.edu.uy, [a]fmassa@iesta.edu.uy

[b]Área de Epidemiología y Estadística, Comisión Honoraria de Salud Cardiovascular,Montevideo, Uruguay
[b]matias.mw@gmail.com

Resumen

En los estudios epidemiológicos en Salud Bucal que analizan los componentes del CPO (piezas dentales cariadas, perdidas y obturadas), es habitual hacerlo en forma separada, lo que lleva a una pérdida de información relevante que está contenida en la estructura multivariada de los datos. Por ese motivo resulta fundamental encontrar una forma de estudiar los 3 componentes del CPO en conjunto. En el presente trabajo, se propone una alternativa gráfica que se denominará CPO-grama, basado en el análisis de datos composicionales, que permitirá explorar las relaciones 2 a 2 de cada componente del CPO pero también su valor global, permitiendo visualizar en forma simultánea cada componente, el CPO y otros atributos de cada persona estudiada. Se aplica esta metodología en el relevamiento en población que se asiste en Facultad de Odontología, Udelar durante 2015-2016. Se logran identificar patrones de comportamiento, que permiten decir que las personas del estudio se caracterizan por tener un fuerte predominio del componente de dientes perdidos, que se asocia con un elevado nivel de CPO. Ese patrón no aparece estar asociado al sexo aunque sí con la edad o el ingreso de las personas. Se propone, considerando que se trata de datos composicionales, evaluar como sería el procedimiento de clustering, buscando crear grupos y de contrastar otros atributos cuantitativos, mediante curvas de nivel.

Keywords: CPO, Datos composicionales, Estructura multivariada.

1. Introducción

En los estudios epidemiológicos que analizan los componentes del CPO, es habitual hacerlo en forma separada, lo que lleva a una pérdida relevante de información que está contenida en la estructura multivariada de los datos. Como ya se manejó en capítulos anteriores un mismo valor de CPO de 12 puede estar indicando situaciones muy diversas, como de una persona con 8 piezas obturadas y 4 con caries, y de otra con 5 cariadas y 7 perdidas. En ambos casos, los niveles de enfermedad son importantes (tienen 12/28 % de su piezas afectadas, es decir 'no sanas') pero no se sabe si la carga de enfermedad es la misma, ya que las piezas obturadas ponen de manifiesto enfermedad pasada.

2. Objetivos

Se propone encontrar una forma de estudiar los 3 componentes del CPO (se deja de lado en principio el componente S), sin pérdida de información que refleje toda la esencia del CPO, por lo cual se presenta una alternativa gráfica que se denominará *CPO-grama*, que permitirá explorar las relaciones 2 a 2 de cada componente del CPO pero también su valor global, con la posibilidad de poder visualizar en forma simultánea cada componente, el CPO y otros atributos de cada persona analizada. Para eso es necesario previamente hacer una breve introducción a la metodología estadística que sustenta el método visual. Por lo tanto esta nueva forma de análisis es de tipo descriptivo a diferencia de otras técnicas que en capítulos anteriores intentaban modelizar

los componentes del CPO.

3. Metodología

Tradicionalmente, un conjunto de datos se llaman composicionales si éstos representan proporciones o partes de un total: porcentajes de trabajadores en diferentes sectores, porciones de los elementos químicos en un mineral, concentración de diferentes tipos de células en la sangre de un paciente, porciones de especies en un ecosistema o en una trampa, concentración de nutrientes en una bebida, porciones del tiempo de trabajo dedicado a diferentes tareas, porciones de tipos de fallas, porcentajes de votos para partidos políticos, etc.

El análisis sobre este tipo de datos, es un caso particular de lo que se denomina análisis de datos composicionales. Las partes individuales de la composición se denominan componentes. Cada componente tiene una cantidad, representando su importancia dentro del conjunto. La suma sobre las cantidades de todos los componentes se llama la cantidad total. Las porciones son las cantidades individuales divididas por esta cantidad total. Es decir, las variables originales se transforman en porcentajes que tienen una suma constante de 100 % por individuo.

