ANÁLISIS DE DATOS COMPOSICIONALES: AGUAS, CONTAMINANTES, RECURSOS, SOCIOLOGÍA...

J.J. Egozcue [1], R. Tolosana-Delgado [2],E. Jarauta-Bragulat [1], M.I. Ortego [1],

J.L. Díaz-Barrero[1]

[1]Dep. Matemática Aplicada III e-mail: juan.jose.egozcue@upc.edu,
[2]Dep. Ing. Hidràulica, Marítima y Ambiental (LIM), e-mail: raimon.tolosana@upc.edu

Plabras clave: símplex, concentraciones, proporciones, transformación logística

Resumen: *Los datos composicionales describen las partes de un todo. Se presentan en forma de vectores proporciones, porcentajes, concentraciones químicas etc. y aparecen en multitud de campos de la ciencia y la ingeniería. A pesar su uso frecuente en modelización, su análisis es problemático y se pueden obtener conclusiones sin sentido si no se utilizan métodos adecuados. Los métodos de análisis composicional se basan en los principios de invariancia por escala y coherencia subcomposicional. Se utilizan logaritmos de cocientes entre proporciones como coordenadas (transformaciones logísticas), que luego se analizan como variables reales ordinarias. Este tipo de análisis requiere técnicas específicas de fácil implementación. Se describen algunas herramientas descriptivas y de modelización a través de ejemplos.*

1. DATOS COMPOSICIONALES

Los metales pesados que contaminan un suelo se expresan en partes por millón (ppm de masa). Los análisis de aguas minerales se presentan en mg/litro. Las partículas en suspensión en el aire se describen por su diámetro y el grado de contaminación se expresa por la proporción de cada unos de esos diámetros. En gestión de transportes y movilidad interesa la proporción de viajeros que van a diversas estaciones o bien el modo o el motivo de su desplazamiento. La estrucura de una cartera de valores se describe por los porcentajes invertidos en diferentes valores. En las elecciones interesa el porcentaje de votos a cada candidatura. Las probabilidades de cada suceso excluyente suman la unidad. Todas estas descripciones numéricas se caracterizan porque reparten un todo (total metales pesados, total de soluto, partículas en suspensión, total de viajeros, total invertido, total de votos, la probabilidad unitaria) entre sus partes y su interés no reside en ese total sinó en la distribución sobre las partes. Pero las unidades indicadas no dan una información relevante. Nadie se preocupa si una probabilidad o proporción, en lugar de expresarse en tanto por uno, se da en porcentaje. O si una concentración química en lugar de expresarse en ppm se expresa en fracción molar. Se entiende que dos analistas trabajando en diferentes unidades deberían llegar a idénticas conclusiones. Por tanto, si una composición se multiplica con una constante positiva, el nuevo vector debe contener idéntica información, la que expresan los cocientes entre las distintas partes. Esta conclusión constituye el principio de *invariancia bajo cambio de escala*.

Los cocientes entre proporciones o partes se convierten en el centro del análisis composicional. Sin embargo, los cocientes de partes, que son cantidades positivas, tienen escala relativa. Si en una votación un candidato obtiene el 2% de los votos y en la siguiente votación obtiene el 4%, podemos decir que se ha doblado sus votos y ha obtenido un éxito

electoral. En cambio, un candidato que obtuvo el 30% de votos y consigue ahora el 32%, habiendo mejorado su posición, podemos decir que mantiene su apoyo. Por razón de la escala relativa se suelen tomar *logaritmos de los cocientes de proporciones* de forma que su escala se transforma en absoluta (diferencia de logaritmos).

Cuando se tratan composiciones químicas nunca se analizan todos los elementos sino una parte reducida de ellos. Parece lógico exigir que los resultados del análisis concernientes a un grupo de elementos (una subcomposicición) no dependa de cuáles fueron el resto de las concentraciones medidas. Esto da lugar al principio de *coherencia subcomposicional* que exige que la información no cambie cuando se extrae una subcomposición. Cualquier descriptor de datos composicionales como concentraciones medias, variabilidad de una muestra, distancias entre composiciones deben cumplir los principios mencionados si no se quiere incurrir en contradicciones, que en ocasiones pueden ser inquietantes.

Las composiciones se pueden representar en el símplex: vectores de componentes positivas que suman una constante. Se puede dotar al símplex de una estructura de espacio Euclídeo [2]. La operación de adición es la perturbación que consiste en multiplicar las composiciones componente a componente y después normalizar a suma constante. La perturbación puede asimilarse a un filtrado (lineal) en química o al incremento/decremento multiplicativo del valor de una cartera. También se establece una métrica que conforma la llamada geometría de Aitchison en el símplex.

