ANÁLISIS DE DATOS COMPOSICIONALES PARA EL ESTUDIO DE MATERIALES ARQUEOLÓGICOS

Denisse L. Argote Espino[1] y Pedro A. López García[2]

(1) Escuela Nacional de Antropología e Historia. México, D.F. México. E-mail: dplopez@prodigy.net.mx

(2) Dirección de Estudios Arqueológicos, Instituto Nacional de Antropología e Historia. México, D.F. México. E-mail: efenfi@gmail.com

RESUMEN

En este trabajo se expone algunos problemas relacionados con el manejo de datos composicionales, poniendo un especial énfasis en lo que respecta al análisis y a la clasificación de los datos. Se mencionan algunos aspectos teóricos relacionados con la estadística multivariable clásica, la cual es utilizada comúnmente para analizar este tipo de datos, y haciendo una breve referencia sobre los supuestos teóricos subyacentes de estos modelos. Se presentan las diferentes transformaciones utilizadas para el tratamiento de los datos composicionales y se considera el problema de la ocurrencia de ceros y de su tipología. Además se trata el problema relacionado con la imputación de valores faltantes y de la detección de valores extremos (outliers), utilizando la representación gráfica utilizada para la identificación de grupos químicos en el análisis de los datos. Para la aplicación de la teoría, se utiliza el análisis químico de 73 muestras de cerámica obtenido por medio de análisis de espectroscopía de EDX (Energy Dispersive X-Ray Spectroscopy) con la finalidad de observar la posibilidad de determinar tipos cerámicos con patrones composicionales similares. Se

comenta el uso del paquetes estadísticos utilizados en el manejo de datos composicionales, entre estos se encuentran el programa CoDaPack, el programa R y varias librerías implementadas en el programa Compositions.

Palabras clave: Datos composicionales, transformación de datos, ocurrencia de ceros, imputación de valores, puntos aberrantes.

INTRODUCCIÓN

Durante los últimos años, se ha ido consolidando el interés en la utilización de métodos analíticos para el análisis de materiales arqueológicos, sobre todo en lo que respecta a la cerámica, en donde el principal objetivo es la identificación de grupos de artefactos con composición química similar. Con estos métodos es posible recuperar el contenido químico de las muestras de componentes mayores o menores, dando lugar a una nueva interpretación de los datos. Cabe mencionar que, aunque este trabajo se enfoca en el análisis composicional de cerámica, los temas tratados se han aplicado tanto al análisis de vidrios arqueológicos (Baxter *et al.*, 2005; Beardah y Baxter, 2005; Jackson y Baxter, 1999) como de cerámica (Buxeda i Garrigós, 1999; Glascock, 1992). El propósito de este trabajo será, entonces, el de abordar las características comunes del análisis de datos composicionales con lo que respecta al manejo de los datos, explicando el por qué los datos composicionales no pueden ser directamente analizados utilizando técnicas estadísticas estándares con los datos crudos.

En la actualidad un gran número de investigaciones han optado por incursionar en el análisis químico de muestras cerámicas con el fin de responder preguntas sobre la composición de las vasijas, el tipo de inclusiones utilizadas en la manufactura de las piezas y realizar estudios de

procedencia de los materiales. Para ello, las muestras se han procesado principalmente con las técnicas instrumentales de Espectrometría de Masas con fuente de Plasma de Acoplamiento Inductivo (ICP-MS), Emisión de Rayos X Inducida por Partículas (PIXE), Activación Neutrónica (NAA), Fluorescencia de Rayos X (XRF) y Espectroscopía por Energía Dispersiva (EDS). Al aplicar estos análisis sobre cualquier tipo de material, se obtiene una matriz de datos de dimensiones $n \times p$ cuyo contenido describe de manera cuantitativa la composición química de n artefactos, en donde n son los renglones que corresponden a cada una de las muestras analizadas y p se refiere a los elementos químicos detectados. Estos valores pueden ser expresados como porcentajes, proporciones o ppm de cada elemento.

Una vez obtenidas estas matrices, muchos investigadores acostumbran aplicar técnicas estadísticas multivariadas tradicionales (análisis de cúmulos, componentes principales, análisis de factores, etcétera), con el fin de poder determinar los grupos químicos de las muestras y con ello realizar inferencias sobre diferentes aspectos relacionados con vida diaria de las poblaciones estudiadas. Sin embargo, siguiendo estos procedimientos comunes de análisis, se puede estar muy lejos de resolver correctamente un problema de clasificación por una gran serie de inconvenientes existentes, sobre todo el referente a la restricción de la suma constante, el cual está presente cuando se trabaja con datos composicionales brutos.

EL ESPACIO DE MUESTRA

El *Simplex* unitario

La teoría del análisis composicional de datos tiene varias décadas de haber sido establecida y con el paso de los años sigue incrementándose el número de investigadores cuya aportación la ha solidificado como una teoría formal. En un estudio de datos composicionales se establece que los datos composicionales consisten de vectores cuyos componentes son las proporciones o porcentajes de un todo. Es importante señalar que los datos composicionales son multivariables en esencia y, al no tener en cuenta el carácter composicional de los datos, se pueden provocar sesgos importantes en la clasificación (Egozcue y Pawloski-Glahn, 2011; Van den Boogaart y Tolosana-Delgado, 2013). El uso de log-cocientes hace posible trabajar con datos en el espacio de los números reales para poder utilizar las técnicas estadísticas multivariables estándar. Fue Aitchinson (1986) quien introduce la metodología desarrollada de log-cocientes y expone las ventajas de este procedimiento y discute las cuestiones relacionadas con el modelado estadístico y su representación gráfica para buscar grupos químicos composicionales.

En primer lugar se establece que, para el caso de datos composicionales, el espacio de muestra es el *Simplex*; este se trata de un espacio cerrado, el cual es muy diferente al espacio Euclidiano real asociado con datos sin restricciones. Una composición se define como un vector D con componentes positivos $\mathbf{x} = [x_1,\ldots, x_D]$, cuya sumatoria es igual a una constante k definido como:

$$S^D = \left\{ x = [x_1, x_2, ..., x_D] \middle| x_i \rangle 0, i = 1,2,...,D; \sum_{i=1}^{D} x_i = k \right\} \qquad (1)$$

Debido a esta definición, los datos composicionales están sujetos a la restricción de que su suma debe ser igual 1 para proporciones, 100 para porcentajes o 10^6 para datos en ppm (partes por millón), y sólo pueden variar de > 0 a k, a diferencia de los datos que son libres de variar en los intervalos de $-\infty$ a $+\infty$ en un espacio real (Pawlosky-Glahn y Egozcue, 2006). Las restricciones de

no negatividad y de suma constante que caracterizan este tipo de datos implican que las técnicas multivariables habitualmente utilizadas no son adecuadas para su análisis y modelización, ya que se basan en matrices de covarianza y de correlación de vectores de observaciones de datos.

En Estadística, la matriz de varianza y covarianza es definida positiva y el determinante de ésta debe ser siempre no negativo, supuesto que no se cumple cuando los datos son composicionales. Por otro lado, los métodos estadísticos tradicionales están diseñados para operar con datos completos o sin restricciones, además de que se asume la existencia del supuesto de que los datos a analizar poseen una distribución multivariable normal, la cual es una generalización de la distribución normal unidimensional a dimensiones superiores y que se define por la media del vector y la matriz de varianzas y covarianzas, o en su caso en matrices de correlación. Algunas aplicaciones acostumbran transformar los datos a logaritmo base 10 con la creencia de que este procedimiento garantizará llevar los datos a una base log-normal, para luego aplicar una técnica multivariada como el análisis de componentes principales (Glascock *et al*., 2004; Makowski *et al.*, 2008). Sin embargo esto no garantiza la normalidad de los datos.

Los datos que se obtienen en el análisis cerámico tienen carácter composicional, es decir, describen cuantitativamente las partes que forman un todo. La constante *k* representa la suma de las partes; *k* puede ser 1 o 100. En estos casos, sólo los cocientes entre las partes son de interés. Siguiendo este razonamiento, las partes de una composición pueden ser interpretadas como probabilidades o porcentajes (Hron *et al*., 2010). Como una regla general, cuando se trata con datos composicionales, primero se deben expresar los datos en términos de componentes de log-cocientes y después aplicar la metodología multivariada apropiada para vectores de datos sin restricciones, los

cuales ahora se encuentran en un espacio real y libres de la restricción de la suma constante (Reyment, 2006).

Completamente composicional

Aunque los datos composicionales son presentados como porcentajes, cada vector de datos no suma el 100%. Esto significa que existe un componente adicional implícitamente definido, al cual suele llamarse la parte residual y que completa el 100% (Pawlosky-Glahn y Egozcue, 2006). Si se utiliza la transformación de log-cocientes, lo conveniente es respetar la restricción de la suma constante y añadir una variable extra o se pueden cerrar los datos. Realizando cualquiera de estas dos operaciones se puede hablar de datos completamente composicionales. El primer caso se logra al especificar una variable residual por diferenciación de 1 o 100 menos el total del vector, de tal manera que la suma del cada vector sea del 1 o el 100%. Para el segundo caso, si los datos son estandarizados, se pueden obtener datos completamente composicionales por medio del operador de cerradura (Egozcue *et al.*, 2011; Baxter *et al.*, 2005).

El operador de cerradura se define por la siguiente expresión:

$$Cx = \left(\frac{kx_1}{\sum_{i=1}^{D} xi}, \frac{kx_2}{\sum_{i=1}^{D} xi}, \ldots, \frac{kx_D}{\sum_{i=1}^{D} xi} \right) \qquad (2)$$

Las componentes del vector cerradura se denominan *partes*, referidas al total *k* (Egozcue *et al.*, 2011). Esta operación es requerida para una definición formal de una subcomposicion y para la transformación inversa.

Transformaciones

Básicamente existen tres transformaciones para datos composicionales: el log-cociente aditivo o *alr*, el log-cociente centrado o *clr* y el log-cociente isométrico o *irl* (Aitchinson, 1986; Egozcue *et al.*, 2003). Aitchinson (1981) demostró que los efectos de la restricción de la suma constante en las matrices de covarianzas y de correlación desaparecen si los datos originales son expresados como cocientes de logaritmos. Estas trasformaciones retienen la estructura apropiada de cualquier conjunto de datos composicionales (Kucera y Malmgren, 1998) y una composición puede ser representada como un vector real.

La transformación log-cociente aditivo se define como (Aitchinson 1986):

$$alr(\mathrm{x}) = \ln\left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, \ldots, \frac{x_{D-1}}{x_D}\right) \qquad (3)$$

Si se utiliza la transformación *alr* sus coordenadas son *D-1*. Hay que tener en cuenta que el *alr* posee un problema de falta de simetría debido al uso de una de las partes como denominador común, violando con esto el principio de invariancia por permutación de las partes, además de que las coordenadas obtenidas por medio de esta transformación no pueden mapearse en ejes ortogonales (Egozcue y Pawlowsky-Glann, 2006).