Mas formalmente desde el punto de vista estadístico un dato *composicional*, [1], [2], [3], [4], [5] [6], [7] es un vector x cuyas componentes $(x_1, x_2, ..., x_D)$, estrictamente positivas, representan *partes* de un *todo*, por lo que x se encuentra sujeto a la siguiente restricción:

$$\sum_{i=1}^{D} x_i = k$$

- Al multiplicar una composición por una constante, la composición obtenida es la misma;
- Todos los vectores de D componentes positivas que son proporcionales; resultan equivalentes y representan la misma composición;
- Por lo general, se selecciona un *representante* de la composición.
- Se define un Operador *clausura C* - Correspondencia entre un vector $w = (w_1, w_2, ..., w_D)$ de componentes positivas y su dato *composicional* asociado;

$$x = (x_1, x_2, ..., x_D) \to C(w) = k\left(\frac{w_1}{\sum_{i=1}^{D} w_i}, \frac{w_2}{\sum_{i=1}^{D} w_i}, ..., \frac{w_D}{\sum_{i=1}^{D} w_i}\right) \quad (1)$$

- Las componentes del vector clausurado se denominan *partes*, sobre el *total* k, y definen el siguiente espacio (*simplex*):

$$S^D = \{(x_1, x_2, ..., x_D) / x_i > 0, i = 1, ..., D, \sum_{i=1}^{D} x_i = k\} \quad (2)$$

4. Visualización a través de *GT*

Los *GT* son un tipo de gráfico baricéntrico que permiten trabajar a la vez con 3 variables que tienen la característica de tener una suma constante por observación; son un caso particular (para 3 variables) de lo que ya se presentó.

En un GT que también se conoce como **ternary plot**, las proporciones de las tres variables a, b, y c deben sumar una constante, K. De esta manera hay solamente 2 variables que pueden fluctuar libremente debida a la restricción de que $a + b + c = K$ para todas las observaciones- sólo hay dos grados de libertad - es posible representar gráficamente la intersección de las tres variables en sólo dos dimensiones.

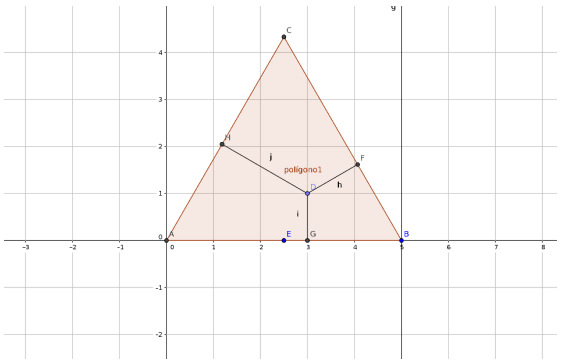

Fig. 1. Ejemplo de gráfico triangular básico

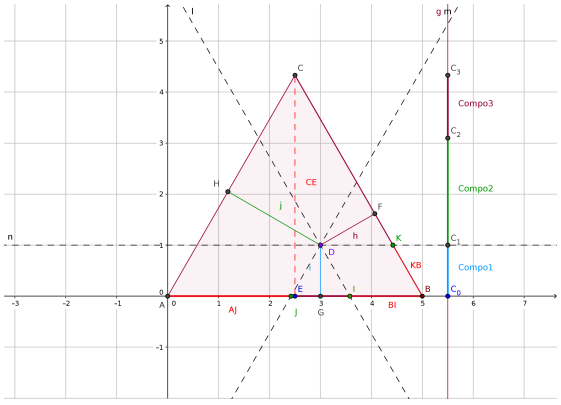

Fig. 2. Ejemplo de gráfico triangular y sus componentes

Si se analiza la posición que ocupa un punto cualquiera interior al triángulo equilátero que aparece en la figura 2, interesa ver cual es la relación entre las longitudes de los segmentos $\bar{HD}, \bar{GD}, \bar{FD}$; la misma relación prevalece entre las longitudes de los segmentos $\bar{AJ}, \bar{BI}, \bar{KB}$. Puede observarse por otra parte que las relaciones antes mencionados pueden verse en las proyecciones perpendiculares que de estas se hacen en el eje perpendicular a la base del triángulo y que está a la derecha y situado por fuera, en los que se representan

