LAS ORGANIZACIONES POLÍTICAS EN LAS ELECCIONES PRESIDENCIALES PERUANAS DE 2011 USANDO ANÁLISIS DE COMPONENTES PRINCIPALES[1]

BAZAN, Jorge Luis[2] *SULMONT, David*[3] *CALDERÓN, Arturo*[4]

RESUMEN: El propósito de este trabajo es presentar una aplicación de la metodología de datos composicionales en datos de la elección presidencial peruana de la primera vuelta del 2011 a nivel provincial y llamar la atención acerca de la importancia de este tipo de metodologías que reconocen de mejor manera la naturaleza real de los datos. Para los aspectos metodológicos fueron considerados los datos multipartidarios como análisis de datos composicionales. Como resultados importantes en este trabajo fue que los dados electorales son de carácter composicional, siendo la proporción de votos el componente más importante, que incluso el número de votos existentes.
PALABRAS CLAVES: Organizaciones Políticas, Elecciones Presidenciales, Componentes Principales.

SUMMARY: The purpose of this paper is to present an application of compositional data methodology on data from the Peruvian presidential election in the first round of the 2011 provincial and draw attention to the importance of such methodologies recognize best so the real nature of the data. For methodological aspects were considered multiparty data as compositional data analysis. As important results in this work was that the electoral dice are compositional in nature, with the proportion of votes the most important, that even the existing number of votes.

KEY WORDS: Political Organizations, Presidential Elections, Principal Components.

1. INTRODUCCIÓN

Una de las dificultades de conjunto de datos con un alto número de variables es su consiguiente alto número de dimensiones. Para una persona promedio es difícil visualizar un espacio de más de tres dimensiones. Análisis de componentes principales (ACP) es una técnica multivariada usada para reducir el número de dimensiones de conjunto de datos multidimensionales. Esta metodología es capaz de identificar variables inherentes a la estructura inicial de un conjunto de datos y establece un significado físico entre variables y objetos, (ver Jolliffe 2002).

Las votaciones por circunscripciones electorales, o datos electorales multipartidarios, por ejemplo los votos válidos de los partidos políticos en las elecciones peruanas a nivel departamental en la elección del 2011, son ejemplos de datos multivariados que presentan ciertas particulares: la suma de sus componentes (las votaciones de los diferentes partidos), siempre positivas, es una constante (el total de votos válidos). Este tipo de datos son llamados datos composicionales (Aitchison, 1986).

[1] **Agradecimientos**. Este trabajo fue elaborado como parte del Proyecto DGI 2010-0173 de la Pontificia Universidad Católica del Perú.
[2] Departamento de Ciencias. Pontificia Universidad Católica del Perú
[3] Departamento de Ciencias Sociales. Pontificia Universidad Católica del Perú
[4] Departamento de Ciencias. Pontificia Universidad Católica del Perú

Para llamar la atención del lector acerca de este tipo de datos proponemos el siguiente ejemplo. Consideremos los votos de las provincias (A, B) entre tres partidos políticos excluyentes (X, Y, Z) en dos procesos electorales (2006, 2011). Considere que los votos para cada partido en miles de votos en la primera elección es [10, 80, 30] para la provincia A y [20, 80, 20] para la provincia B. En la elección 2011 imagine que estos votos son, respectivamente, [20, 70, 30] y [30, 70, 20]. Note en la provincia A los votos para el partido X se incrementa en un 100 % de una elección a otra, mientras que para el partido Y se reduce en un 12.5% y no cambia para el partido Z. Por otro lado, en la provincia B se incrementan los votos dedicados al partido X en un 50%, mientras que se reduce en un 12.5% el voto dedicado al partido Y y tampoco se modifican las votaciones dedicadas a Z. Una medida de diferencia adecuada debería detectar un mayor cambio en la distribución de votos de la provincia A que en el de la provincia B entre las dos elecciones. Sin embargo, si medimos esta diferencia entre periodos utilizando la distancia euclídea habitual obtenemos un valor de 14.14214, el mismo para las dos provincias, con lo cual podríamos equivocadamente concluir que en ambas provincias se ha producido las mismas variaciones de un proceso electoral frente a otro. Un ejemplo similar en el contexto del análisis de los patrones de actividad de dos operarios es proporcionado por Palarea-Albaladejo, Martín-Fernández y Gómez-García (2007).

Katz y King (1999) llamaron la atención que los datos electorales multipartidarios son esencialmente composicionales. Lo que sucede es que frente a los datos composicionales, un cambio en una componente (la votación en un determinado partido político) conlleva un cambio en, al menos, una de las demás componentes (de los otros partidos políticos) y de esta manera cualquier medida basada de forma más o menos explícita en la distancia Euclídea no se comportará bien en este contexto.

Otros ejemplos de este tipo de datos ocurren cuando se trabaja con datos que representan partes de un total: proporciones, porcentajes, partes por millón, o similar. Es el caso de, por ejemplo, la distribución del gasto familiar en diferentes componentes, la composición de una cartera de inversión, el empleo del tiempo diario en distintas actividades, la distribución de ventas en distintas regiones, y como hemos adelantado, la distribución de votos en diferentes partidos que se presentan en un determinado proceso electoral. Los datos composicionales no son raros, aparecen no sólo en ciencias políticas, sino también en geoquímica, genética, ecología, ciencias de la alimentación y muchas otras áreas.

Las restricciones de no negatividad y de suma constante que caracterizan este tipo de datos implican que las técnicas multivariantes habitualmente utilizadas no son adecuadas para su análisis y modelización. La cuestión clave es que la geometría del espacio muestral euclídea debe ser substituida por una geometría diferente.

Cuando hay una necesidad de analizar e interpretar una elección por circunscripciones electorales (departamentos, provincias, distritos por ejemplo), la variabilidad dentro de cada circunscripción electoral deben tenerse en cuenta y el análisis estadístico y la interpretación debe hacerse con cuidado. En particular la restricción de que la suma sea la unidad no debe ser ignorado o mal incorporado en los modelos estadísticos, no hacerlo puede llevar a resultados inadecuados o irrelevantes (Aitchison, 1986). Por otra parte, los conceptos más básicos, como la covarianza y la correlación, los cuales suelen ser base para análisis complejos no tienen interpretación simple como lo hacen cuando se aplica a datos que no son composicionales Esto tiene implicaciones fuertes para la mayoría de los procedimientos multivariantes que se basan en la matriz de covarianza / correlación, en particular, el análisis de componentes principales (ACP) que es el enfoque de este trabajo.