2. CORRELACIÓN ESPÚREA

El cálculo de la correlación entre variables es la base de casi todos los métodos estadísticos. Un ejemplo muestra cómo la correlación entre concentraciones de solutos en un acuífero puede cambiar dependiendo de qué subcomposición se analice. Se analizan aguas profundas en 23 pozos para estudiar su salinidad (Moeller et al. 2008) y se obtienen las concentraciones en mg/l de los iones Na, K, Mg, Ca, Cl, SO4, HCO3. Pueden obtenerse las correlaciones de cada par de elementos. Tomemos como ejemplo los tres pares que pueden obtenerse de Cl, Na, Mg. A continuación supongamos que no se ha medido HCO3 y reducimos cada composición a tanto por cien en peso. La tabla 1 muestra las correlaciones entre las concentraciones de los pares de elementos en los dos casos mencionados. Un observador, ante la correlación entre Cl y Na (cuando se expresan en mg/l), concluye satisfecho que los incrementos de Cl corresponden a incrementos proporcionales de Na (sal marina, salmuera, ...). En cambio quién observara el % en masa de la subcomposición (sin HCO3) llegaría a la conclusión de que los incrementos de Cl corresponden a decrementos en Na. Cada uno de los pares encierra sorpresas de este tipo. Estamos ante una flagrante violación del principio de coherencia subcomposicional.

Tabla 1. Correlaciones espúreas en un análisis de aguas

	Cl Na	**Cl Ca**	**Na Ca**
mg/l	**0.99**	**0.73**	**0.65**
% masa	**-0.37**	**-0.85**	**0.00**

3. TRANSFORMACIONES LOGÍSTICAS Y ANÁLISIS EXPLORATORIO

Aitchison [1] propone una solución al análisis de datos composicionales de acuerdo con los principios que se han descrito. Consiste en estudiar el comportamiento de log-cocientes

que sean invariantes por escala, llamados log-contrastes. Los más simples son los log-cocientes de dos partes. Esto da lugar a transformaciones de los vectores de proporciones en vectores de log-contrastes que contienen toda la información de las muestras. El análisis estadístico se realiza sobre los vectores transformados. Por sus características estas transformaciones son generalizaciones de la tranformación logística o logit utilizadas en diversos campos. Una de las más importantes es la llamada transformación clr (*centered log-ratio*) que consiste en dividir cada parte entre la media geométrica de todas las partes y tomar logaritmo. Esta transformación permite utilizar la métrica euclídea ordinaria sobre los datos clr-transformados de forma que las operaciones de perturbación equivalen a sumas y las distancias corresponden a la geometría de Aitchison en el símplex. Con ello la operativa de análisis se reduce a la ordinaria.

Como se ha mencionado, la geometría de Aitchison en el símplex es una geometría euclídea. Esto permite utilizar todas las propiedades y herramientas de estos espacios que nos son familiares. Es decir, disponemos de productos escalares para hacer proyecciones ortogonales y de ejes ortogonales que permiten trabajar en coordenadas ([3] cap. 3). Simplificando, los pasos de un análisis composicional son: expresar los datos composicionales mediante sus coordenadas; realizar el análisis estadístico de las coordenadas; e interpretar los resultados, deshaciendo la transformación a coordenadas si es necesario.

Una alternativa composicional al análisis de correlación consiste en examinar el comportamiento de log-cocientes entre dos partes. Si la varianza del log-cociente es pequeña indica que ambas partes están muy asociadas y viceversa. El conjunto de todas estas varianzas puede sustituir a las matrices de covarianza-correlación ordinarias. Por otra parte los valores medios de estos log-cocientes, corresponden al concepto de media en el símplex: se define *centro* de una muestra composicional como la media geométrica de cada parte a lo largo de la muestra, normalizada a la suma constante. Estos conceptos permiten hacer los resúmenes estadísticos necesarios en cualquier estudio ([3] cap. 2).