los 3 segmentos $C_0\bar{C}_1, C_1\bar{C}_2, C_2\bar{C}_3$, que no son más que homotecias de los segmentos $\bar{AJ}, \bar{BI}, \bar{KB}$. Si los 3 lados del triángulo se usan para representar variables estadísticas, cuya suma tiene sentido (magnitudes económicas, físicas, etc) y es constante, para el caso de un triángulo cuyo lado tiene longitud 100, se está en presencia de lo que se llama *GT*. En este caso el punto interior al triángulo permite evaluar que % de cada variable estadística se tiene de la variable suma. Si por ejemplo se está analizando magnitudes continuas cada punto interior al triángulo indica que valor tiene en cada una de las 3 variables (son 3 coordenadas) las que a su vez representan que proporción de la variable suma tiene cada una de éstas.

Los GT tienen su origen en disciplinas como la geología, donde interesa ver que fraccción o proporción de un mineral componen un compuesto como puede ser la arena que tiene, u otros elementos. Tienen la ventaja de ser de fácil interpretación y no son muchas los paquetes estadísticos [8], [9] que tiene implementados este tipo de gráfico. La desventaja que puede tener este tipo de gráfico es que cuando la cantidad de observaciones es muy numerosa, puede ser difícil su interpretación, aunque si pueden mostrar esencialmente patrones de dispersión en los gráficos. Por lo tanto una vez presentada la forma de construcción de un GT, a continuación en la seccción 5 se presenta una aplicación de los mismos como solución al problema presentado en la seccción 1.

5. Aplicación

Los datos provienen del estudio sobre personas que demandan atención en la Facultad de Odontología de la Universidad de la República, Uruguay y que son evaluados por los odontólogos del Servicio de registros de la Facultad, desarrollado en el marco del proyecto 'I+D' de la CSIC 2014. Se aplica una muestra de 602 personas que consultan en el período que corresponde a mayo 2015-junio 2016, los que se seleccionan mediante muestreo sistemático, a los que se les aplica un cuestionario sociodemográfico y un examen completo de la boca, en donde se evalúa el estado de las piezas dentales y de la mucosa, además de medidas antropométricas, de PA y de glicemia. El tamaño muestral se determinó para poder medir prevalencias de hasta 25% con un margen de error $\delta = 0.05$ y un nivel de confianza $1 - \alpha = 0.95$ y cubrir hasta una tasa de no respuesta del 90%. Finalmente de los 640 originalmente calculados se obtuvieron 602, que representa una fracción de muestreo de alrededor del 15% del total de personas que consultan anualmente.

6. Resultados

Se presentan las distribuciones univariadas de los 3 componentes del CPO y el CPO global

En función de los valores para el mínimo del CPO

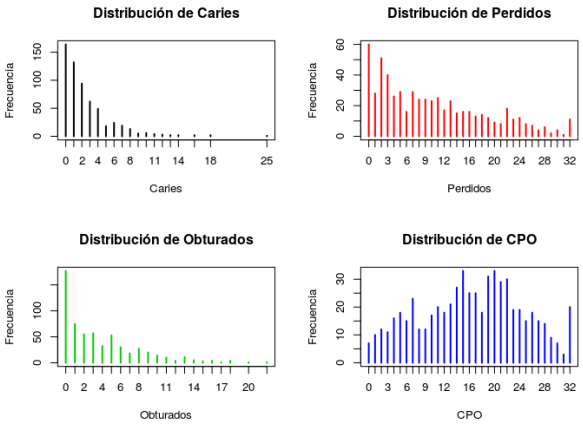

Fig. 3. Distribución de cada componente del CPO en forma univariada

Componente	n	$\bar{x}$	mediana	mínimo	máximo
Caries	602.00	2.50	2.00	0.00	25.00
Perdido	602.00	10.17	8.00	0.00	32.00
Obturado	602.00	3.66	2.00	0.00	22.00
CPO	602.00	16.33	17.00	0.00	32.00