Como indican Palarea-Albaladejo, Martín-Fernández y Gómez-García (2007), en este contexto, hay que replantearse, entre otros, el concepto de independencia. Además, una fila cualquiera de la matriz de covarianzas de una muestra de vectores de proporciones siempre tiene al menos un elemento negativo y su suma es igual a 0. Esto implica que la matriz de covarianzas es singular y que las correlaciones no varían libremente en el habitual intervalo [-1,1]. Este tipo de problemas ya había sido anticipado por Pearson (1897), citado en Mosimann (1962) quien llamó “correlación espúrea” a la generada entre variables que representan parte de un total y que suman una cantidad constante.

Estos hechos hacen que el concepto de independencia sea cuestionado en este contexto y pone en cuestionamiento la aplicación de métodos como el análisis de componentes principales, basado en la información de las covarianzas o las correlaciones de los datos.

Como indica Winzer (1999), otro problema relacionado con el anterior, es que los gráficos de componentes principales suelen mostrar formas curvas ó triangulares, y no elípticas. Esto evidenciaría que la relación entre las variables no es lineal y, por lo tanto, plantea la inconveniencia del uso de medidas como la covarianza o la correlación.

No es hasta los años 80 con la publicación de la monografía de Aitchison (1986) cuando se dispone de una obra de referencia sobre los fundamentos teóricos y sobre una metodología específica para el análisis estadístico de tales vectores de proporciones, que desde entonces suelen denominarse datos composicionales.

Recientemente, Rodrigues y Lima (2009) han presentado el análisis de la elección de la unión europea usando análisis de componentes principales ignorando y tomando en cuenta la naturaleza composicional de este tipo de datos.

El propósito de este trabajo es presentar una aplicación de la metodología de datos composicionales en datos de la elección presidencial peruana de la primera vuelta del 2011 a nivel provincial y llamar la atención acerca de la importancia de este tipo de metodologías que reconocen de mejor manera la naturaleza real de los datos.

En la sección 2 resumiremos los fundamentos sobre los que se asienta el análisis de datos composicionales aplicado a los datos multipartidarios. A continuación, en la sección 3, destacaremos los aspectos más relevantes de la metodología basada en transformaciones para trasladar los datos al espacio real donde poder aplicar las técnicas habituales. Al mismo tiempo, pondremos de relieve algunos de los problemas prácticos que pueden surgir. La sección 4 presenta la aplicación de las herramientas del análisis de datos composicionales en el campo del análisis político y específicamente en el análisis de datos electorales de multi partidos de la elección peruana del 2011. Finalmente en la última sección presentamos una discusión de los resultados encontrados, así como sugerimos líneas de investigación futura.

2. DATOS MULTIPARTIDARIOS COMO ANÁLISIS DE DATOS COMPOSICIONALES

Nosotros ahora identificamos las características estadísticas de los datos electorales multipartidarios.

Considere V_{ij} que denota la proporción de los votos en la circunscripción electoral (departamento, provincia, distrito, mesa) i $(i = 1, \cdots, n)$ para el partido j $(j = 1, \cdots, p)$.

Dos fundamentales características de los datos de votos multipartidarios son que cada proporción está en el intervalo unitario

$$V_{ij} \in [0, 1] \; para \; todo \; i \; y \; j \tag{1}$$

Y que el conjunto de proporciones de votos para todos los partidos en una circunscripción suma uno:

$$\sum_{j=1}^{p} V_{ij} = 1 \qquad para \; todo \; i \tag{2}$$

Las variables que satisfacen estas dos restricciones se dice que están en una región generalmente referida como el simplex el cual se puede definir como el siguiente conjunto

$$S_i^p = \left\{ (v_{i1}, \cdots, v_{iJ}) : \sum_{j=1}^{p} V_{ij} \leq 1, V_{ij} \geq 0, \;\; para \; todo \; i \right\}$$

Para entender mejor el espacio muestral simplex lo presentaremos gráficamente para el caso de dos partidos. Considere los resultados de la segunda vuelta de la elección presidencial 2011 (ver cuadro 1 en el Anexo 1), nosotros usamos

V_{iF} para los votos de la candidata Keiko Fujimori y V_{iH} para los votos del candidato Humala, donde i representa el departamento del país (ver datos en Anexo). Obviamente, nosotros podemos fácilmente representar ambas variables pero podemos escoger una, digamos, V_{iH} porque la otra es meramente $V_{iF} = 1 - V_{iH}$. La figura 1, plotea V_{iF}. por V_{iH} Debido a la restricción de la ecuación 2, todos las fracciones de votos en las circunscripciones caen en una simple segmento de recta, y debido a la restricción 1, la línea finaliza en los ejes.

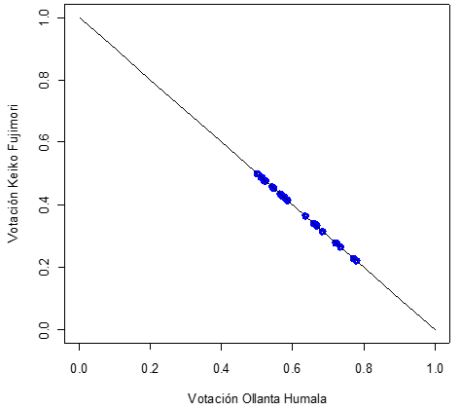

Figura 1. El Simplex para dos partidos en la Elección Presidencial Peruana de 2011 Segunda vuelta

La misma lógica se puede aplicar para elecciones con multi partidos $J > 2$ como ocurre en las elecciones presidenciales en primera vuelta del 2011 a nivel de departamentos, aunque claro, las representaciones graficas se tornan mas complicadas. Por ejemplo, considere ahora V_{iF} para los votos de la candidata Keiko Fujimori, V_{iH} para los votos del candidato Humala y V_{iO} para los otros candidatos (ver datos en Cuadro 1 en Anexo).