Con la finalidad de superar la deficiencia hallada en la transformación *alr*, Aitchinson (1986) define la transformación log-cociente centrada (*clr*), la cual está dada por:

$$\mathrm{v} = clr(\mathrm{x}) = \ln\frac{x_1}{g_m(x)}, \frac{x_2}{g_m(x)}, \ldots, \frac{x_D}{g_m(x)} \qquad (4)$$

con $$g_m(x) = \left(\prod_{j=1}^{n} \right) x_{ij}^{\ 1/n}, i = 1,2,...,D. \quad (5)$$

en donde $g_m(x)$ representa a la media geométrica. Esta expresión es simétrica en las partes, por lo que es utilizada para generar los biplots composicionales. Sin embargo, cabe aclarar que la suma de las componentes es de cero, por lo que se recomienda cautela al trabajar con estas coordenadas ya que las matrices de covarianzas y de correlación son singulares, es decir, su determinante es cero (Pawlowski-Glahn y Egozcue, 2006).

La transformación isométrica (*ilr*) traslada la geometría del simplex a un espacio multivariable real, caracterizado por la transformación relativa de ángulos y distancias en el simplex a ángulos y distancias en el espacio real (Egozcue *et al*., 2003). Esta transformación no sufre del problema de singularidad como es el caso de la transformación *clr*; el inconveniente es que la interpretación de los resultados es mucho más difícil, por lo que para interpretar los datos aplicando alguna técnica multivariada se debe volver a transformar los datos a sus valores originales o aplicar balances por medio de una partición secuencial binaria. La transformación isométrica está dada por:

$$ilr_V^{-1}(x) = C[\exp(x \bullet V^t)] \quad (6)$$

donde C es el operador del cierre para la constante *k*. Las variables *ilr* son coeficientes de una base ortonormal en el hiperplano formadas por la transformación *clr*.

Matriz de variación

A diferencia de la exploración de datos sin restricciones, en el análisis de datos composicionales la exploración se hace por medio de la matriz de variación, calculando el centro de los datos y por

medio de la varianza total. Con la matriz de variación es posible darse cuenta de la dispersión de los datos; esta matriz se define como T = [τ_{ij}] y en forma extendida es igual a:

$$T = [\tau_{ij}] = \begin{pmatrix} \mathrm{var}\left[\ln\frac{X_1}{X_1}\right] & \mathrm{var}\left[\ln\frac{X_1}{X_2}\right] & \cdots & \mathrm{var}\left[\ln\frac{X_1}{X_D}\right] \\ \mathrm{var}\left[\ln\frac{X_2}{X_1}\right] & \mathrm{var}\left[\ln\frac{X_2}{X_2}\right] & \cdots & \mathrm{var}\left[\ln\frac{X_2}{X_D}\right] \\ \vdots & \vdots & \cdots & \vdots \\ \mathrm{var}\left[\ln\frac{X_D}{X_1}\right] & \mathrm{var}\left[\ln\frac{X_D}{X_2}\right] & \cdots & \mathrm{var}\left[\ln\frac{X_D}{X_D}\right] \end{pmatrix} \qquad (7)$$

en donde τ_{ij} = var[ln(X_i/X_j)] representa la varianza común de los log-cocientes de las partes *i* y *j*; esta matriz se basa en el aporte de varianza por cada uno de los log-cocientes.

Varianza total

Una medida de la variabilidad total de la matriz X está dada por

$$\mathrm{var}\,tot[X] = \frac{1}{2D}\sum_{l=1}^{D}\sum_{j=1}^{D}\mathrm{var}\left[\ln\frac{X_i}{X_j}\right], \qquad (8)$$

en donde 'var*tot*' representa la varianza total del conjunto de datos, es decir, la suma de todos los elementos de la matriz de variación *T* dividida entre 2 (Daunis-i-Estadella *et al*., 2006). Si para el caso normal se calcula la matriz de varianzas y covarianzas, se obtienen aportes negativos de la variabilidad. De hecho, el sesgo negativo en la estructura de las covarianzas puede ser explicado como una consecuencia del fundamento euclidiano de las estadísticas clásicas, en donde la escala es

absoluta y no relativa (Pawlosky-Glahn y Egozcue, 2006). Debido a este hecho, en muchas aplicaciones los resultados obtenidos no tienen nada que ver con la realidad y por lo tanto serían inválidos.

MATERIALES Y MÉTODO

La tipología cerámica para el Estado de Tlaxcala, México, fue establecida por los arqueólogos Ángel García Cook y Leonor Merino (García Cook, 1996), quienes además lograron describir las diferentes secuencias de ocupación para toda la región. La tipología se estableció principalmente con base en el acabado de superficie y el tipo de cocción de la cerámica, conjuntamente con una descripción de la forma.

Para este trabajo se analizaron un total de 73 muestras de cerámica de diferentes tipos provenientes del sitio arqueológico de los Teteles de Ocotitla (Tabla 1). Estas muestras están fechadas para la fase Tenanyecac (200 a.C. a 200 d.C.). El conjunto de datos consiste de la composición química de 10 variables. En la tabla 2 se presentan los porcentajes de cada componente asociado con cada una de las variables, incluyendo la presencia de valores faltantes; cabe señalar que en esta tabla los valores que aparecen en negritas originalmente correspondían a ceros y posteriormente fueron imputados, hecho que se explica más adelante. En la segunda parte de la tabla 2 (datos completamente composicionales) se cerraron los datos utilizando la ecuación 2, de tal forma que los datos cumplen con la restricción de la suma constante, es decir la sumatoria es igual a 1 o al 100%. Para este trabajo se consideraron los elementos mayores con el objetivo de estudiar su fórmula química, obtenidos a través de la técnica de EDX (Energy dispersive X-ray spectroscopy). Entre estos elementos mayores se trata de poder discriminar los elementos de origen

natural y locales de aquellos que fueron añadidos deliberadamente. De igual manera se trató de ver si era posible detectar muestras no locales. Del análisis se pudieron obtener las concentraciones O, Na, Mg, Al, Si, P, K, Ca, Ti y Fe en porcentajes.

Tabla 1. Número de muestras de diferentes tipos cerámicos utilizados en el análisis de texturas

Café	2
Bicromos	5
Anaranjado delgado	4
Blanco	5
Acabado burdo	4
rojos	5
Rojo interior	4
Rojo exterior	7
Negro_b	5
Negros	3
Café pulido	7
Café_g	5
Café_bpg	5
Café_alisado_dot	5
Café alisado	8

Tabla 2. Concentraciones de las partes en porcentajes e imputación de valores con el algoritmo IRMI (método robusto *model-based*); los valores imputados están en negritas

ID (TIPO)	O	Na	Mg	Al	Si	P	K	Ca	Ti	Fe
A_burdo	52.62	1.74	0.42	10.43	25.72	0.56	1.12	2.58	0.47	4.35
A_burdo	53.87	1.42	0.68	9.83	23.65	0.85	1.16	2.91	0.93	4.7
A_burdo	55.88	1.21	0.51	9.72	23.76	0.35	1.2	2.26	0.68	4.43
A_burdo	51.9	1.8	0.91	10.24	25.76	0.39	1.23	3.15	0.81	3.81
A_naranjad	53.93	1.84	0.86	9.74	25.46	0.18	1.97	1.93	0.34	3.73
A_naranjad	48.28	1.29	0.54	8	22.27	0.42	1.03	2.35	0.22	3.46
A_naranjad	52.31	1.16	0.55	10.23	27.79	0.43	1.62	2.15	0.35	3.42
A_naranjad	52.51	1.19	0.33	9.38	25.84	0.68	2.11	2.58	1.21	4.19
Bicromo	52.41	1.28	0.43	9.58	27.98	0.27	1.18	2.76	0.47	3.63
Bicromo	51.19	1.84	0.57	9.54	28.32	0.39	1.47	2.4	0.8	3.49
Bicromo	53.8	1.44	0.65	9.41	25.28	0.95	0.87	2.59	1.09	3.92
Bicromo	53.97	1.04	0.6	9.51	25.81	0.47	1.23	2.5	0.76	3.92
Bicromo	54.79	1.23	0.79	10.13	24.2	0.65	1.08	2.91	0.52	3.7
Blanco	53.5	1.46	0.63	9.97	24.98	0.35	0.98	2.72	0.72	4.38

Blanco	55.5	2.01	0.8	10.29	23.61	0.57	1.13	2.49	0.52	3.08
Blanco	54.77	1.56	0.81	8.97	26.77	0.3	1.27	1.91	0.32	3.31
Blanco	53.77	1.63	0.87	9.75	25	0.55	1.29	2.29	0.57	4.28
Blanco	51.37	1.49	0.46	9	23.43	0.28	1.04	2.35	0.58	3.52
Café	53.05	1.16	0.61	9.23	21.1	0.61	1.38	2.17	0.39	3.34
Café	49.51	1.67	0.57	9.04	22.43	0.37	1.57	2.07	0.57	4.38
Café_alisado	47.75	**0.95**	0.39	7.98	18.71	**0.43**	0.96	1.58	0.45	4.46
Café_alisado	49.43	1.39	0.54	8.59	20.92	0.71	1.39	2.45	0.34	3.5
Café_alisado	50.81	1.33	0.44	8.42	24.19	0.24	1.02	2.19	0.45	3.56
Café_alisado	52.09	1.55	0.7	9.55	26.51	0.58	1.66	2.92	0.4	4.06
Café_alisado	51.84	1.88	0.49	10.28	25.74	0.59	1.59	2.63	0.78	4.19
Café_alisado	52.56	1.56	0.38	9.54	25.64	0.45	1.42	2.55	0.86	5.04
Café_alisado	52.43	1.49	0.61	10.31	25.36	0.6	1.42	2.92	0.63	4.23
Café_alisado	52.64	1.85	0.56	9.94	26.03	0.47	0.94	2.7	1.14	4.2
C_alisado_dot	55.19	1.36	0.5	9.48	25.95	0.24	0.97	2.32	0.49	3.5
C_alisado_dot	54.27	2.15	0.81	9.75	25.4	0.32	1	2.4	0.54	3.36
C_alisado_dot	54.15	1.54	0.75	9.11	25.26	0.37	1.69	2.22	0.44	4.47
C_alisado_dot	50.04	1.61	0.63	9.3	28.17	0.42	1.98	2.12	**0.60**	4.23
C_alisado_dot	54.15	1.33	0.76	9.72	25.98	0.53	1.43	2.6	0.38	3.13
Café_bpg	54.97	1.4	0.57	9.87	25.37	**0.50**	1.18	1.7	0.55	4.4
Café_bpg	46.19	2	0.59	10.38	30.08	**0.25**	1.34	3.14	0.68	5.6
Café_bpg	48.3	1.42	0.69	10.94	26.75	0.37	1.28	3.14	**0.48**	5.5
Café_bpg	54.3	1.21	0.46	10.19	23.85	0.79	1.27	2.31	0.84	4.78
Café_bpg	54.07	1.41	0.42	9.43	25.98	0.38	0.99	2.92	0.67	3.29
Café_g	53.15	1.4	0.44	9.39	27.68	0.37	1.03	2.13	0.29	4.12
Café_g	55.4	2.25	0.54	9.33	24.2	**0.74**	1	2.45	1.35	3.49
Café_g	54.07	1.9	1	9.41	26.59	**0.51**	1.51	1.69	0.44	3.39
Café_g	51.9	1.28	0.63	11.48	24.67	**0.22**	1.04	2.03	0.38	6.59
Café_g	52.43	1.54	0.64	9.25	26.68	0.25	1.28	2.49	0.35	4.55
Café_pulido	53.18	1.67	0.68	9.46	26.52	0.91	1.39	2.7	0.4	3.1
Café_pulido	52.95	1.46	1.04	9.41	27.29	0.34	0.81	2.38	0.33	3.98
Café_pulido	49.74	1.52	0.59	9.7	26.35	0.54	0.91	3.5	1.72	5.43
Café_pulido	54.89	1.15	**0.50**	8.31	29.08	0.36	2.03	1.55	0.4	2.23
Café_pulido	53.53	1.79	0.72	9.99	25.65	0.42	1.33	2.68	0.45	3.44
Café_pulido	53.39	1.63	0.51	10.1	25.6	0.85	1.31	2.74	0.49	3.36
Café_pulido	54.45	1.51	0.51	8.97	25.69	0.91	1.15	2.25	0.48	4.08
Negros	55.33	1.52	0.35	9.97	25.06	0.66	0.75	1.83	0.46	4.08
Negros	54.52	1.22	0.64	10.28	25.65	0.44	0.82	2.32	0.44	3.66
Negros	51.4	1.27	0.38	10.03	28.16	0.29	0.89	2.92	0.78	3.88
Negros_b	55.97	0.84	0.74	9.17	26.31	0.5	1.4	1.62	0.48	2.97
Negros_b	55.32	1.32	0.58	9.2	25.59	0.49	1.67	2.2	0.7	2.81