y que se pueda operar como se presenta en (1), es necesario quitar las observaciones que corresponden a individuos totalmente sanos, es decir que tiene CPO=0. Quedan finalmente 595 encuestados, los que se analizan a continuación. Como la Figura 3, solo muestra la distribución univariada, sigue sin resolverse, el problema de no perder la estructura mutivariada, por ejemplo saber si la personas con bajo nivel de Caries, tienen bajo nivel de piezas perdidas o ambos atributos son independientes, por lo cual se da un paso más en la visualización y se obtiene la Figura 4.

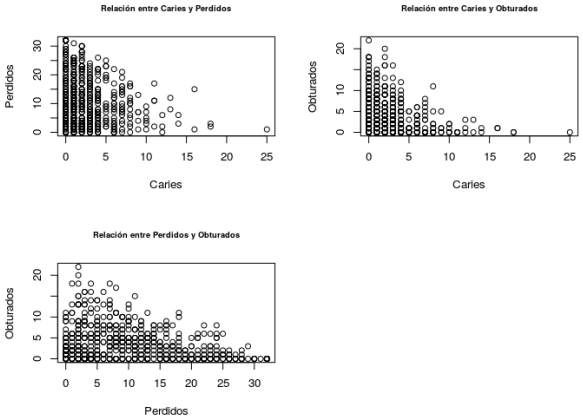

Fig. 4. Relaciones 2 a 2 entre componentes del CPO por separado

La Figura 5 podría ayudar a ver la relación entre

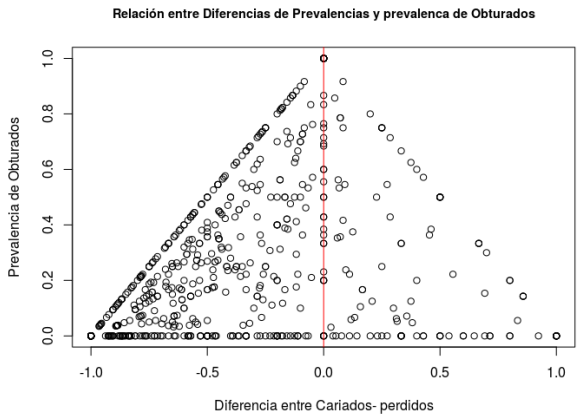

Fig. 5. Relaciones entre diferencias entre prevalencias de los componentes del CPO

los 3 componentes, ya que el gráfico presentado es un gráfico de dispersión que debe interpretarse del siguiente modo: Al tener en el eje de las abcisas la diferencia entre prevalencias el eje de simetría pintado en rojo indica que los puntos que están por debajo de 0, son personas que tienen más proporción de dientes perdidos que de dientes cariados. Y luego de que sabe de que lado del eje están cuanto más elevado sea su coordenada en el eje vertical indica cuanta más proporción de dientes Obturados tiene, con un evidente descenso en la diferencia de prevalencias de perdidos y cariados. Es claro que hay un patrón de los puntos pero resulta de difícil interpretación. Incluso los puntos que caen sobre los lados del triángulo que se formaron son de más compleja interpretación, por lo cual se opta por trabajar con *GT*, definidos antes, los que por ser gráficos que revelan el patrón de asociación de los componentes del CPO, se denominarán *CPOgr*, de ahora en adelante.