En este caso obtenemos un tetraedro.

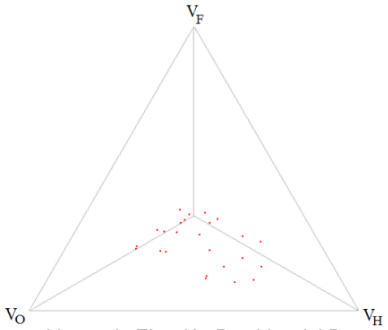

Figura 2. El Simplex para tres partidos en la Elección Presidencial Peruana de 2011 Primera vuelta

3. TRANSFORMACIÓN DE LOS DATOS, USO DEL CONTRASTE LOGARÍTMICO

Una metodología adecuada para el análisis de datos composicionales como los votos de elecciones multipartidarias debe tener en cuenta algunos principios lógicos necesarios y las características del símplex como espacio muestral sobre el que se definen. La idea principal es que las composiciones, los votos por los partidos, sólo proporcionan información sobre la magnitud relativa de sus partes, y no pueden justificarse interpretaciones que involucren a las magnitudes absolutas. Se asume que el valor de la suma de las partes es irrelevante, en este caso el total de votaciones. Por lo tanto, cualquier aseveración sobre las votaciones de los partidos debe hacerse en términos de los cocientes entre ellos, los cuales medirán dicha relación relativa. Así, una función aplicable sobre las votaciones de los partidos deberá ser invariante por cambios de escala y expresable en términos de cocientes entre los partidos. Trabajar con cocientes asegura además un principio lógico básico: la coherencia subcomposicional.

Una vez que se ha puesto de relieve la necesidad de centrar la atención en los cocientes entre las partes surge la pregunta sobre qué tipo de cocientes utilizar. La piedra angular de la metodología propuesta por Aitchison (1986) es la transformación de una composición definida sobre el simples S^p en un vector que involucre los cocientes entre las partes y que esté definido sobre el espacio real. Si esa transformación es biyectiva se establece una correspondencia uno a uno entre las composiciones en el símplex y los correspondientes vectores transformados reales. De esta manera, cualquier problema que afecte a composiciones queda expresado en términos de tales vectores transformados, con lo que se tiene la posibilidad de resolverlo utilizando las técnicas multivariantes habituales en espacios reales.

En este caso considere n provincias y P partidos políticos y las proporciones de votos conseguidas $V_{ij}, i = 1, \cdots, n\ , j = 1, \cdots, P$ las cuales se asumen que son completamente composicionales (si no lo fueran, se pueden usar algunas transformaciones como las sugeridas por Barceló-Vidal (2003),

Aitchison propone fundamentalmente la transformación log-cociente centrada (clr) sobre R^p, definida como

$$clr(V_i) = \left[ln\left(\frac{V_{ij}}{g_i}\right), \cdots, ln\left(\frac{V_{ip}}{g_i}\right)\right] \qquad (3)$$

Donde $g_i = \left[\prod_{j=1}^{p} V_{ij}\right]^{1/p}$ es la media geométrica.

La transformación log cociente centrada puede también escribirse de la siguiente manera

$$clr(V_i) = \left[ln(V_{ij}) - ln(g_i), \cdots, ln(V_{ip}) - ln(g_i)\right]$$

con $ln(g_i) = \frac{\left[\sum_{j=1}^{p} \ln(V_{ij})\right]}{p}$, denotemos ahora los valores transformados de las proporciones de votos por $V_{ij}^{*} = ln\left(\frac{V_{ij}}{g_i}\right)$ y que note que se cumple que $\sum_{j=1}^{p} V_{ij}^{*} = 0\ para\ todo\ i.$

El hecho de tomar logaritmos de los cocientes sólo responde a una conveniencia matemática. Los log-cocientes son más manejables y además permiten que se cumplan algunas propiedades sencillas. Por ejemplo, no existe una relación exacta entre las varianzas $var(V_{im}/V_{in})$ y $var(V_{in}/V_{im})$ y, sin embargo, sí se cumple que $var\Big(ln\big(V_{im}/V_{in}\big)\Big)$=$var\Big(ln\big(V_{in}/V_{im}\big)\Big)$

Existen otras transformaciones propuestas como las transformaciones alr y ilr las cuales presentan ciertas características que las hacen deseables. Para detalles ver Barceló-Vidal (2003)

Por ejemplo la transformación alr está más vinculada a modelizaciones paramétricas ya que a partir de ella es posible definir una clase de distribuciones, la normal logística aditiva, (Aitchison y Shen, 1980), sin embargo, por otro lado presentan dificultades en la interpretación y en la suposición de que siguen un determinado modelo paramétrico. En contraste, la transformación clr está más vinculada a contextos no paramétricos, ya que a partir de ella es posible especificar la distancia de Aitchison en términos de la distancia Euclídea entre los vectores clr-transformados, esto es,

$$d_a(x, x^{'}) = d_e(clr(x), clr(x^{'}))$$

La transformación clr es simétrica e isométrica, pero la imagen de S^p queda realmente restringida a un subespacio de R^p y la matriz de covarianzas de los datos clr-transformados es singular, del mismo modo que ocurría con la matriz de covarianzas aplicada directamente sobre las composiciones.

Después de la aplicación de esta transformación logratio centrado en las P variables en cada de las n observaciones suman cero. Un ejemplo de este resultado se puede encontrar en Kucera y Malmgren (1998). Si alguno de los elementos de los datos originales es igual a cero, la transformación logratio no se puede realizar de la misma manera. Sin embargo, Aitchison (1986) presenta un procedimiento de substitución del cero (o error de redondeo de reemplazo, Kucera y Malmgren 1998), donde todos los elementos cero se sustituyen por un valor $0{,}0 + \epsilon$ y al resto se debe restar $\frac{\epsilon \times r}{p-r}$, con el fin de mantener la suma de cada composición de la muestra constante donde r y p denotan el número de ceros y el número de variables respectivamente. Por ejemplo si tenemos 8 variables y en 3 de ellas hay ceros, en

estas debemos substituir el 0 por digamos, 0.005 y para las otras 5 variables debemos restar las proporciones de $0.005 \times 3/5$

Otras opciones para reemplazar los ceros redondeadas en los conjuntos de datos de composición se pueden encontrar en Palarea-Albaladejo y Martín-Fernández (2008) y Butler y Glasbey (2008).