Negros_b	50.93	1.31	0.31	10.04	26.06	0.24	1.87	3.15	0.92	5.17
Negros_b	55.42	1.16	0.46	10.23	24.19	0.42	1.05	2	1.03	4.04
Rojo_ext	52.62	0.99	0.5	9.41	26.91	**0.46**	0.77	2.81	0.5	5.5
Rojo_ext	54.09	1.52	0.54	10.09	25.06	0.85	0.85	2.43	0.42	4.15
Rojo_ext	53.51	1.2	0.62	9.7	25.47	0.73	1.19	2.56	0.89	4.12
Rojo_ext	52.57	1.96	0.61	9.61	26.22	0.5	0.64	2.16	0.69	4.49
Rojo_ext	52.26	1.84	0.48	10.45	26.83	0.36	0.75	2.2	0.28	4.23
Rojo_ext	53.5	1.53	0.83	9.68	25.54	0.44	1.31	2.36	0.77	4.02
Rojos	51.54	1.52	0.61	10.04	26.67	0.27	1.98	2.7	0.69	3.96
Rojos	51.1	1.52	0.74	10.08	26.18	0.52	2.25	2.55	0.47	4.59
Rojos	51.28	1.79	0.76	9.26	26.85	0.53	1.36	2.49	0.71	4.96
Rojos	53.2	1.83	0.87	9.74	25.44	0.57	1.82	2.35	0.41	3.76
Rojos	51.76	0.41	0.78	7.08	22.95	0.40	2.04	2.95	0.59	4.15
Rojo_ext	49.14	**1.31**	0.74	7.9	24.92	**0.06**	1.72	0.95	0.36	5.07
Rojo_int	53.21	0.9	1.1	8.43	26.46	**0.47**	1.13	2.39	0.89	5.5
Rojo_int	47.43	1.09	0.62	8.75	21.47	**0.47**	0.6	2.03	0.46	3.16
Rojo_int	50.58	1.45	0.3	8.58	22.85	**0.42**	0.86	2.4	0.75	3.98
Rojo_int	49.26	1.41	0.52	8.42	24.37	**0.30**	1.06	2.29	0.64	3.99

Datos completamente composicionales

O	Na	Mg	Al	Si	P	K	Ca	Ti	Fe
52.6147385	1.73982602	0.419958	10.4289571	25.7174283	0.55994401	1.11988801	2.57974203	0.469953	4.34956504
53.87	1.42	0.68	9.83	23.65	0.85	1.16	2.91	0.93	4.7
55.88	1.21	0.51	9.72	23.76	0.35	1.2	2.26	0.68	4.43
51.9	1.8	0.91	10.24	25.76	0.39	1.23	3.15	0.81	3.81
53.9407882	1.84036807	0.86017203	9.74194839	25.465093	0.18003601	1.97039408	1.93038608	0.34006801	3.73074615
54.9510585	1.46824494	0.61461416	9.10539495	25.3471432	0.47803323	1.1723196	2.67470977	0.25039836	3.93808331
52.3047695	1.15988401	0.54994501	10.2289771	27.7872213	0.429957	1.61983802	2.14978502	0.349965	3.41965803
52.4995001	1.18976205	0.32993401	9.37812438	25.834833	0.67986403	2.10957808	2.5794841	1.20975805	4.18916217
52.4152415	1.28012801	0.430043	9.5809581	27.9827983	0.270027	1.18011801	2.76027603	0.470047	3.63036304
51.1848815	1.83981602	0.56994301	9.5390461	28.3171683	0.389961	1.46985301	2.39976002	0.79992001	3.48965103
53.8	1.44	0.65	9.41	25.28	0.95	0.87	2.59	1.09	3.92
54.0727382	1.04197976	0.60114217	9.5281034	25.8591324	0.4708947	1.23234145	2.50475904	0.76144675	3.92746218
54.79	1.23	0.79	10.13	24.2	0.65	1.08	2.91	0.52	3.7
53.6663657	1.46454007	0.63195907	10.0010031	25.0576788	0.35108837	0.98304745	2.72845822	0.72223894	4.39362022
55.5	2.01	0.8	10.29	23.61	0.57	1.13	2.49	0.52	3.08
54.7754775	1.56015602	0.81008101	8.97089709	26.7726773	0.30003	1.27012701	1.91019102	0.320032	3.31033103
53.77	1.63	0.87	9.75	25	0.55	1.29	2.29	0.57	4.28
54.9294269	1.59324209	0.4918734	9.62360992	25.0534645	0.2994012	1.11206159	2.51283148	0.6201882	3.76390077
57.0184867	1.24677558	0.65563199	9.92046432	22.6784179	0.65563199	1.48323302	2.33233018	0.41917455	3.58985383
53.7101323	1.81167281	0.61835539	9.80689954	24.3328271	0.40138859	1.70318941	2.24560642	0.61835539	4.75157301
57.0762611	1.13554865	0.4661726	9.53860865	22.3643318	0.51398518	1.14750179	1.8885967	0.53789147	5.33110208
55.3775487	1.55724849	0.60497423	9.62357159	23.4371499	0.79542908	1.55724849	2.7447905	0.3809097	3.92112929

54.8407987	1.43550998	0.47490556	9.08796546	26.1090124	0.2590394	1.10091743	2.36373448	0.48569887	3.8424177
52.0795841	1.54969006	0.69986003	9.54809038	26.5046991	0.57988402	1.65966807	2.91941612	0.39992002	4.05918816
51.8348165	1.87981202	0.489951	10.2789721	25.7374263	0.58994101	1.58984102	2.62973703	0.77992201	4.18958104
52.56	1.56	0.38	9.54	25.64	0.45	1.42	2.55	0.86	5.04
52.43	1.49	0.61	10.31	25.36	0.6	1.42	2.92	0.63	4.23
52.3937494	1.84134568	0.55738031	9.89350055	25.9082313	0.46780133	0.93560267	2.68736936	1.13466706	4.18035234
55.19	1.36	0.5	9.48	25.95	0.24	0.97	2.32	0.49	3.5
54.27	2.15	0.81	9.75	25.4	0.32	1	2.4	0.54	3.36
54.15	1.54	0.75	9.11	25.26	0.37	1.69	2.22	0.44	4.47
50.4944501	1.62462159	0.63572149	9.38446014	28.4258325	0.42381433	1.99798184	2.13925328	0.60544904	4.26841574
54.1445855	1.32986701	0.75992401	9.7190281	25.9774023	0.52994701	1.42985701	2.59974003	0.379962	3.12968703
54.6910755	1.39289623	0.56710775	9.81991842	25.2412695	0.49746294	1.17401254	1.69137399	0.54720923	4.37767386
46.074813	1.99501247	0.58852868	10.3541147	30.0049875	0.24937656	1.33665835	3.13216958	0.67830424	5.58603491
48.8520279	1.43622939	0.69788611	11.0650349	27.0557297	0.37422879	1.29462931	3.17588753	0.48548599	5.56286032
54.3	1.21	0.46	10.19	23.85	0.79	1.27	2.31	0.84	4.78
54.3089594	1.41623142	0.42185617	9.47167537	26.0948172	0.38167939	0.99437525	2.93290478	0.67296103	3.30453998
53.15	1.4	0.44	9.39	27.68	0.37	1.03	2.13	0.29	4.12
54.9875931	2.23325062	0.53598015	9.26054591	24.0198511	0.73449132	0.99255583	2.43176179	1.33995037	3.46401985
53.7956422	1.89035917	0.99492588	9.36225251	26.4550791	0.5074122	1.50233808	1.68142473	0.43776739	3.37279873
51.7860706	1.27719018	0.62861704	11.4547994	24.6158451	0.21951706	1.03771702	2.0255438	0.37916584	6.57553383
52.7146592	1.54836115	0.64347476	9.30022119	26.8248542	0.25135733	1.28694953	2.503519	0.35190026	4.5747034
53.1746825	1.66983302	0.67993201	9.45905409	26.5173483	0.90990901	1.38986101	2.69973003	0.39996	3.09969003
52.9552955	1.46014601	1.04010401	9.41094109	27.2927293	0.340034	0.81008101	2.38023802	0.330033	3.98039804
49.74	1.52	0.59	9.7	26.35	0.54	0.91	3.5	1.72	5.43
54.6169154	1.14427861	0.49751244	8.26865672	28.9353234	0.35820896	2.0199005	1.54228856	0.39800995	2.21890547
53.53	1.79	0.72	9.99	25.65	0.42	1.33	2.68	0.45	3.44
53.4006801	1.63032607	0.51010202	10.1020204	25.605121	0.85017003	1.31026205	2.74054811	0.49009802	3.36067213
54.45	1.51	0.51	8.97	25.69	0.91	1.15	2.25	0.48	4.08
55.3244676	1.51984802	0.349965	9.9690031	25.0574943	0.65993401	0.74992501	1.82981702	0.459954	4.07959204
54.5254525	1.22012201	0.64006401	10.2810281	25.6525653	0.440044	0.82008201	2.32023202	0.440044	3.66036604
51.4	1.27	0.38	10.03	28.16	0.29	0.89	2.92	0.78	3.88
55.97	0.84	0.74	9.17	26.31	0.5	1.4	1.62	0.48	2.97
55.3864638	1.3215859	0.58069684	9.21105326	25.6207449	0.49058871	1.67200641	2.20264317	0.70084101	2.81337605
50.93	1.31	0.31	10.04	26.06	0.24	1.87	3.15	0.92	5.17
55.42	1.16	0.46	10.23	24.19	0.42	1.05	2	1.03	4.04
52.3738429	0.98536877	0.49766099	9.36597989	26.7841147	0.45784811	0.76639793	2.79685478	0.49766099	5.47427093
54.09	1.52	0.54	10.09	25.06	0.85	0.85	2.43	0.42	4.15
53.5153515	1.20012001	0.62006201	9.7009701	25.4725473	0.73007301	1.19011901	2.56025603	0.89008901	4.12041204
52.860734	1.97083962	0.61337355	9.66314731	26.3650075	0.50276521	0.64353947	2.1719457	0.69381599	4.51483157
52.4277689	1.8459069	0.48154093	10.4835474	26.9161316	0.3611557	0.7524077	2.2070626	0.28089888	4.24357945
53.5107021	1.53030606	0.83016603	9.68193639	25.545109	0.44008802	1.31026205	2.36047209	0.77015403	4.02080416
51.5503101	1.52030406	0.61012202	10.0420084	26.6753351	0.27005401	1.98039608	2.70054011	0.69013803	3.96079216
51.1	1.52	0.74	10.08	26.18	0.52	2.25	2.55	0.47	4.59
51.2851285	1.79017902	0.76007601	9.26092609	26.8526853	0.53005301	1.36013601	2.49024902	0.71007101	4.96049605
53.2053205	1.83018302	0.87008701	9.7409741	25.4425443	0.57005701	1.82018202	2.35023502	0.410041	3.76037604
55.5901622	0.44033938	0.83771883	7.60390935	24.6482655	0.4295994	2.19095693	3.16829556	0.63365911	4.45709376