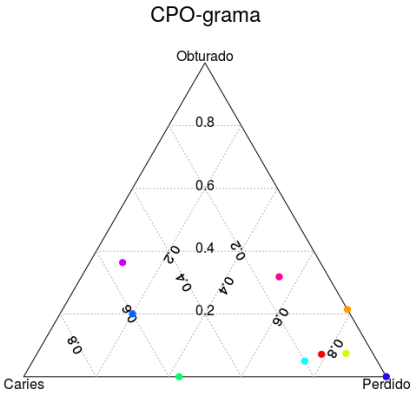

Fig. 6. Ejemplo de CPO-grama para 10 individuos

Se opta entonces por hacer un análisis gradualista tratando de entender lo que representan los *CPOgr*. La Figura 6 presenta como es la relación por ejemplo de 10 individuos para los 3 componentes y en función de la explicación dada en la Figura 2, para cada componente se traza una línea parela a la base opuesta al vértice del componente estudiado, por ejemplo Perdido, con lo cual de los 10 puntos hay una persona que tiene 60% de Caries, 20% de obturado y por construcción (las 3 proporciones deben sumar 100), tiene 20%. Por otra parte si se observa el individuo que está en el vértice derecho abajo donde aparece Perdido en color azul oscuro, se puede aseverar que es una persona que tiene el 100% de sus piezas cariadas. Cuando las observaciones se encuentran en algunos de los lados del tríangulo equilátero en el que se basa el *CPOgr*, se puede decir que éstos tiene 0% del componente opuesto al cateto donde se encuentran, tal es el caso de una observación en color naranja que tiene 0% de Caries , 20% de Obturados y 80% de piezas perdidas, mientras que la observación en verde más intenso se caracteriza por no tener piezas obturadas repartiendo el CPO entre 40% de piezas pedidas y el resto en piezas con Caries.

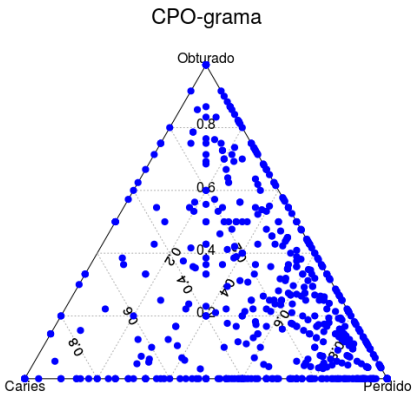

Fig. 7. CPO-grama completo

En las Figuras 8 y 9 aparecen los resultados a través de los *CPOgr*, que permiten encontrar patrones de comportamiento asociadas a otros atributos como son el sexo y por ejemplo el nivel de CPO, ya que una restricción que tienen los *GT*, que al ser representaciones gráficas de datos composicionales, mantienen la invarianza y pueden haber 2 *CPOgr* iguales a pesar de que los niveles de CPO sea uno la mitad que el otro.

Para poder contrastar la asociación de C, P y O, se elabora una nueva variable que consiste en categorizar en CPO en 4 categorías que podrían asimilarse a cuartiles y que se muestran en la Tabla 1.

Sexo	[1,11]	(11,17]	(17,22]	(22,32]	Total
Masculino	75	66	50	54	245
Femenino	91	83	91	85	350
Total	166	149	141	139	595