Tomando en cuenta esta transformación la matriz de varianza covarianza de logratio centrada (Rodrigues y Lima, 2009) $P \times P$ es dada por

$$\Gamma = [\sigma_{jK}] = \left[cov\left(V_{ij}^*, V_{ik}^*\right) : j, k = 1, \ldots, p = P\right]$$

Para mas detalles acera de las propiedades de logratios, logcontrastes y análisis de componentes principales de logcontraste ver Aitchison (1986).

4. EL ANÁLISIS DE LOS DATOS ELECTORALES PERUANOS DEL 2011 USANDO ACP

La idea central del ACP es reducir la dimensionalidad de un conjunto de datos constituida por un gran número de variables relacionadas entre sí, conservando lo más posible la variación presentado en los datos originales. Esta reducción se logra mediante la transformación de las variables originales en un nuevo conjunto de variables, los componentes principales, que no están correlacionados ¿con? combinaciones lineales de las variables originales. Si el primero (o algunos) los componentes principales conserva la mayor parte de la variación presente en las variables originales, es posible construir biplots (Bradu y Gabriel 1978, Gabriel 1971) y por lo tanto la visualización de las variables que subyacen a la estructura original. Para más detalles, véase, por ejemplo, Jolliffe (2002).

ACP ha sido estudiada como una técnica para reducir la dimensión de los conjuntos de datos y en el reconocimiento de patrones de datos composicionales dimensional, la mayoría de ellos se aplica a datos geoquímicos (Chayes y Trochimczyk 1978; Thi Henestrosa y Martín-Fernández, 2005), sobre la base de una estructura de covarianza cruda (Aitchison, 1986).

Los datos usados en la presente aplicación consisten en los resultados (número de votos) de la elección presidencial peruana de la primera vuelta en el año 2011 para cada una de los 24 departamentos y una provincia constitucional del país. En este proceso participaron 11 organizaciones electorales los cuales obtuvieron los siguientes porcentajes de votos según la Oficina Nacional de Procesos Electorales: Gana Perú (GP: 31.699 %), Fuerza 2011 (F2011: 23.551 %), Alianza por el Gran Cambio (APEGC: 18.512 %), Perú Posible (PP: 15.631 %), Alianza Solidaridad Nacional (ASN: 9.832 %), Fonavistas del Perú (0.253 %), Despertar Nacional (0.147 %). Adelante (0.118 %), Fuerza Nacional (0.115 %), Justicia, Tecnología, Ecología

(0.077 %), Partido Descentralista Fuerza Social (0.064 %) (Datos publicados en http://www.elecciones2011.onpe.gob.pe/resultados2011/1ravuelta/).

Tomando en cuenta estos resultados y el bajo porcentaje alcanzado para las organizaciones políticas ubicadas entre el 6to y 11vo lugar, estas se agruparon en una categoría general denominada "Otras organizaciones electorales", como en el cuadro 1.

Cuadro 1. Resultados electorales en primera y segunda vuelta en la Elección Presidencial Peruana 2011 a nivel de departamentos y una provincia Constitucional

DEPARTAMENTOS	PRIMERA VUELTA							
	GP	F2011	APEGC	PP	ASN	OTROS	VOTOS VALIDOS	Otros sin incluir GP y F2011
1 AMAZONAS	58.275	45.738	5.830	26.938	4.607	746	142.134	38.121
2 ANCASH	160.841	108.375	56.762	156.214	33.609	4.463	520.264	251.048
3 APURIMAC	76.198	38.892	10.054	17.152	3.657	1.968	147.921	32.831
4 AREQUIPA	348.717	83.172	188.885	63.553	38.784	5.378	728.489	296.600
5 AYACUCHO	137.663	57.462	12.820	20.774	6.017	2.441	237.177	42.052
6 CAJAMARCA	181.247	195.586	38.577	127.781	26.405	4.766	574.362	197.529
7 CALLAO	112.051	113.670	148.890	86.030	54.735	5.085	520.461	294.740
8 CUSCO	346.424	60.097	67.027	44.407	29.673	5.738	553.366	146.845
9 HUANCAVELICA	84.191	28.305	6.854	26.553	3.471	2.430	151.804	39.308
10 HUANUCO	123.278	59.652	25.199	57.072	11.897	2.507	279.605	96.675
11 ICA	130.689	116.872	66.107	58.795	47.556	1.704	421.723	174.162
12 JUNIN	213.959	149.561	90.010	69.449	31.847	4.155	558.981	195.461
13 LA LIBERTAD	201.029	233.152	117.648	149.451	108.012	7.774	817.066	382.885
14 LAMBAYEQUE	160.103	165.113	53.310	70.490	141.091	2.587	592.694	267.478
15 LIMA	1.134.353	1.213.824	1.434.317	825.075	693.205	34.092	5.334.866	2.986.689
16 LORETO	99.744	71.506	36.473	114.841	14.996	3.218	340.778	169.528
17 MADRE DE DIOS	26.132	7.970	4.859	9.614	2.156	250	50.981	16.879
18 MOQUEGUA	46.437	11.871	19.830	12.653	5.783	576	97.150	38.842
19 PASCO	34.693	35.148	14.361	22.075	6.804	876	113.957	44.116
20 PIURA	249.210	256.116	91.880	127.909	69.481	4.994	799.590	294.264
21 PUNO	364.235	90.749	51.566	42.442	23.775	8.084	580.851	125.867
22 SAN MARTIN	111.844	106.821	23.747	49.531	15.679	2.327	309.949	91.284
23 TACNA	102.307	18.248	32.600	13.044	11.688	969	178.856	58.301
24 TUMBES	29.500	37.623	9.456	22.961	5.924	449	105.913	38.790
25 UCAYALI	70.194	55.610	16.047	28.135	7.848	1.453	179.287	53.483
TOTAL	**4.603.314**	**3.361.133**	**2.623.109**	**2.242.939**	**1.398.700**	**109.030**	**14.338.225**	**6.373.778**

GP: Gana Perú, F2011: Fuerza 2011, APEGC: Alianza por el Gran Cambio, PP: Perú Posible, ASN: Alianza Solidaridad Nacional, Otros (Fonavistas del Perú, Despertar Nacional, Adelante, Fuerza Nacional, Justicia, Tecnología, Ecología y Partido Descentralista Fuerza Social.