53.3145275	1.42128675	0.80286427	8.57111859	27.0369969	0.0650971	1.86611696	1.03070413	0.39058262	5.50070522
52.9558121	0.89570064	1.09474522	8.3897293	26.3335987	0.46775478	1.12460191	2.3785828	0.88574841	5.47372611
55.0999071	1.26626394	0.72026022	10.1649628	24.9419145	0.54600372	0.69702602	2.35827138	0.53438662	3.67100372
54.876858	1.57317999	0.32548552	9.30888575	24.7911468	0.45567972	0.93305848	2.60388413	0.81371379	4.31810784
53.3925862	1.52828962	0.56362454	9.12638196	26.4144808	0.325168	1.14892695	2.48211576	0.69369174	4.32473445

Se propone como objetivo explorar si la composición de las pastas de estas muestras es la misma y obtener información de la existencia de diferentes fuentes de abastecimiento de la cerámica ya que, de acuerdo con García Cook (1996), la variante de los tipos de esta región depende básicamente del acabado de superficie y de la cocción.

Algunos de los cálculos fueron realizados en el entorno estadístico R, desarrollado por R Development Core Team 2011 (Derechos Reservados). También se utilizó el programa de análisis composicional CoDaPack (Comas-Cufi y Thió-Henestrosa, 2011). Para la imputación se utilizó la función Irmi en el programa "compositions" que opera con R. El despliegue gráfico para ver el comportamiento de datos faltantes se hizo con el paquete VIM de R; la identificación de puntos aberrantes se hizo con la función de OutCoda (Templ *et al*., 2011).

Presencia de ceros en las proporciones: exploración de valores faltantes

La mayoría de los datos composicionales contienen casos nulos o ceros. Los valores nulos también se les menciona de manera indistinta como datos faltantes, también denominados perdidos o incompletos. De acuerdo a las transformaciones propuestas por Aitchinson (1986), el caso de la ocurrencia de ceros se vuelve un problema serio debido a que el logaritmo de cero es indefinido. Los datos composicionales que registran la ausencia de valores no se pueden tratar con los log-

cocientes directamente, por lo que debe pensarse bien el procedimiento a seguir cuando se enfrenta esta situación.

Ceros en un componente generalmente se explican de manera doble. Primero, la variable no es realmente cero pero, debido a la falta de herramientas o técnicas de medición adecuadas, es a menudo imposible o demasiado caro obtener algún valor computable para la variable, por lo que se redondea a cero. En este caso está justificado imputar un valor "pequeño" con el fin de que los datos se puedan trabajar dentro del proceso de la transformación log-cociente. Segundo, la variables realmente pueden tomar el valor cero al declararse la ausencia del elemento en algunos casos (ceros esenciales). Por otro lado, no podemos modificar el valor obtenido, ya que podríamos alterar los datos reales originales, siendo que estos probablemente pertenezcan a una población diferente de la que se está estudiando.

La problema de ceros en datos composicionales ha dado origen a su estudio de tal manera que, para poder hacer un tratamiento estadístico adecuado de estos, es necesario estudiar si presentan algún patrón de distribución específico con el fin de utilizar un tratamiento adecuado o, si es el caso, utilizar un método de sustitución. En el estudio y modelado de datos nulos se ha podido establecer una tipología de la ocurrencia de ceros en un conjunto de datos. Rubin (1976) estableció una taxonomía para datos faltantes; esta taxonomía depende de las razones del por qué los datos están perdidos. Advierte que la ausencia de datos debe analizarse como un fenómeno estocástico, por lo que los ceros deben considerarse como una variable aleatoria con distribución de probabilidad conjunta, la cual da cuenta del porcentaje de omisión existente y de su relación con las observaciones completas. En la literatura se han establecido tres distintos tipos de datos faltantes:

(1) Datos faltantes completamente al azar o MCAR (*Missing Completely at Random*). Un MCAR se refiere a que el proceso de la pérdida no depende de las otras variables explicativas en el conjunto de datos. Con MCAR, cualquier observación tiene igual probabilidad de perderse (Mandeville, 2010); es decir, que las variables son independientes y el proceso de que ocurra un dato faltante depende exclusivamente del azar. La probabilidad de ocurrencia de un valor faltante es la misma para todos los individuos en la muestra. El MCAR resulta ser un caso ideal porque el tratamiento de los datos existentes no conduce a sesgo en los parámetros estimados.

(2) Datos faltantes al azar o MAR (*Missing at Random*). Ocurre cuando el valor no fue observado por razones estocásticamente independientes de su valor real (Van den Boogart y Tolosana-Delgado, 2013), entendiendo con esto que el resultado depende básicamente de un proceso aleatorio o no determinista. El caso de encontrar un MAR es mucho más común que los datos MCAR, pero en muchas ocasiones son tratados de forma similar.

(3) Datos faltantes no al azar o NMAR (*Not Missing at Random*). En este caso, incluso teniendo en cuenta toda la información observada disponible, el motivo de observaciones faltantes aún depende de las propias observaciones que no se ven. Incluso después de que los datos observados se toman en cuenta, las diferencias sistemáticas permanecen entre los valores faltantes y los valores observados. En este tipo de valores existe dependencia (estocástica o determinística) entre las observaciones; es decir, un factor puede influir en la ocurrencia de otro factor.

Un método estándar para tratar con valores nulos es simplemente eliminar del análisis cualquier dato que haya registrado un valor de cero. Si una muestra no presenta un componente, se procede a su eliminación y posteriormente se utiliza cualquier análisis convencional basado en datos

completos; pero en este caso se reduce el tamaño de muestra *n*. Este método no resulta ser muy conveniente debido a que, entre mayor es el número de valores nulos, mayor será la eliminación de muestras y, por ende, mayor la perdida de información.

Otra de las estrategias para lidiar con este problema incluye el reemplazo de ceros por una pequeña proporción, de tal forma que todos los ceros son eliminados. Como una solución al problema del redondeo de ceros, Aitchinson (1986) sugirió la reducción del número de componentes en la composición por medio de la fusión (*amalgamation*). Esto es, eliminando las componentes con observaciones de cero al combinarlas con algún otro componente. Esta aproximación no es apropiada cuando el objetivo es modelar la composición original o si el modelo sólo incluye tres componentes, por lo que una aproximación más lógica es reemplazar los ceros redondeados por métodos no paramétricos que no distorsionen seriamente la estructura de covarianzas de los datos (Martín-Fernandez *et al*., 2003). Otra alternativa consiste en recurrir a la sustitución de valores mediante el cálculo del promedio por variable, lo cual no resulta ser muy conveniente debido a que existen variables que en ocasiones presentan valores extremos debido a la distancia que existe entre el valor mínimo y máximo registrados.

Por otro lado, si los datos son multivariables y contienen valores nulos en distintas variables se hace aún más difícil distinguir entre los mecanismos que dan origen a la presencia de los valores nulos. La situación puede ser aún peor cuando existen puntos aberrantes, datos muy heterogéneos o la presencia de datos demasiado sesgados. Por esta razón, la información que pueda ser extraída sobre el comportamiento de los valores nulos puede indicar si algunas partes de los datos deben ser calibrados para los valores que faltan, o si debe realizarse un procedimiento de imputación, es decir, la estimación y la sustitución de los valores nulos (Filzmoser *et al*., 2011; Little y Rubin, 2002). Si el fin es aprovechar toda la información disponible de una muestra, es indispensable abordar el

problema de la presencia de ceros o de valores nulos. Los valores nulos forman parte de un conjunto de observaciones con características especiales que incluye datos agrupados, agregados, redondeados, censurados o truncados; en otras palabras, datos con información especial (Heitjan y Rubin, 1991).

Diversas técnicas de imputación paramétricas y no-paramétricas han sido propuestas para reemplazar el redondeo de ceros y el modelado de ceros esenciales para cuando se trabaja con log-cocientes. Muchos procedimientos de imputación múltiple asumen que los datos se distribuyen normalmente así que, incluyendo variables que no se distribuyen normalmente, pueden introducir sesgos. No se recomienda imputar datos en situaciones en que la omisión en una o más variables alcance porcentajes superiores al 20%. Como procedimiento estándar se puede recurrir a técnicas visualización de datos incompletos; estas técnicas permiten explorar simultáneamente los datos y la estructura de los valores faltantes. Las herramientas de visualización que se proponen en esta sección no se basan en ningún supuesto teórico del modelo estadístico. La utilidad de este procedimiento es tratar de explicar el comportamiento y distribución de los valores faltantes en los datos, permitiendo además identificar posibles estructuras de los datos y de su relación con la información disponible (Filzmore *et al*., 2011). Herramientas de visualización apropiados para los valores que faltan deben ser útiles para distinguir entre los tres mecanismos de valores faltantes.

Tomando en cuenta las 73 muestras analizadas y obteniendo una estadística de valores nulos, se puede apreciar en la Figura 1 tres gráficos de pastel que despliegan en porcentaje el número de variables perdidas. En este caso el gráfico de la izquierda muestra que, de las 10 variables, solo 4 de estas tienen valores nulos. En el gráfico del centro se muestra cuantos casos no poseen valores, los cuales son 15 o un poco más del el 20.55%. El 78.45% corresponde a todos los casos que sí

registran valores. El gráfico de la derecha indica que, de todos los valores, sólo el 2.192% de valores está ausente, es decir, un total de 16 observaciones (las 10 variables multiplicadas por 73 casos da un total de 730 valores, así que hay 16 casos con valores faltantes). Esto indica un pequeño porcentaje mucho menor del 5%.

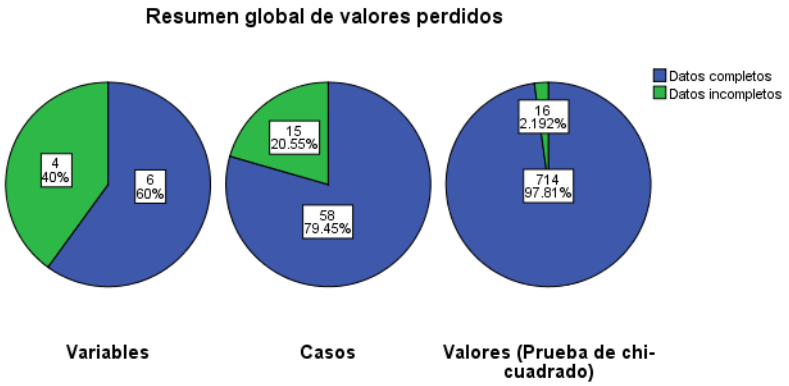

Figura 1. Representación gráfica de valores faltantes en el conjunto de datos.