Tabla 1. Distribución de personas por categoría de CPO según sexo

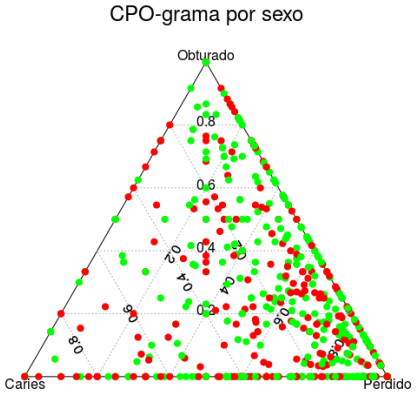

Fig. 8. CPO-grama por sexo

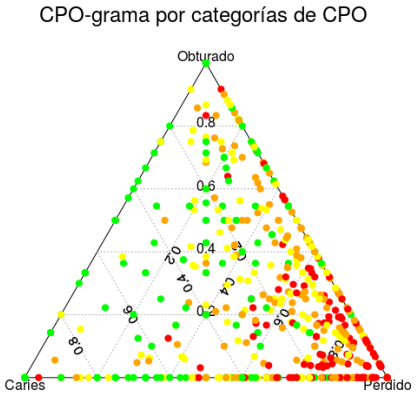

Fig. 9. CPO-grama por nivel de CPO

7. Discusión

Los resultados se presentan mediante una estrategia gradualista permitiendo al investigador en Biomedicina ir captando cada vez con más detalle el fenómeno del CPO, partiendo de distribuciones univariadas, donde se nota un gran exceso de 0, aspecto ya conocido e identificado en los anteriores trabajos donde de trabajó con el estudio *RPAFO2015*, pero donde queda claro que las distribuciones univariadas son insuficientes para comprender exactamente el problema, del mismo modo que las distribuciones 2 a 2 donde siempre resta por ver que sucede con el tercer componente del CPO que no se manejó en ese gráfico bivariado. El esfuerzo a través del gráfico en forma de embudo invertido que aparece en la Figura 5 para el investigador que no conoce los *GT*, no alcanza aún a contestar un fenómeno que es trivariado representado en una imagen plana, algo que si se logra con el *GT*.

Si ahora se presta atención a la Figura 7, puede verse que existe un patrón muy marcado en las relaciones de los componentes del CPO, donde hay un predominio del porcentaje de piezas perdidas, con una zona muy cargada en el vértice derecho inferior; por otra parte si se examinan los lados del *GT*, se observa que la densidad es menor en el lado que corresponde al que une el componente Caries con Obturado, lo que significa que son pocas las personas del estudio que solo tiene piezas cariadas y Obturadas. Por último antes de pasar a un análisis más detallado que tome en cuenta otros atributos, puede decirse que la densidad en la parte central de la Figura 7 es baja, lo que debe interpretarse que son pocas las personas que reparten por igual Perdidos y Cariados y dientes Obturados.

Si ahora se evalúan atributos extras como el sexo la Figura 8 muestra que no hay un patrón claro de que algún componente concentre más individuos de un sexo que de otro.

Cuando ahora el contraste de la asociación de componentes del CPO se hace en función del nivel del mismo (aspecto hasta ahora no se ha tenido en cuenta), surge un claro patrón de que no solamente predomina el componente de diente perdidos, sino que además los puntos en colores naranja (CPO en el tramo $(17,22]$) y rojo (CPO en el tramo $(22,32]$), son los que allí aparecen dando una idea de la carga de enfermedad donde está concentrada. En oposición a este realidad, surge que los puntos verdes (CPO=$(17,22]$) se dan en los lados del triángulo opuesto al vértice de Perdidos, indicando que son personas con Caries o piezas Obturadas y bajo nivel de CPO, es decir muchas piezas sanas. Algunos otros puntos verdes se dan en general lejanos al componente de perdidos y en la zonas más baricéntrica del gráfico.

8. Conclusiones y pasos a futuro

En este trabajo se presenta una alternativa gráfica para la visualización del CPO sin perder la esencia de la estructura multivariada del mismo. Se hace a través de un tipo de gráfico llamado *GT*, el que se denominará *CPOgr*, y que no es más que un caso particular del *ADC*. En este nuevo tipo de herramienta visual (nueva para el trabajador del área biomédica pero ya muy usada en otras disciplinas), se logra resolver el objetivo, sin pedir información extra, y no perder la estructura multivariada del fenómeno. Se logran identificar patrones de comportamiento, que permiten decir que las personas del estudio *RPAFO2015*, se caracterizan por tener un fuerte predominio del componente de dientes perdidos, que se asocia con un elevado nivel de CPO. Es muy probable que este patrón pueda no estar asociado al sexo aunque si lo pueda estar con la edad o por ejemplo el ingreso de las personas. Esto se podría corroborar en lugar de constrastar por CPO (en términos de colores se haga por nivel de ingreso), pero nuevamente hay una restricción al quitar el atributo de nivel de CPO para sustituirlo por otro.