Como Rodrigues y Lima (2009) han observado, existen tres posibles ACP: considerando la frecuencia de los datos electorales (el número de votos); transformando los datos electorales a proporciones; y transformando los datos

electorales de proporciones a log contraste centrados. En el primer caso no se reconoce la naturaleza composicional de los datos y corresponde al proceso común que podría considerarse. Nótese en este caso que la suma de votos cambia por cada provincia pero la distribución de votos es composicional entre las agrupaciones políticas (el número de votos de una determinada agrupación política modifica el número de votos de las otras). Este análisis es denominado aquí como un ACP directo. El ACP directo de datos composicionales se enfrenta a las dificultades generales antes mencionadas en combinación con el hecho de que este tipo de datos muestran a menudo considerable curvatura, lo cual no es compatible con el hipótesis de linealidad asumida por el ACP, y conduce a resultados poco satisfactorios como se indica en Rodríguez y Lima (2009).

En el segundo caso, la suma de las proporciones de las votaciones de todas las proporciones es la misma e igual a 1. Este análisis es denominado por Rodríguez y Lima como ACP crudo. ACP crudo es un ACP aplicado a datos donde las votaciones electorales son reemplazadas por distribución relativa (proporciones) del total en cada fila. Rodríguez y Lima (2009) indican, citando a Aitchison (1986), que se debe realizar los correspondientes diagramas de dispersión de los pares $\{(a_m^{'}p_r, a_n^{'}p_r) : r = 1, \ldots, n\}$, donde p_r es el vector con la proporción de votación electoral de la provincia r con el eje principal crudo $\{(a_m^{'}p, a_n^{'}p) : m, n = 1, \ldots, p, \forall m \neq n\}$. Nótese que en este caso la matriz varianza-covarianza $K = [cov(p_m, p_n)]$ de las proporciones de votos de cada par de agrupaciones m y n, tiene $p = P - 1$ valores propios mayores que cero, digamos $\lambda_1 > \ldots > \lambda_p > 0$ con correspondientes valores propios estandarizados $a_1, \ldots, a_p$.

Si en este diagrama de dispersión los puntos tienen una estructura elíptica homogénea, el ACP crudo es una buena solución. Sin embargo, si se observan patrones curvos, este método es inapropiado y produce resultados inapropiados.

En el tercer caso, el ACP es aplicado a los datos composicionales luego de realizar la transformación clr y es denominado por Rodrigues y Lima (2009) como el ACP de contraste. Como indican Rodrigues y Lima (2009) en esta alternativa se encuentra las funciones no lineales de los componentes y consecuentemente un mejor ajuste para los datos como es el caso de la transformación clr.

Para realizar los diferentes ACP nosotros consideramos diferentes paquetes del programa R, en particular el paquete compositions (van den Boogaart and Tolosana-Delgado 2008). Los resultados son mostrados a continuación:

4.1 ACP directo

Como primer resultado es mostrado la matriz de correlaciones entre las organizaciones participantes, como en el cuadro 2 abajo, en él se presentan las correlaciones según el número de votos, la proporción de votos y las transformaciones clr. Este cuadro permitirá comparar los resultados que se obtendrán al hacer el Análisis de Componentes Principales (ACP)

Cuadro 2. Matriz de correlaciones entre las organizaciones políticas participantes de las elecciones presidenciales peruanas de primera vuelta del 2011

a) Considerando los votos obtenidos

	GP	p	F2011	p	APEGC	p	PP	p	ASN	p	OTROS	p
GP	1.00000		0.91930	**	0.93199	**	0.90805	**	0.91077	**	0.97035	**
F2011	0.91930	**	1.00000		0.97069	**	0.98482	**	0.98203	**	0.96413	**
APEGC	0.93199	**	0.97069	**	1.00000		0.97102	**	0.97825	**	0.96685	**
PP	0.90805	**	0.98482	**	0.97102	**	1.00000		0.96997	**	0.96507	**
ASN	0.91077	**	0.98203	**	0.97825	**	0.96997	**	1.00000		0.95292	**
OTROS	0.97035	**	0.96413	**	0.96685	**	0.96507	**	0.95292	**	1.00000	

**: p<0.01 *:p <0.05: correlación significativa

b) Considerando las proporciones de votos obtenidos

	GP	p	F2011	p	APEGC	p	PP	p	ASN	p	OTROS	p
GP	1.00000		-0.61112	**	-0.29396		-0.56015	**	-0.57269	**	0.43508	*
F2011	-0.61112	**	1.00000		-0.40166	*	0.34186		0.14135		-0.23125	
APEGC	-0.29396		-0.40166	*	1.00000		-0.21476		0.35003		-0.24951	
PP	-0.56015	**	0.34186		-0.21476		1.00000		-0.09134		-0.07309	
ASN	-0.57269	**	0.14135		0.35003		-0.09134		1.00000		-0.42065	*
OTROS	0.43508	*	-0.23125		-0.24951		-0.07309		-0.42065	*	1.00000	

**: p<0.01 *: p<0.05: correlación significativa

c) Considerando las transformaciones clr de las proporciones de votos obtenidos

	GP	p	F2011	p	APEGC	p	PP	p	ASN	p	OTROS	p
GP	1.00000		-0.25668		-0.26678		-0.32887		-0.62369	**	0.52727	**
F2011	-0.25668		1.00000		-0.68197	**	0.45195	*	-0.12718		-0.13061	
APEGC	-0.26678		-0.68197	**	1.00000		-0.47832	*	0.48684	*	-0.37899	
PP	-0.32887		0.45195	*	-0.47832	*	1.00000		-0.26961		-0.12392	
ASN	-0.62369	**	-0.12718		0.48684	*	-0.26961		1.00000		-0.73104	**
OTROS	0.52727	**	-0.13061		-0.37899		-0.12392		-0.73104	**	1.00000	

**: p<0.01 *: p<0.05: correlación significativa

GP: Gana Perú, F2011: Fuerza 2011, APEGC: Alianza por el Gran Cambio, PP: Perú Posible, ASN: Alianza Solidaridad Nacional, Otros (Fonavistas del Perú, Despertar Nacional, Adelante, Fuerza Nacional, Justicia, Tecnología, Ecología y Partido Descentralista Fuerza Social.