En la Tabla 3 se presenta la frecuencia de ceros en la matriz de datos original. De acuerdo con el gráfico de la Figura 1, se puede apreciar cuantos valores nulos se tienen en cada una de las variables y como aparecen estos valores nulos en las diferentes variables.

Tabla 3. Variables con valores nulos y su frecuencia.

Na	2
Mg	1
P	13
Ti	3

El paquete VIM, que opera en conjunto con el programa R para la visualización e imputación de valores perdidos (Templ *et al*., 2012), introduce técnicas de visualización de los valores nulos. Se puede utilizar para la preparación y manipulación de datos, para la exploración de los valores faltantes así como para la estimación estadística, con el fin de detectar algún patrón en la

secuencia de los datos faltantes y determinar si se puede identificar la categoría de estos valores de acuerdo con la tipología presentada en la sección de arriba. De estas técnicas se dispone del Gráfico de Agregación.

Gráfico de agregación

Con este tipo de gráfica es posible detectar si existe una cierta combinación de variables con datos faltantes. En la gráfica de la izquierda de la Figura 2 se pueden observar las frecuencias de datos faltantes en cada una de las variables en forma de un histograma, representado la frecuencia por cada barra. Resulta evidente que la variable con mayor número de valores nulos es P, le sigue Ti y con menos ocurrencias Ng y Mg. En la gráfica del lado derecho se pueden observar las combinaciones de valores faltantes y de los no faltantes (un rectángulo rojo indica la presencia de valores nulos y su variable correspondiente; los rectángulos en gris representan datos completos).

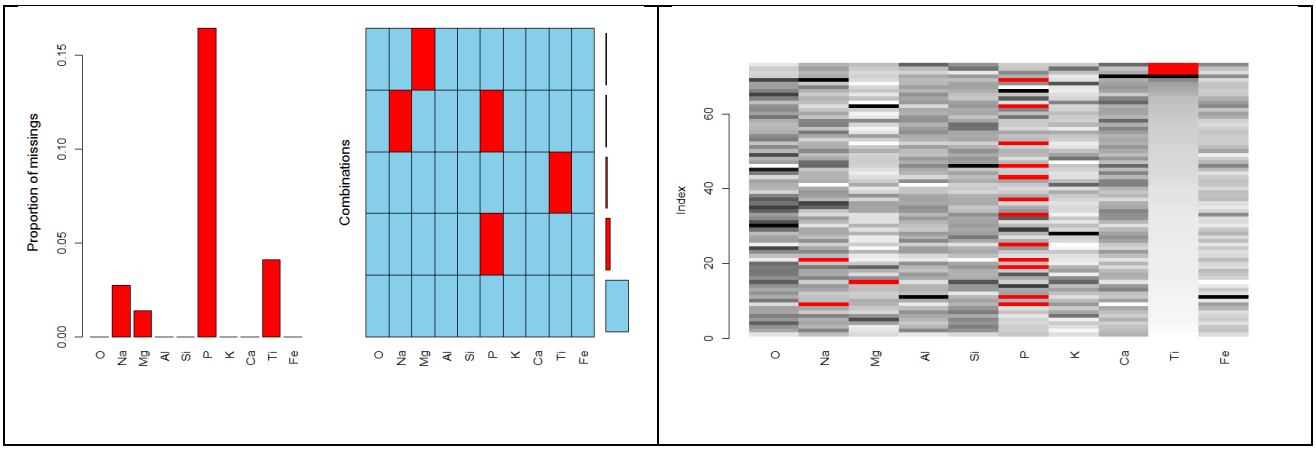

Figura 2. Histograma y grafico de agregación mostrando patrones de valores faltantes.

En la Figura 2 se observa que las ausencias presentan una ocurrencia aleatoria y en cualquier celda pueden existir datos faltantes; es decir, las omisiones no están dispuestas en una forma

predeterminada (por ejemplo, patrones monotónico creciente, faltante condicional). No existe un patrón en la distribución de las líneas en rojo; si hubiera una concentración en orden creciente (la parte izquierda del gráfico) y una concentración de orden decreciente (en el lado derecho) a través de la secuencia, sería un indicador de un patrón sistemático de valores perdidos. Como es visible, solamente se tienen islas de valores nulos y no existe un patrón en cómo están organizadas las líneas en el gráfico. Esto es un indicador de que los valores nulos están registrados debido a un proceso aleatorio.

Imputación de valores faltantes

Varios Métodos de imputación de datos se han desarrollados durante los últimos años. En Estadística, la imputación es el proceso de sustitución de los datos faltantes con los valores sustituidos. Dado que los datos faltantes pueden crear problemas para el análisis de datos, la imputación es vista como una manera de evitar riesgos implicados con la eliminación por lista de casos que tienen valores perdidos. La mayoría de paquetes estadísticos descartan cualquier caso por defecto. La imputación conserva todos los casos mediante la sustitución de los datos faltantes con un valor probable basándose en otra información disponible. Una vez que todos los valores faltantes se han imputado, el conjunto de datos se puede analizar utilizando técnicas estándar de datos completa (ver *http://en.wikipedia.org/wiki/Imputation_(statistics)*).

La imputación es una técnica más sofisticada que únicamente excluye valores o los sustituye con valores pequeños. La calidad del modelo de imputación llega a influir en los resultados finales; es por esta razón que es muy importante considerar el diseño del modelo de imputación. Entonces la pregunta es, de acuerdo al comportamiento de los datos, qué modelo de imputación utilizar. En el

caso de valores multivariables existen tres aproximaciones para imputar valores: métodos de imputación basados en distancias o algoritmo KNN (K nearest neighbour), métodos basados en covarianzas y métodos basados en el modelado como imputación por regresión (Filzmore *et al.*, 2011). Sin embargo, la mayoría de los métodos existentes suponen que los datos se originan a partir de una distribución normal multivariable. Este supuesto se convierte en inapropiado tan pronto existen valores atípicos o puntos aberrantes en los datos. Si el caso, se debe a distribuciones sesgadas o multimodales. Dado que ésta es una situación muy frecuente cuando se trabaja con datos reales, es preferible trabajar con métodos de imputación basada en estimaciones robustas.

El objetivo es proponer el uso de un algoritmo automático llamado IRMI para imputación. Es un método iterativo conocido como *model-based*, el cual utiliza métodos robustos para la estimación y forma parte de los paquetes programados en el programa R.

Algoritmo IRMI

El algoritmo conocido como IRMI para imputación Iterativa es un método robusto basado en modelos (*model-based*). Fue implementado como una función en R en el paquete VIM, cuyas mejoras se encuentran en la estabilidad de los valores iniciales y la robustez de los valores imputados (Templ *et al*., 2011). Este algoritmo reduce la influencia de observaciones atípicas en la estimación de los parámetros de regresión. En cada paso de la iteración, una variable se utiliza como variable de respuesta y las variables restantes sirven como variables predictoras, de tal forma que toda la información multivariable es utilizada para la imputación en la variable de respuesta. A continuación se describen los pasos para aplicar el algoritmo IRMI (tomado directamente de Templ *et al*., 2011).

(1) Inicializa los valores faltantes utilizando una técnica simple de imputación (en este caso se utiliza la media por defecto).

(2) Ordena las variables de acuerdo a la cantidad original de valores perdidos. Asumiendo que las variables ya están ordenadas, es decir, $M\ (x_1) \geq M\ (x_2) \geq \ldots \geq M\ (x_p)$ en donde $M\ (x_j)$, denota el número de celdas que faltan en la variable x_j. Siendo $I = (1, \ldots, p)$.

(3) Sea $l = 1$, denótese $ml \subset \{1, \ldots, n\}$, con los índices de las observaciones originalmente faltantes en la variable *xl*, y $o_l = \{1, \ldots, n\} \setminus ml$ y los índices correspondientes a las celdas con valores observados de *xl*. Sea $X^{ol}_{I\setminus\{l\}}$ y $X^{ml}_{I\setminus\{l\}}$, denotar las matrices correspondientes a las observaciones y a los valores faltantes de *xl*, respectivamente. Además, la primer columna de $X^{ol}_{I\setminus\{l\}}$ y de $X^{ml}_{I\setminus\{l\}}$ consisten de unos (1), teniendo cuidado con el término de intersección en el problema de regresión

$$X^{ol}_{l} = X^{ol}_{I\setminus\{l\}}\beta + \varepsilon, \tag{9}$$

con coeficientes de regresión desconocidos β y un término del error ε. La distribución de la respuesta X^{ol}_{l} es considerada en cada uno de los ajustes de regresión.

(4) Si la respuesta es continua, como en este caso, el enlace es μ, aplicando un método robusto de regresión.

(5) Se estiman los coeficientes de regresión β con el modelo correspondiente en el paso 4, y se utilizan los coeficientes de regresión estimados $\hat{\beta}$ para reemplazar los valores faltantes X^{ml}_{l} por medio de la ecuación:

$$\hat{X}^{ml}_{l} = X^{ml}_{k}\hat{\beta} \tag{10}$$

(6) Realizando los pasos 4 - 5 a su vez para cada $l = 2, \ldots, p$.

(7) Repetir los pasos 3 – 6 hasta que los valores imputados se estabilicen, por ejemplo hasta

$$\sum_i (\hat{x}_{l,i}^{ml} - \tilde{x}_{l,i}^{ml})^2 \langle \delta, \quad \text{para todo } i \in m_l \text{ y } l \in I \tag{11}$$

para una constante pequeña de δ, en donde $X_{l,i}^{ml}$ es el i-*ésimo* valor imputado de la iteración actual y $\tilde{x}_{l,i}^{ml}$ i-*ésimo* valor imputado de la iteración anterior,

Templ *et al*. (2011) asegura que el algoritmo converge en pocas iteraciones y que realmente desde la segunda iteración no hay mejoría significativa. Con este algoritmo, los errores para proporcionar correctas covarianzas son incluidos de una manera adecuada en la iteración final. El método de imputación debe ser "adecuado". Esto implica que la variabilidad que afecta a los valores imputados debe ser incorporada con el fin de tratar los errores estándar de manera consistente (ver valores imputados en la segunda parte de la tabla 2).

Exploración de datos Composicionales: Análisis estadístico

Una vez imputados los datos es conveniente realizar una exploración de los mismos mediante algunas estadísticas descriptivas para datos composicionales, por medio de las cuales se puede tener una idea de la importancia de aquellas variables con mayor variabilidad dentro de la composición. Como una primera aproximación se pueden comparar los centros de las muestras para determinar la identificación de los componentes que pueden discriminar más en la composición. El centro de los datos, así como el análisis de la variabilidad relativa de cada variable (ecuación 3) puede apreciarse

en la Tabla 4, en la que los valores con mayor variabilidad se han remarcado en negritas. La matriz de variación muestra el porcentaje de variación total expresada como pares de log-cocientes de los componentes, así como las medias de los log-cocientes.