Queda entonces como desafío lograr en una imagen plana que maneja 3 variables (con la restricción de que tienen suma constante), lograr incorporar más información de la estructura multivariada. Ya el contraste por colores se logra y se propone tratar por ejemplo que una quinta variable como el ingreso, se pueda considerar al variar la forma del punto graficado; por otro lado si se quiere considerar además una variable de tipo cuantitativa se podría solapar al *CPOgr*, curvas de nivel en función de la variable cuantitativa, para poner de manifiesto un patrón, si existiese. Resta entonces lograr que en este caso ese planteo se logre resolver mediante alguna subrutina de graficación más potente. Actualmente hay una serie de librerías del R que funcionan en combinación con la librería *ggplot*, [10], que elaboran gráficos de alto nivel y que podrían tal vez resolver ese problema de combinar múliples atributos y curvas de nivel. Por último, considerando que se trata de datos composicionales, resta por evaluar como sería el procedimiento de clustering, buscando crear grupos, sabiendo de las restricciones en los datos establecidas en la ecuación (1) y (2).

Agradecimientos

Los autores quieren agradecer a la Prof. Agregada. Dra. Susana Lorenzo-Erro y la Prof. Asociada Magíster Anunzziatta Fabruccini del Servicio de Edidemiología y Estadística de la Facultad de Odontología de la Universidad de la República, Montevideo, Uruguay coordinadoras del relevamiento en el cual se basa este trabajo.

.

Referencias

[1] J. Aitchison, "The statistical analysis of compositional data." *Journal of the Royal Statistical Society. Series B (Methodological).*, vol. 44, no. 2, pp. 139 – 177, 1982.

[2] J. Aitchison, *The Statistical Analysis of Compositional Data*, ser. Monographs on Statistics and Applied Probability. Springer Netherlands, 1986.

[3] J. Aitchison and J. J. Egozcue, "Compositional data analysis: Where are we and where should we be heading?." *Mathematical Geology*, vol. 37, no. 7, pp. 829–850, 2005.

[4] C. Barceló-Vidal, J. Martín-Fernández, and V. Pawlowsky-Glahn, "Mathematical foundations of compositional data analysis." *Ross, G., ed., Proceedings of IAMG01, The sixth annual conference of the International Association for Mathematical Geology: Cancun, México, 20p.*, 2001.

[5] J. Egozcue, R. Tolosana-Delgado, E. Jarauta-Bragulat, M. Ortego, and J. Díaz-Barrero, "Análisis de datos composicionales: aguas, contaminantes, recursos, sociología..." in *Recerca i innovació a l'escola de camins.*, 2011.

[6] K. G. van den Boogaart, R. Tolosana, and M. Bren, *compositions: Compositional Data Analysis*, 2014, r package version 1.40-1. [Online]. Available: `https://CRAN.R-project.org/package=compositions`

[7] A. Amador Rescalvo, "Un modelo de regresión para datos en el simplex d-dimensional." Master's thesis, Universidad Autónoma Metropolitana, Ciudad de México, México., 2017.

[8] D. Chessel, A. B. Dufour, and J. Thioulouse, "The ade4 package — I: One-table methods," *R News*, vol. 4, no. 1, pp. 5–10, June 2004. [Online]. Available: `#http#`

[9] D. Meyer, A. Zeileis, and K. Hornik, *vcd: Visualizing Categorical Data*, 2016, r package version 1.4-3.

[10] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. [Online]. Available: `http://ggplot2.org`