La estadística composicional es siempre multivariante, frecuentemente con una dimensión considerable. La representación intuitiva de una muestra es, en estas circunstancias, un desafío. Una de las herramientas más útiles es el llamado *biplot* que permite la representación simultánea de datos y variables que se basa en el análisis de componentes principales. En el caso composicional se procede a tomar clr de los datos y a continuación realizar el biplot ([3] cap. 5, 8). La figura 1 representa el *biplot*-composicional (86% varianza explicada) de los resultados de las elecciones al *Parlament de Catalunya* en 2010 ([3] cap. 2). Se puede interpretar que la primera componente principal corresponde a el no apoyo al nacionalismo catalán; la segunda parece representar el conservadurismo del *status quo*. Las uniones de los rayos representan la varianza de los log-cocientes simples y el ángulo entre ellas su correlación. Por ejemplo se sugiere que el log-cociente (CiU, C’s) está poco correlacionado con el de (PP, ICV); también que las relación (PSC, PP) está muy correlacionada con la de (ERC, CiU, C’s), etc.

4. MODELOS DE EVOLUCIÓN

Como se ha mencionado los métodos estadísticos tradicionales pueden aplicarse a los vectores de coordenadas. Por ejemplo, una coordenada composicional puede tomarse como variable respuesta y tratar de predecirla a partir de una variable explicativa externa a la composición con métodos de regresión.

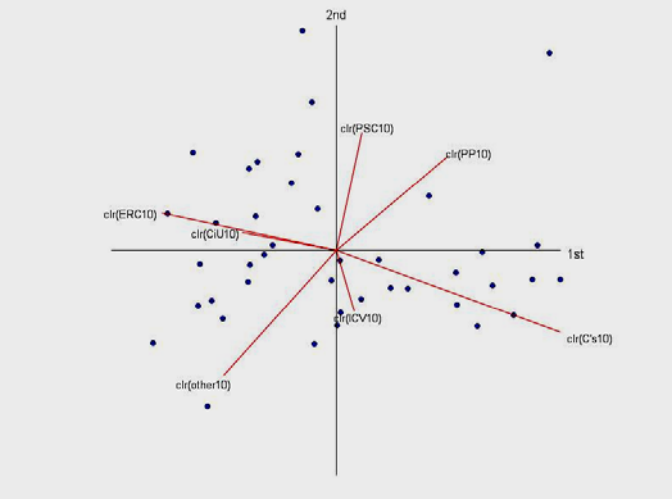

Figura 1. Biplot compositional de los votos a candidaturas en las elecciones al Parlament de Catalunya 2010. Los rayos son proporcionales a la variabilidad de las componentes clr de las candidaturas. Los puntos representan *vegueries.*

El símplex, como espacio euclídeo, admite la definición de derivada e integral y por tanto ecuaciones diferenciales que pueden modelar la evolución en el tiempo ([3] cap. 12). Por ejemplo, la ecuación que iguala la derivada con una constante será una recta en el símplex. La figura 2 muestra un ajuste (por regresión) del petróleo ya consumido, las reservas conocidas y tres hipótesis de reservas desconocidas. El ajuste es a una recta en el símplex, después de descartar ecuaciones diferenciales de orden superior.

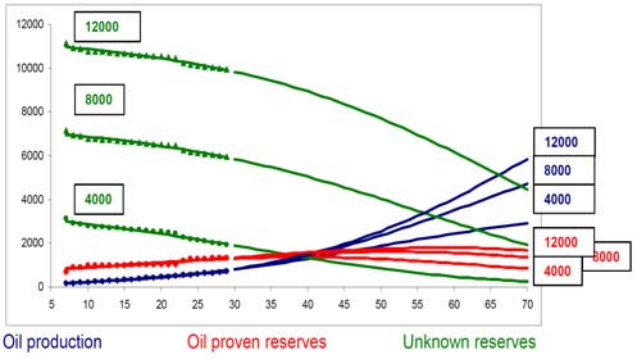

Figura 2. Regresión de la composición (petróleo consumido, reservas conocidas, reservas desconocidas) en Mbarriles. Predicción a 50 años siguiendo una recta del símplex.

REFRENCIAS

[1] Aitchison, J., 1986, *The statistical analysis of compositional data. Monographs on statistics and applied Probability*, Chapman & Hall, London (Reprinted in 2003 with additional material by Press Blackburn).
[2] Pawlowsky-Glahn, V. and J.J. Egozcue: Geometric Approach to Statistical Analysis on the Simplex, *Stochastic Environmental Research and Risk Assessment*, 15, 5, 384-398, 2001.
[3] Pawlowsky-Glahn, V. and A. Buccianti (Ed.) (2011): *Compositional data analysis. Theory and applications*, Wiley, West Sussex, UK.