Tabla 4. Centro de los datos y matriz de variación

Centro de los datos

O	Na	Mg	Al	Si	P	K	Ca	Ti	Fe
0.0094	0.0529	**0.0726**	0.0084	0.0111	**0.1499**	**0.0801**	0.0261	**0.1191**	0.0343

Matriz de Variación

Xi / Xj	**O**	**Na**	**Mg**	**Al**	**Si**	**P**	**K**	**Ca**	**Ti**	**Fe**	***clr* variances**
O	-	0.0628	0.0765	0.0063	0.0065	**0.1798**	0.0869	0.0428	0.1568	0.0398	0.0094
Na	-3.6073	-	0.128	0.0488	0.0572	**0.2417**	0.156	0.0911	**0.211**	0.0963	0.0529
Mg	-4.4891	-0.8818	-	0.0845	0.0788	**0.2699**	0.1322	0.1243	**0.2784**	0.1172	0.0726
Al	-1.7103	1.897	2.7788	-	0.0075	**0.1837**	0.0959	0.0341	0.1541	0.0324	0.0084
Si	-0.7287	2.8786	3.7604	0.9816	-	**0.2061**	0.0825	0.0397	0.1595	0.0373	0.0111
P	-4.7882	-1.1809	-0.2991	-3.0779	-4.0595	-	**0.3047**	0.1686	**0.26**	**0.2478**	0.1499
K	-3.7615	-0.1542	0.7276	-2.0512	-3.0328	1.0267	-	0.1309	**0.2495**	0.126	0.0801
Ca	-3.0991	0.5083	1.39	-1.3888	-2.3703	1.6891	0.6624	-	0.1341	0.0586	0.0261
Ti	-4.5395	-0.9322	-0.0504	-2.8292	-3.8108	0.2487	-0.778	-1.4404	-	0.1512	0.1191
Fe	-2.5763	1.031	1.9128	-0.866	-1.8476	2.2119	1.1852	0.5227	1.9632	-	0.0343
Media ln(Xi/Xj)											0.5638

En la diagonal superior de la Tabla 4 se puede apreciar que la mayor variabilidad composicional se asocia con las partes de P con Mg, K con P y Ti con Mg (valores altos). Los componentes que retienen la mayor variabilidad resultan ser los idóneos para obtener una subcomposicion de tres partes (Thió-Henestrosa y Martín-Fernández, 2006).

En muchas aplicaciones de datos composicionales es costumbre presentar los datos en diagramas de dispersión, involucrando pares de componentes o cocientes entre dos componentes por eje. Sin embargo, representar las partes de una composición en pares de marginales no es del todo apropiado, ya que las partes de una composición están intrínsecamente vinculadas unas con otras: el conjunto de datos es multivariable en naturaleza, no por decisión del analista (Van den Boogaart y Tolosana-Delgado, 2013). Por esta razón es más conveniente utilizar un despliegue multivariable para poder apreciar de forma simultánea a las partes de la composición. Esto puede lograrse mediante un biplot, que es una gráfica exploratoria y en la que se puede observar al mismo tiempo las variables y las muestras. Las muestras son desplegadas como puntos y las componentes como rayos. El biplot se construye al obtener una descomposición de los valores singulares de la matriz de covarianzas utilizando la transformación *clr*.

El biplot posee la ventaja de poder desplegar en la proyección tanto los datos como las variables: los datos son desplegados como puntos, mientras que las variables se despliegan como vectores o rayos. La interpretación se basa en los enlaces entre los rayos: cada rayo representa una variable *clr* y su longitud se asocia con la varianza explicada en la proyección; las direcciones de los rayos indican aquellas observaciones con un mayor dominio de la parte compositiva. En la Figura 3 se aprecia que la mayor parte de la varianza está representada por los componentes de Ti, Mg y P; de igual forma, la varianza entre Na, Al y K debe ser pequeña. Esto sugiere que los componente que no tienen Ti están poco relacionados con los que sí lo tienen. Ti es un metal abundante en la naturaleza, se encuentra en las cenizas de animales y plantas, por lo que es de pensarse que para la manufactura de estas muestras se utilizaron materiales orgánicos.

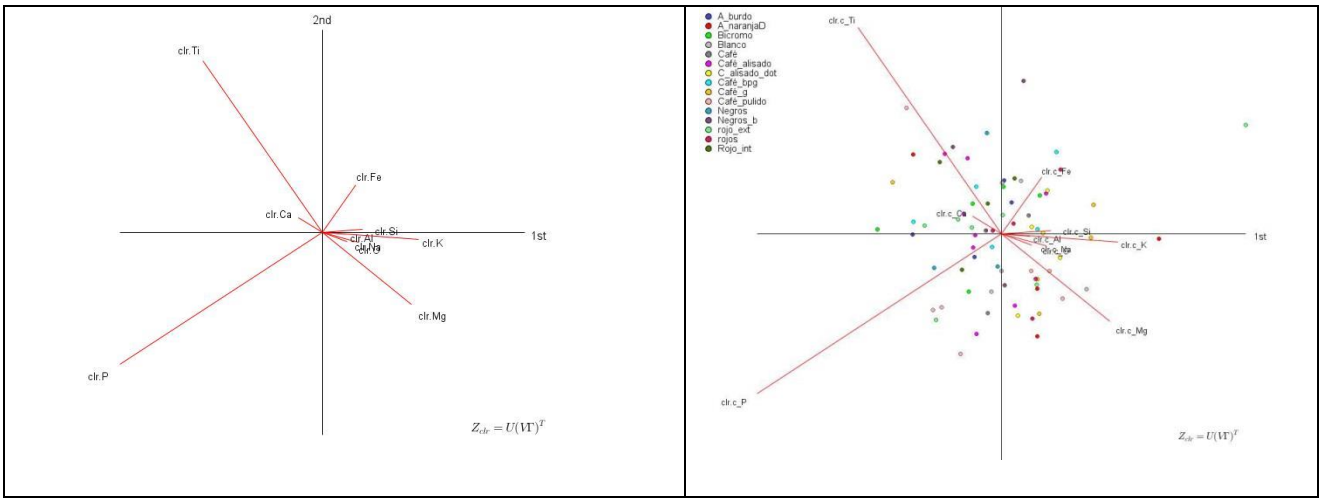

Figura 3. Biplot de las composiciones usando la transformación *clr*: (izq.) solo las partes y (der.) con las observaciones.

El biplot muestra que la proyección de las observaciones se encuentran dispersas a lo largo del mismo y siguen el mismo patrón, por lo que no existe una clara diferenciación que sugiera grupos bien definidos. El conjunto completo de enlaces, especificando todas las varianzas relativas, determina la estructura de covarianza de la composición y proporciona información directa sobre la variabilidad subcompositional y de independencia. Cabe señalar que, en el análisis del biplot composicional, el porcentaje de variación explicada por las componentes (Tabla 5) para la segunda componente es de casi el 60% (59.43) e incluyendo la tercer componente es del 73% (73.62).

Tabla 5. Porcentaje de variación explicada por las componentes

0.3498	0.5943	0.7362	0.8373	0.9185	0.9617	0.9914	0.9967	1.0000

Como se había establecido en el análisis descriptivo, los rayos con mayor variabilidad son los mismos P, K y Ti (rayos más largos) y los más próximos son Na, Al, K y O en el primer eje. La subcomposición entre las diez partes que retienen la mayor variabilidad es [P, K y Ti] y explica casi el 60%. Cuando $D = 3$, entonces S^3 puede representarse gráficamente mediante un diagrama ternario que consiste en un triángulo equilátero en donde los puntos datos son representados mediante el escalamiento interno con una altura igual a 1. En la Figura 4 se representa el triplot de la variabilidad entre las componentes [P, K y Ti].

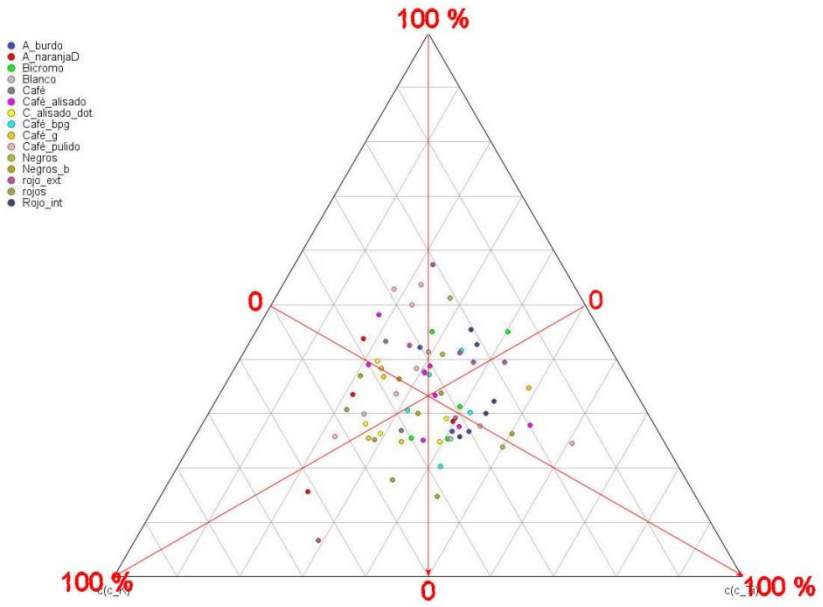

Figura 4. Representación de un gráfico ternario de la variabilidad entre las componentes [P, K y Ti].

En un gráfico ternario es factible que se puedan centrar los datos por medio de una operación de perturbación con motivos de despliegue. En este gráfico es claro que en la distribución de los datos no se pueden apreciar grupos definidos entre los tipos cerámicos composiciones. Las muestras distribuidas cerca del baricentro representarían una composición más homogénea, pero se observa una gran dispersión de los mismos. También es clara la presencia de algunos valores atípicos, los cuales se encuentran más cercanos a la esquina del componente K (parte baja del

gráfico) y algunos más hacia el componente Ti. Este resultado hace pensar que no había una formula común o estandarizada entre los artesanos para manufacturar la cerámica de este sitio, sino que al parecer se utilizaban cantidades variables de las materias primas. Es decir, la mezcla al parecer fue la misma (arcilla proveniente de las mismas fuentes o yacimientos), pero las cantidades agregadas a los componentes de la materia prima fueron dispuestas en algunas ocasiones con mayores concentraciones y en otras en menores cantidades en los porcentajes de las mezclas.

Esto es más evidente si se observa la proyección de las líneas rojas en el diagrama ternario. Conforme se desplaza uno hacía cualquiera de los ejes (por ejemplo, hacia el componente K), las cantidades de este componente varían de 0.20% a 0.50%; yendo en dirección al eje de proyección con el componente Mg, las concentraciones varían de 0.20% a 0.60% y lo mismo pasa con el otro componente que es P. Cabe aclarar que todos estos componentes, como el Mg, forman parte de numerosos compuestos, en su mayoría óxidos y sales; este último es un componente insoluble. Por otro lado, el componente K (potasio) abunda en la naturaleza y se encuentra combinado en fosfatos inorgánicos y junto a organismos vivos, pero nunca en estado nativo. Esto nos indica que en su mayoría se mezclaban cenizas, materia orgánica y sales minerales a la arcilla, pero en porcentajes que no eran agregados de manera proporcional.