El cuadro 3 muestra las cargas y las proporciones acumuladas de la varianza explicada de los 6 componentes principales.

De aquí, las cargas en el Cuadro 3 no tienen las mismas coordenadas de la Figura 3 ya que ellas tienden a dar solamente la dirección de cada carga componente.

Aunque los primeros dos componentes principales explican el 98.6 % de la varianza original de los datos (cuadro 2), la estructura que se muestra en la Figura 3 no es clara desde que los departamentos provincias se superponen y las cargas tienen al menos la misma dirección. Este hecho no ayuda a comprender la estructura de los datos y análisis adicionales se deben realizar.

Cuadro 3. Cargas de los Componentes para un ACP directo

Variables	Cargas de los Componentes					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
GP	-0.398	0.777	-0.107	0.224	0.347	0.236
F2011	-0.411	-0.28	0.234	0.497	0.312	-0.595
APEGC	-0.411	-0.122	-0.503	-0.653	0.252	-0.27
PP	-0.409	-0.318	0.563	-0.261	0.213	0.548
ASN	-0.409	-0.334	-0.511	0.407	-0.395	0.372
OTROS	-0.411	0.3	0.32	-0.206	-0.719	-0.281
Desviación Estándar	2.40474	0.36413	0.19474	0.15979	0.11197	0.09284
Proporción de la Varianza	0.96380	0.02210	0.00632	0.00426	0.00209	0.00144
Proporción Acumulada	0.96380	0.98590	0.99222	0.99647	0.99856	1.00000

El biplot mostrado en la Figura 3 muestra las cargas de los componentes y los escores de los dos primeros componentes principales de los datos fila, el número de votos. Las flechas asociadas con los cargas componentes en el biplot tienen la dirección dada por las primeras dos columnas en el cuadro 2 y su longitud es proporcional a la importancia relativa de la correspondiente organización electoral en el modelo de ACP.

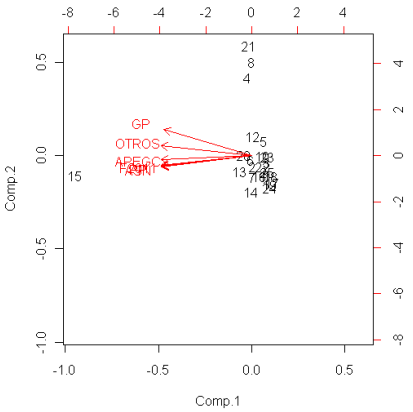

Figura 3. Biplots para los dos primeros componentes principales de un ACP directo para los votos por organizaciones políticas en la elección presidencial peruana de 1ra vuelta a nivel de 24 departamentos y 1 provincia constitucional

Adicionalmente, como se muestra en la figura 4, y se corrobora con los resultados en el cuadro 2, las correlaciones entre las votaciones de las organizaciones políticas son altas y significativas, evidenciando aparentemente una alta correlación entre las votaciones lo que no permite distinguir las preferencias electorales entre las organizaciones políticas.

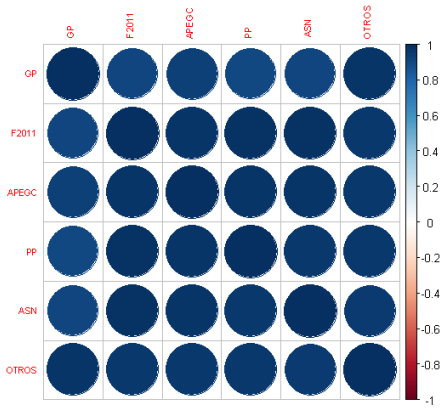

Figura 4. Correlaciones entre los votos por organizaciones políticas en la elección presidencial peruana de 1ra vuelta a nivel de 24 departamentos y 1 provincia constitucional

4.2 ACP crudo

El cuadro 4 muestra las cargas de los componentes y las proporciones acumuladas de la varianza explicada de los 6 componentes principales crudos.

Cuadro 4. Cargas de los Componentes para un ACP crudo

Variables	Cargas de los Componentes					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
GP	0.613		0.224	-0.211	0.184	-0.703
F2011	-0.403	0.448	0.409	0.161	-0.512	-0.426
APEGC	-0.141	-0.631	-0.453		-0.49	-0.369
PP	-0.308	0.452	-0.646	-0.169	0.361	-0.352
ASN	-0.42	-0.379	0.269	0.469	0.57	-0.252
OTROS	0.414	0.227	-0.296	0.825		
Desviación Estándar	1.56091	1.31251	0.91171	0.78344	0.62920	0.00000
Proporción de la Varianza	0.40607	0.28712	0.13854	0.10230	0.06598	0.00000
Proporción Acumulada	0.40607	0.69319	0.83172	0.93402	1.00000	1.00000

La Figura 5 muestra el biplot para los dos primeros componentes principales crudos. Aunque las proporciones explicadas de varianza obtenidas por los dos primeros componentes principales crudos son menores que los obtenidos usando el

ACP directo (alrededor de 69.3 % para los dos primeros y 83.2 % para los tres primeros componentes principales), el biplot de la Fig. 5 es más auto explicativo. De hecho podemos observar que las organizaciones políticas se ubican en tres grupos. Los partidos F2011 y PP en el lado izquierdo superior, los partidos ASN y APEGC en el lado izquierdo inferior, y el partido GP en el lado derecho junto a OTROS.