Esto permite interpretar que, aunque la cerámica poseía una misma combinación básica de arcilla y temperante, no se logró alcanzar un grado de estandarización en la producción de la cerámica de este sitio, ya que la variabilidad en las muestras refleja que la manufactura de los diferentes tipos es bastante heterogénea. Esto permite descartar una producción centralizada por un poder político. Más bien, todo parece indicar que se trataba de cerámica local con un bajo grado de especialización, manufacturada por artesanos poco especializados; no habría que descartar que

algunas variantes puedan deberse a las condiciones posdeposicionales o de diagénesis. Una duda que surge es que la tipología establecida para esta región abarca un periodo de tiempo de casi 400 años y asume que la producción fue la misma en este periodo de tiempo. La heterogeneidad de las muestras quizás tenga que ver con los cambios que los artesanos van realizando a través del tiempo, de generación en generación no se seguía el mismo patrón en la confección de la receta de las mezclas o que se deba más concretamente a que existen muestras que presentan otra distribución, que puede ser la razón principal por la cual se presentan valores atípicos en el biplot.

Detección multivariada de valores extremos

La detección de valores extremos es una norma en los análisis de estadística multivariada. La presencia de estos casos permite obtener conclusiones acerca de la calidad de los datos y de si presentan cierta tendencia que se desvía de la normalidad. En la distribución de las muestras en el biplot composicional y en el diagrama ternario se pueden observar valores de datos extremos (*outliers*); por esta razón es necesario reconocer cuales de estos datos son los que ocasionan la falta de ajuste del modelo. La definición de un valor extremo es una observación atípica fuera de la masa principal del conjunto de datos (Van den Boogart y Tolosana-Delgado, 2013); estos se comportan de manera diferente en comparación con el resto de las observaciones. Un valor extremo en muchas ocasiones puede ser el resultado de un error en la medición o pueden ser valores que presentan una mayor variabilidad con respecto al conjunto de datos y, cuando se les detecta, su ocurrencia puede deberse a que corresponden a poblaciones diferentes. Para tener una idea más clara sobre los diferentes tipos de valores extremos y su tratamiento ver Van den Boogart y Tolosana-Delgado (2013). Como medida precautoria, los valores extremos deben ser identificados para poder decidir su origen. Una manera sencilla de detección de valores extremos es mediante el cálculo de la

distancia robusta de Mahalanobis $MR(x_i)$ de una composición x_i (Filzmoser y Hron, 2008). Esta medida de distancia multivariable asigna a cada observación una distancia con respecto al centroide, teniendo en cuenta la estructura de varianza, aclarando que para obtener esta métrica se utiliza la transformación *ilr* de los datos (Filzmoser y Gschwandtner, 2011). La distancia de Mahalanobis está dada por la siguiente expresión:

$$MR(x_i) := (ilr(x_i) - m_r)^t \sum_R^{-1} (ilr(x_i) - m_R) \qquad (12)$$

donde m_R es la estimación robusta de la media y $\sum_R$ es la estimación de la varianza. Para determinar si algún punto dato se desvía demasiado de la masa de datos de la estructura de covarianzas, se ha establecido un punto de corte o umbral para reconocer valores extremos de aquellos valores regulares de la muestra. Este valor resulta ser a partir del uso de un cuartil de la distribución X^2 con *D-1* grados de libertad, el cual por lo regular es de 0.95 o del 0.975. De esta manera, valores con una distancia robusta de Mahalanobis grande son considerados puntos extremos. Esta rutina puede obtenerse con el programa OutCoda (Filzmoser y Gschwandtner, 2011) que corre bajo el programa R. Aplicando esta métrica a los datos, se obtienen los resultados presentados en la Tabla 6, donde se denota a los puntos aberrantes de la muestra como V y los valores regulares como F. De estos valores, un total de 23 observaciones exceden el límite del umbral establecido por el cuartil de la X^2 con D-1 *g.l.* En la Tabla 7 se puede observar a las 23 muestras con valores atípicos (el número de la segunda fila corresponde a su identificador o tipo cerámico abreviado).

Tabla 6. Índice de los valores extremos

[1]	V	F	F	F	F	F	V	V	F	F	F	F	F	F	F	F	F	F	F	F
[21]	F	F	F	F	F	F	F	F	F	F	F	F	F	F	V	V	F	F	V	V
[41]	F	V	F	F	V	F	V	F	F	F	V	V	F	V	V	V	F	V	V	F
[61]	V	V	F	F	F	F	F	V	V	V	V	F	F							

Tabla 7. Valores extremos numerados de acuerdo a su identificador

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
AB	AD	AD	CB	CB	CB	CG	CG	CP	CP	CP	N	N	NB	NB	NB	RE	RE	RE	R	R	RE	RI
1	7	8	35	36	39	40	42	45	47	51	52	54	55	56	58	59	61	62	68	69	70	71

Con el programa *mvoutlier* es posible desplegar diferentes gráficas que proporcionan información sobre la estructura de los datos y de la presencia de puntos aberrantes. Las gráficas mostradas en la Figura 5 se obtienen al calcular una descomposición robusta en componentes principales y generando un biplot composicional, con la ventaja de mostrar los valores atípicos mediante símbolos '+' (gráfico de la izquierda), donde se puede apreciar la distribución de datos y la existencia de los puntos aberrantes. En la gráfica de la derecha se muestra la misma proyección del biplot de los datos, únicamente que en esta ocasión se despliegan los outliers con el número de identificación.

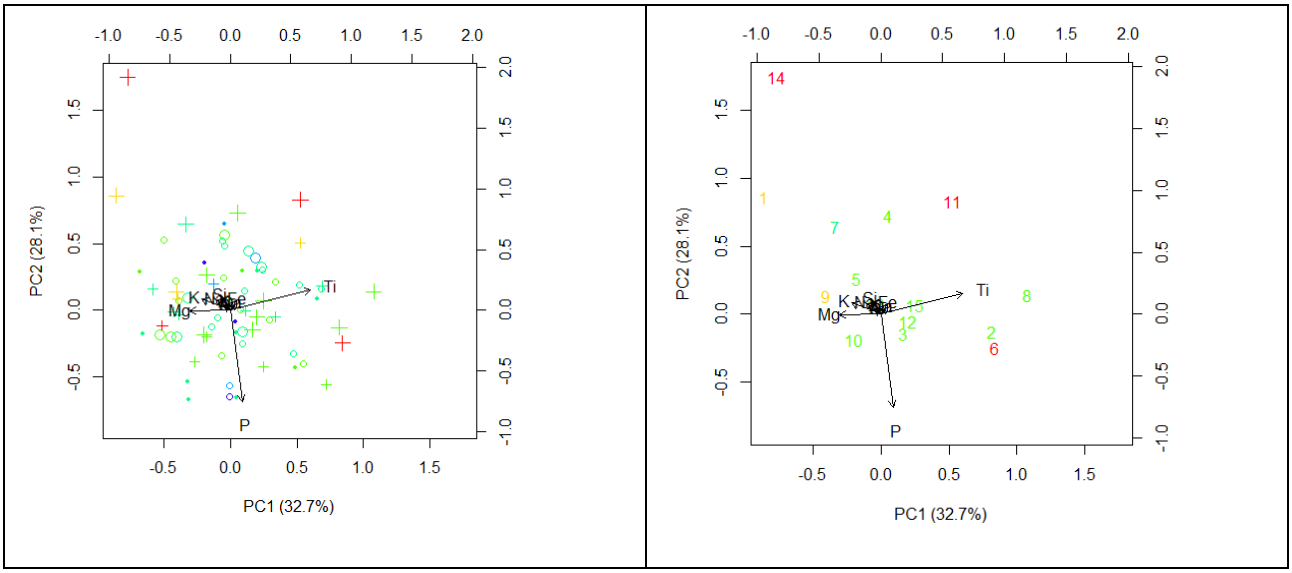

Figura 5. Biplot composicional de valores atípicos. A la izquierda todos los datos son desplegados diferenciando valores atípicos con un signo '+'; en la gráfica de la derecha sólo se representa a los *outliers* con su número de identificación.

Utilizando la distancia de Mahalanobis (de acuerdo al índice de las observaciones), junto con el valor de corte de la X^2, es posible graficar los valores que exceden el umbral, reconocidos como *outliers* potenciales (Figura 6).

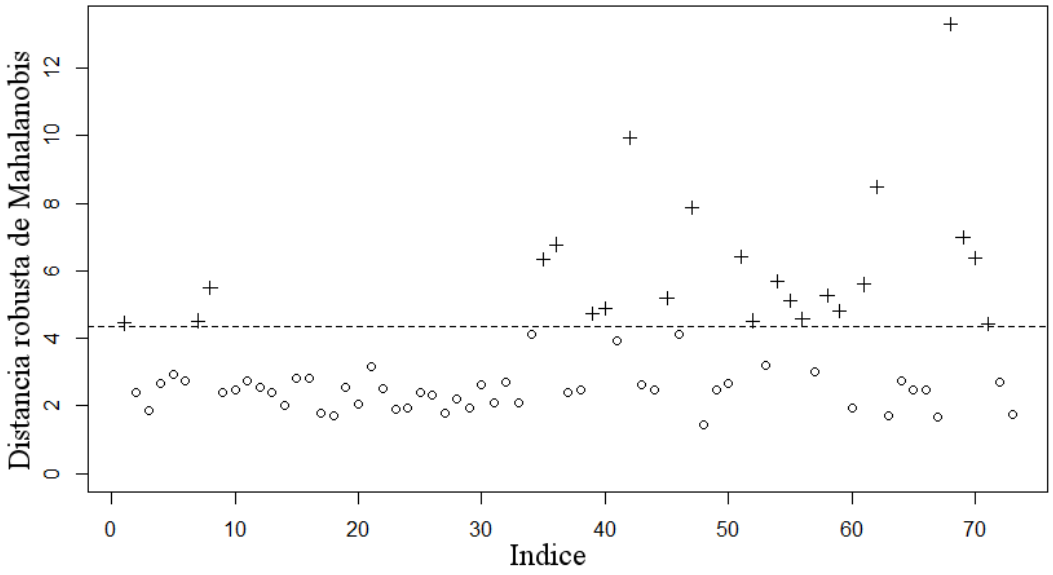

Figura 6. Despliegue gráfico de valores extremos (+) utilizando la distancia de Mahalanobis.

En esta Figura 6 se puede apreciar una fuerte relación entre la mayor parte de los elementos, ya que sus cocientes son casi constantes (rayos concentrados casi en el mismo lugar). Los componentes P y Mg son dominados en un sentido relativo por algunos de los valores aberrantes, lo mismos que el componente Ti. Esto da a entender que estos componentes fueron administrados de manera más abundante en algunas muestras y también puede sugerir la existencia de diferentes fuentes del material o material no local. El componente Ti es casi dominado por las observaciones número 8, 2 y 6, pero las observaciones 1, 7, 4 y 11 tienen valores muy inusuales en comparación con el resto.

Una interpretación más detallada de los valores extremos multivariados es posible mediante un gráfico univariado de las variables *ilr*. En este gráfico se puede observar que los puntos extremos no necesariamente se encuentran en los extremos de la nube de datos de cada una de las variables de manera independiente, sino que pueden encontrarse distribuidos en todo el conjunto de datos (Figura 7).

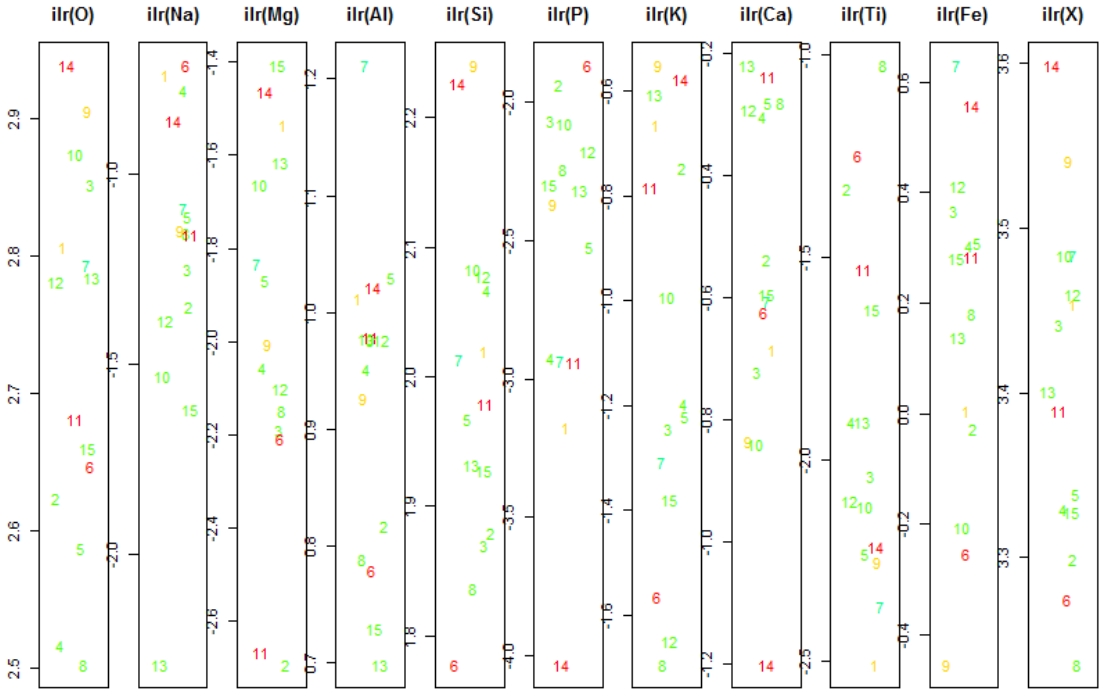

Figura 7. Gráficos de dispersión univariados de las muestras de cerámica. Los valores extremos son mostrados de acuerdo a su número de identificador.

CONCLUSIONES

Se analiza una muestra de cerámica con datos obtenidos por medio del análisis EDX con la finalidad de contrastar la información con la tipología original establecida para la región de Tlaxcala. El método utilizado es el análisis de datos composicionales establecido por Aitchinson (1986) y se pone especial atención en la transformación log-cociente de los datos, a la ocurrencia de

ceros, en el método de imputación utilizado y la detección de valores extremos en la muestra. Los resultados obtenidos permiten discriminar la formula química utilizada en los materiales cerámicos estudiados. Esto evidencia que no se puede hablar de la existencia de tipos cerámicos bien definidos y que en la confección de la cerámica existió una variabilidad muy alta, por lo que es recomendable redefinir la tipología de la zona. Otra motivación del trabajo es proporcionar métodos con buen rendimiento en el análisis de datos composicionales cuando hay pequeñas desviaciones de distribuciones paramétricas, proporcionando los elementos que ofrezcan resultados robustos frente a posibles desviaciones.

APÉNDICE: Programas estadísticos

Programa *Compositions* (Van der Boogaart y Tolosana-Delgado, 2013) que opera bajo programa estadístico R el cual fue desarrollado para trabajar exclusivamente con datos composicionales (Beardah y Baxter, 2005). Este programa contiene rutinas de estimaciones robustas de datos para imputar datos, detectar observaciones atípicas y operar con subcomposiciones. Creación de gráficos ternarios biplots, dendograma y balances. El lenguaje de programación R es de dominio público (R. Development Core Team, 2011) con aplicación para estadística y análisis de datos. El ambiente básico de R puede ser descargado de http://cran.r-project.org. El programa para imputación de datos puede obtenerse de la página http://www.statistik.tuwien.ac.at/public/filz/programs.html (Hron *et al.*, 2010).

BIBLIOGRAFÍA

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Baxter, M.J., Beardah, C.C., y Freestone, I.C. (2005). "Compositional analysis of archaeological glasses". En: (G. Mateu I Figueras y C. Barceló I Vidal, eds.) *CoDaWork'05. Compositional Data Analysis Workshop*. Universitat de Girona, Spain. Disponible en: http://dugi-doc.udg.edu//handle/10256/688.

Beardah, C.C., y Baxter, M. J. (2005). "An R Library for Compositional Data Analysis in Archaeometry". En: (G. Mateu I Figueras y C. Barceló I Vidal, eds.) *CoDaWork'05. Compositional Data Analysis Workshop*. Girona, Spain. Disponible en: http://ima.udg.es/Activitats/CoDaWork05/CD/Session5/Beardah-Baxter.pdf

Buxeda i Garrigos, J. (1999). "Alteration and contamination of archaeological ceramics: The perturbation problema". *Journal of Archaeological Science* 29: 295-313.

Comas-Cufí, M., y Thió-Henestrosa, S. (2011). CoDaPack 2.0: a stand-alone, multi-platform compositional software. En: (J.J. Egozcue, R. Tolosana-Delgado y M.I. Ortego, eds.) CoDaWork'11: 4th International Workshop on Compositional Data Analysis. Universitat de Girona, Sant Feliu de Guíxols, Spain. Disponible en: http://ima.udg.edu/codapack/

Daunis-i-Estadella, Barcelo-Vidal. J y Buccianti, A. (2006). "Exploratory compositional data analisys". En: (A. Buccianti, G. Mateu-Figueras y V. Pawlowsky-Glahn, eds.) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Special Publication vol. 264, Geological Society, London, pp.161-174.

Egozcue, J. J., y Pawlowski-Glahn, V.

- (2006). Simplicial geometry for compositional data. En: (A. Buccianti, G. Mateu-Figueras y V. Pawlowsky-Glahn, eds.) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Special Publication vol. 264, Geological Society, London, pp. 145- 159.
- (2011). "Análisis composicional de datos en Ciencias Geoambientales". *Boletín Geológico y Minero* 122 (4): 439-452.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). "Isometric log-ratio transformations for compositional data analysis". *Mathematical Geology* 35 (3): 279–300.

Egozcue, J., Barceló-Vidal, C., Martín-Fernández, J., Jarauta-Bragulat, E., Díaz-Barrero, J. L., y Mateu-Figueras, G. (2011). "Elements of simplicial linear algebra and geometry". En: (V. Pawlowsky-Glahn y A. Buccianti, eds.) *Compositional Data Analysis: Theory and Applications*. Wiley, UK, pp. 141-157.

Filzmoser, P., y Gschwandtner, M. (2011). Package 'mvoutlier': Multivariate outlier detection based on robust methods. Manual and package, version 1.9.1. Disponible en: http://cran.r-project.org/package=mvoutlier

Filzmoser, P., y Hron, K. (2008). "Outlier Detection for Compositional Data Using Robust Methods". *Mathematical Geosciences* 40 (3): 233–248.

García Cook, Ángel (1996). "El desarrollo cultural prehispánico en el norte del área, intento de una secuencia cultural". En: (L. Mirambell, coord.) *Antología de Tlaxcala*, Vol. I. Gobierno del estado de Tlaxcala – INAH, México, pp. 247-254.

Glascock, M. D. (1992). "Characterization of archaeological ceramics at MURR by Neutron Activation Analysis and Multivariate Statistics". En: (H. Neff, ed.) *Chemical characterization of ceramic pastes in archaeology*. Prehistory Press, Madison, pp. 11-26.

Glascock, M. D., Neff, H. y Vaughn, K. J. (2004). "Instrumental Neutron Activation Analisys and Multivariate Statistics for Pottery Provenance". *Hyperfine Interactions* 154: 95-105.

Heitjan, D. F., y Rubin, D. B. (1991). "Ignorability and Coarse Data". *The Annals of Statistics* 19: 2244-2253.

Hron, K., Templ, M., y Filzmoser, P. (2010). "Imputation of missing values for compositional data using classical and robust methods". *Computational Statistics and Data Analysis* 54 (12): 3095-3107.

Jackson, C. M. y Baxter, M. J. (1999). "Variable Selection in Archaeometry: the Statistical Analysis of Glass Compositional Data". En: (J.A. Barceló, I. Briz y A. Vila, eds.) *New Techniques for Old Times - CAA'98. Computer Applications and Quantitative Methods in Archaeology*. Proceedings of the 26th Conference, Barcelona, March 1998 (BAR International Series 757). Archaeopress, Oxford, pp. 159-162.

Kucera, M., y Malmgren, B. A. (1998). "Logratio transformation of compositional data – a resolution of constant sum constraint". *Marine Micropaleontology* 34: 117-120.

Little, R. J. A. y Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Makowski, K., Ghezzi, I., Guerrero, D., Neff, H., Jiménez, M., Ore, G., y Álvarez-Calderón, R. (2008). "Pachacamac, Ychsma y los Caringas: Estilos e identidades en el valle de Lurín inca". En: (O. Pinedo y H. Tantaleán, eds.) *Arqueología de la costa centro sur peruana*. Avqi ediciones, Lima, pp. 267-316.

Mandeville, P. (2010). "Tema 24: Observaciones perdidas". *CIENCIA UANL* 13 (3): 325-328. Disponible en: http://eprints.uanl.mx/2183/1/tipsbioestadisticos.pdf

Martín-Fernández, J. A., Barceló-Vidal, C., y Pawlowsky-Glahn, V. (2003). "Dealing with zeros and missing values in compositional data sets". *Mathematical Geology* 35 (3): 253-278.

Reyment, R. A. (2006). "On stability of compositional canonical variate vector components". En: (A. Buccianti, G. Mateu-Figueras y V. Pawlowsky-Glahn, eds.) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Special Publication vol. 264, Geological Society, London, pp. 59-66.

Rubin, Donald B. (1976). "Inference and Missing Data". *Biometrika* 63 (3): 581-592.

Pawlowski-Glahn, V. y Egozcue, J. J. (2006). "Compositional data and their analysis: an introduction". En: (A. Buccianti, G. Mateu-Figueras y V. Pawlowsky-Glahn, eds.) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Special Publication vol. 264, Geological Society, London, pp. 1-10.

Templ, M., Kowarik, A., y Filzmoser, P. (2011). "Iterative stepwise regression imputation using standard and robust methods". *Computational Statistics and Data Analysis* 55: 2793- 2806.

Templ, M., Alfons, A., y Filzmoser, P. (2012). "Exploring incomplete data using visualization techniques". *Advances in Data Analysis and Classification* 6 (1): 29-47.

Thió-Henestrosa, S., y Martín-Fernánez, J. A. (2006). "Detailed guide to CoDaPack: a freeware compositional software". En: (A. Buccianti, G. Mateu-Figueras y V. Pawlowsky-Glahn, eds.) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Special Publication vol. 264, Geological Society, London, pp. 101-118.

Van den Boogaart, K. G., y Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Springer-Verlag, Berlin-Heidelberg.

Páginas de internet

http://www.r-project.org/ Nota: R es un programa de código abierto que pone a disposición un poderoso arreglo funciones estadísticas y graficas sin ningún costo.

http://en.wikipedia.org/wiki/Imputation_(statistics)