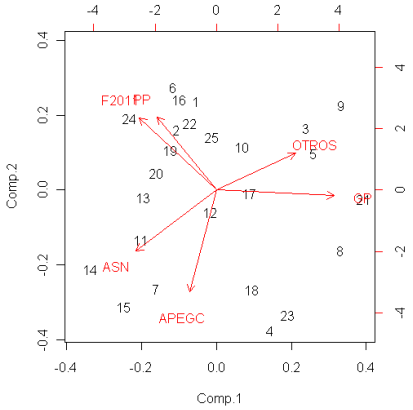

Figura 5. Biplots para los dos primeros componentes principales de un ACP crudo para las proporciones de votos por organizaciones políticas en la elección presidencial peruana de 1ra vuelta a nivel de 24 departamentos y 1 provincia constitucional

En el lado superior izquierdo de la Fig. 5 tenemos, principalmente los departamentos de Cajamarca (6), Loreto (16), Amazonas (1), San Martin (22), Ancash (2), Pasco (19), Tumbes (24) y Piura (20), la mayoría de los cuales que pueden considerarse como departamentos de nivel intermedio de de desarrollo humano[5]. En el lado inferior izquierdo tenemos los departamentos de Ica (11), Lambayeque (14), Lima (15) y Callao (7) que pueden ser considerados como mayores o intermedios niveles de desarrollo humano. Finalmente en el lado derecho tenemos los departamentos de Apurimac (3), Ayacucho (5), Cusco (8), Huancavelica (9) y Puno (21), que forman parte del grupo de departamentos con menores niveles de desarrollo humano.

Si analizamos estos datos composicionales podemos decir que Gana Perú se asocia a los departamentos menos desarrollados, ASN y APEGC con los departamentos de mayores ingresos en tanto que F2011 y PP compiten por los departamentos de ingresos intermedios.

Estas interpretaciones parecen corresponder con las verdaderas preferencias por las organizaciones políticas en los diferentes departamentos. Sin embargo, en el primer componente principal se aprecia una gran influencia del grupo

[5] Véase el Cuadro 5 del índice de desarrollo humano del PNUD en los anexos.

"OTROS" que incluye al resto de organizaciones políticas, pese a ser un grupo minoritario en su proporción de votos. Es posible mejorar los resultados usando un método que remueve las relaciones no lineales entre las variables (organizaciones políticas) como se encuentra usando el ACP de Logcontrast

Nótese que en este caso, como se muestra en la figura 6, y se corrobora con los resultados en el cuadro 2, las correlaciones entre las proporciones de votos de las organizaciones políticas son desiguales observándose algunas significativas y otras no, evidenciando aparentemente una mejor diferenciación entre las proporciones de votos lo que nos permite distinguir las preferencias electorales entre las organizaciones políticas.

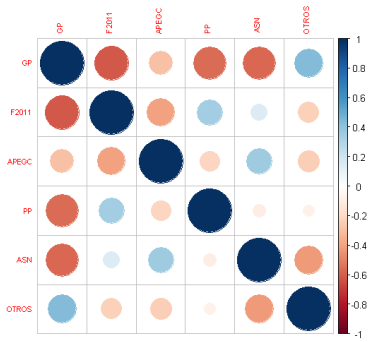

Figura 6. Correlaciones entre las proporciones de votos por organizaciones políticas en la elección presidencial peruana de 1ra vuelta a nivel de 24 departamentos y 1 provincia constitucional

4.3 ACP de Logcontraste

El cuadro 5 muestra las cargas de los componentes y las proporciones acumuladas de la varianza explicada de los 6 componentes principales de log contraste.

Cuadro 5. Cargas de los Componentes para un ACP de log contraste

	Cargas de los Componentes					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
GP	0.385	-0.434	-0.258	0.643		0.421
F2011	0.206	0.564	-0.507	-0.121	0.49	0.358
APEGC	-0.49	-0.329	0.325		0.588	0.447
PP	0.184	0.528	0.673	0.255	-0.191	0.364
ASN	-0.562	0.113	-0.318		-0.587	0.466
OTROS	0.469	-0.306	0.128	-0.706	-0.164	0.38
Desviación Estándar	1.61166	1.43693	0.81175	0.66045	0.49259	0.00000
Proporción de la Varianza	0.43291	0.34413	0.10982	0.07270	0.04044	0.00000
Proporción Acumulada	0.43291	0.77704	0.88686	0.95956	1.00000	1.00000

La figura 7 muestra el biplot para los dos primeros componentes principales de log contraste. Los dos principales componentes explican el 77.7 % de la varianza

total y los primeros tres 88.7 %. Analizando la Fig. 7 podemos observar que la organización política ASN se asocia fuertemente con los departamentos de Lambayeque (14), Lima (15), Ica (11), Callao (7). La organización política APEG con Callao (7) y Moquegua (18), la organización política GP con Puno (21) mientras que las otras organizaciones políticas con Ayacucho (3) y Apurimac (3). Finalmente puede verse que las organizaciones políticas F2011 y PP se asocian con Cajamarca (6), Ucayali (25), Loreto (16), San Martin (22), Amazonas (1).

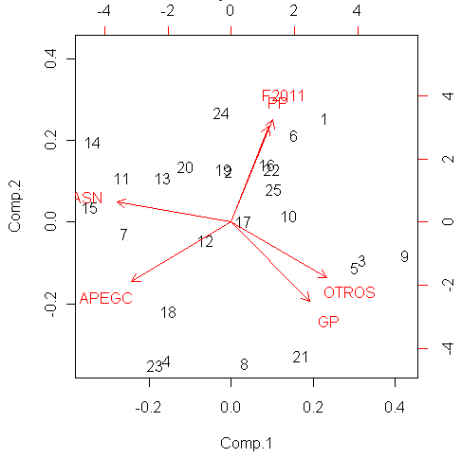

Figura 7. Biplots para los dos primeros componentes principales de un ACP de log contraste para los votos transformados por organizaciones políticas en la elección presidencial peruana de 1ra vuelta a nivel de 24 departamentos y 1 provincia constitucional

Adicionalmente puede observarse que, contrariamente a lo que se indicó, puede decirse que los partidos F2011 y PP se mostraron competitivos entre sí ocupando el mismo espectro el cual fue claramente diferente del especto ocupado por ASN y APEG y GP. Nótese también que GP se encuentra en el lado inverso de las organizaciones ASN y APEG, mientras que F2011 y PP ocupan un lado más central.

En este caso, como se muestra en la figura 8, y se corrobora con los resultados en el cuadro 2, las correlaciones entre las transformaciones de las proporciones de votos de las organizaciones políticas son desiguales observándose algunas significativas y otras no, evidenciando realmente una mejor diferenciación entre las transformaciones de las proporciones de votos lo que nos permite distinguir las preferencias electorales entre las organizaciones políticas.

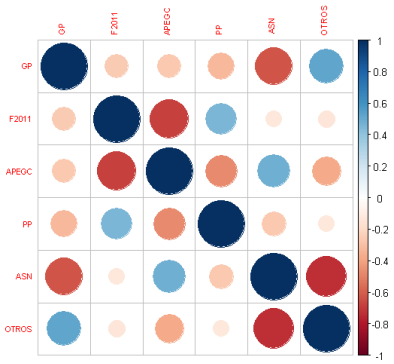

Figura 8. Correlaciones entre las transformaciones de las proporciones de votos por organizaciones políticas en la elección presidencial peruana de 1ra vuelta a nivel de 24 departamentos y 1 provincia constitucional

5. DISCUSIÓN

La naturaleza de los datos electorales es esencialmente composicional desde que la proporción de votos es de importancia primaria, antes que el número de votos. El análisis de datos composicionales es una técnica que fue usada para estudiar los datos de la elección presidencial peruana de la primera vuelta en el año 2011 para cada uno de los 24 departamentos y una provincia constitucional del país.

Debido a que el ACP directo asume relaciones entre las variables, este es inadecuado para datos composicionales y las conclusiones que resulten son deficientes. A diferencia de ello ACP de log contraste basado en datos transformados vía la transformación clr presenta una clara y más útil interpretación de los datos cuando se compara con ACP crudo debido a la no linealidad de las relaciones entre las variables removida por la transformación clr.

El ACP crudo parece resaltar la influencia de factores de índole socioeconómico (el nivel de desarrollo humano relativo de las circunscripciones electorales) en los patrones de la distribución del voto entre las diferentes organizaciones. Sin embargo, el ACP de log contraste, que respeta mejor la estructura composicional de los datos, podría hacer más visibles posibles relaciones de tipo geográfico, algo así como la influencia de “regiones electorales” en los patrones con los patrones de votación, que podrían combinar elementos sociodemográficos (niveles de desarrollo relativo del de la costa versus la sierra o selva) con elementos socioculturales (culturas o historias políticas del norte, del centro o del sur). En todo caso, ello da pistas de variables explicativas que podrían explicar de alguna manera los componentes principales despejados en el análisis realizado.

El presente trabajo es ilustrativo para comprender el comportamiento partidario a nivel de circunscripciones electorales grandes considerando un análisis

conveniente que respeta la naturaleza de los datos de votaciones partidarias. Un análisis más fino puede ser realizado a nivel de provincias o incluso distritos.

REFERENCIAS

AITCHISON, J. y Shen, S. (1980). Logistic-Normal Distributions: Some Properties and Uses. *Biometrika*, Vol. 67, No. 2. pp. 261-272.

AITCHISON J. (1983) Principal component analysis of compositional data. *Biometrika* 70:57-61

AITCHISON J. (1986) The statistical analysis of compositional data London. Springer.

BARCELÓ-VIDAL, C. (2003). When a data sol can be considered compositional? In: CoDaWork03: Compositional Data Analysis Workshop. Girona. Spain, Available at http://ima,udg,es/Activitats/CoDaWorkO.V

BRADU, D. y GABRIEL, K. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics* 20:47-68

BUTLER, A y GTASBEY, C. (2008). A latent Gaussian model for compositional data with Zeros. *Journal of the Royal Statistical Society. Series C, Applied statistics* 57:505-520

CHAYES, F. y TROCHIMCZYK, J. (1978). An effect of closure on the structure of principal component. *Journal of Mathematica Geology* 10:323 – 333

GABRIEL KR (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58:453–467

JOLLIFFE, I. (2002). *Principal component analysis*. Springer. New York

KATZ, JONATHAN, and GARY KING. A Statistical Model for Multiparty Electoral Data. *American Political Science Review* 93 (1999): 15-32

KUCERA, M. y MALMGREN, B. (1998). Logratio transformation of compositional data a resolution of the constant sutn constraint. *Marine Micropaleontology* 34:1 17-120

MOSIMANN, J. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49 (1-2): 65-82.

PALAREA-ALBALADEJO, J., MARTÍN-FERNÁNDEZ, J.A., and GÓMEZ-GARCÍA, J. (2007). Modelización y Análisis de Datos sobre Proporciones, In: Proceedings of XXI Reunión Anual de la Asociación de Economía Aplicada (ASEPELT'07), Valladolid, Spain, June 20-23, p. 259-276. Disponible en http://ima.udg.edu/~jamf/P_M_G_asepelt07.pdf

PALAREA-ALBALADEJO, J. y MARTÍN-FENÁNDEZ, J. (2008). A modified EM alr - algorithm for replacing rounded zeros in compositional data sets. *Computational Geosciences* 34:902-917

RODRIGUES, P. y LIMA, A. (2009). Analysis of an European union election using principal component analysis. Statistical Papers (2009) 50; 895-904. Springer – Verlag.

THI-HENESTROSA, S. y MARTÍN-FERNÁNDEZ, J. (2005) Dealing with compositional data: the freeware coDaPack. En *Marine Micropaleontology* 7:773-793

VAN DEN BOOGAART, K.G. and TOLOSANA-DELGADO, R. (2008) "compositions": a unified R package to analyze Compositional Data, *Computers & Geosciences*. 34 (4), 320-338

WINZER, N. (1999). Efectos de la transformación de log-cocientes centrados en datos composicionales. *Revista de la Sociedad Argentina de Estadística*, 3 (1-2), 114-120. Disponible en http://www.s-a-e.org.ar/Vol%203/Articulo8.pdf