

*Curso de postgrado*  
Manejo estadístico de datos composicionales

Profesora: Dra. Vera Pawlowsky-Glahn  
*Universitat de Girona*  
*Catalunya*

# **Análisis exploratorio de la composición del stock de capital *per capita* en una muestra de países**

Por Juan Manuel Larrosa  
*jarrosa@{criba.edu.ar; yahoo.com}*



Universidad Nacional del Sur  
Bahía Blanca  
Argentina

Diciembre  
2000

## 1. Introducción

Este ensayo busca determinar las posibles relaciones presentes en los componentes del stock de capital per capita de una muestra de países. En un principio no se cuenta con una teoría que describa exactamente cómo estas relaciones debieran ser. Pueden conjeturarse algunas relaciones que serán expuestas más adelante, pero en general el análisis a desarrollar es básicamente exploratorio. Un primer paso, en ese sentido, será el de reducir la dimensionalidad de los datos para percibir algún patrón en la estructura. A partir de ello se realizarán las reformulaciones necesarias para intentar esbozar conclusiones preliminares. En ese sentido se estima conveniente utilizar métodos estadísticos para datos multivariados, como son los datos composicionales, a fin de ir capturando estas interrelaciones.

El trabajo continúa con la sección 2, de definiciones iniciales. La sección 3 describe la fuente y los datos a utilizar. La sección 4 inicia con el análisis de datos y la sección 5 analiza los resultados de la sección precedente con la exposición de algunas conclusiones. Por último, en la sección 5 se esboza un resumen del trabajo realizado.

## 2. Definiciones iniciales

El stock de capital físico de una economía puede ser definido como la suma de todos los activos físicos de la misma, valuados a precios constantes. El conjunto de activos físicos responde a necesidades específicas que debe cumplir el capital en la economía (por ejemplo, infraestructura de caminos, maquinaria de producción, construcción residencia, entre otras), por lo que tienen asociados formas de valoración diferentes. De este modo, podemos definir a  $A_t^i$  como activo físico correspondiente a una categoría de capital  $i$  de un total de  $N$  categorías el período  $t$ , a  $p_t^i$  como el valor del capital asociado a cada categoría específica en el período  $t$  y a  $K_t$  como el stock de capital en el período  $t$ , y obtenemos

$$K_t = \sum_{i=1}^N p_t^i A_t^i$$

Para comprender la relatividad del stock de capital de cada economía se suele normalizar sus resultados basándose en el número de la fuerza laboral correspondiente. De ese modo se obtiene el stock de capital en términos *per capita*, el cual refleja la cantidad de capital por trabajador que posee cada economía para su proceso de producción y asignación. Siendo  $L_t$  el número de trabajadores de la fuerza laboral en el periodo  $t$ , el stock de capital per capita  $k_t$  en el periodo  $t$  se puede definir como

$$k_t = \frac{K_t}{L_t}$$

¿Qué relación existe entre las composiciones ( $A_t^i$ ) de stock de capital? No existen muchos aportes teóricas para responder esta pregunta. En primer lugar, aquellas economías altamente

industrializadas debieran destinar una significativa proporción de su capital a la construcción de inmuebles industriales en general. Sin embargo, a mayor infraestructura productiva se requiere de mayor infraestructura de apoyo (camino y vías de comunicación, construcción residencial así como infraestructura de múltiples propósitos). Los países menos industrializados debieran poseer su mayor stock de capital asignado a viviendas e infraestructura de transporte, dado que usualmente producen materias primas (sin procesar) para la exportación. En medio de ambos extremos, la distribución del uso del capital puede corresponder a cuestiones más particulares de cada economía, más difíciles de llevar al plano de la generalización. Sin embargo, el propio carácter composicional de la variable a analizar hace que esta deducción implique contemplar conceptos diferentes a los de una variable no composicional común.

En este ensayo se pretende observar patrones de asignación del capital en una muestra de economías. Para ello se utilizarán datos composicionales del stock de capital per cápita de dichos países. El análisis es exploratorio y para el mismo se empleará las técnicas de manejo de datos composicionales aportados por Aitchinson (1986). Las conclusiones al mismo son preliminares.

### 3. Los datos utilizados

La información utilizada para el análisis fue extraída de la *Penn World Table 5.6*, una base de datos desarrollada en el *Center for International Comparisons at the University of Pennsylvania* (<http://ssc.upenn.edu>). Actualmente tiene registrado datos anuales de 29 variables económicas de 152 países desde 1950 hasta 1992. Ello representa un total de algo más de 116.000 datos, dado que algunas series no están completas. Para el presente trabajo se ha propuesto el estudio de la composición del stock de capital per cápita de 42 economías para el período 1965 a 1990. Para ello se utilizó una composición de 5 variables, las cuales se describen en la Tabla a continuación:

**Tabla 1. Variables composicionales. Codificación y descripción**

<b>Código</b>	<b>Descripción</b>
<b>KRES</b>	% del stock de capital per cápita asignado a construcción residencial.
<b>KNRES</b>	% del stock de capital per cápita asignado a construcción no residencial.
<b>KDUR</b>	% del stock de capital per cápita asignado construcción de activos durables de producción (maquinaria y equipos).
<b>KOTHR</b>	% del stock de capital per cápita asignado a otras construcciones.
<b>KTRAN</b>	% del stock de capital per cápita asignado a equipos de transporte.

La suma de los porcentajes corresponde al total del stock de capital per cápita de cada economía. Los países que fueron seleccionados son aquellos que poseen series completas para las variables y el período analizado. La lista de economías se enumera a continuación (con el código entre paréntesis):

Argentina (ARG), Australia (AUS), Austria (ÖST), Bélgica (BEL), Canadá (CAN), Chile (CHI), Colombia (COL), Dinamarca (DEN), República Dominicana (DOM), Ecuador (ECU), Finlandia (FIN), Francia (FRA), Alemania Federal (GER), Grecia (GRE), Honduras (HON), Hong Kong (HONG), Islandia (ICE), India (IND), Irlanda (IRE), Israel (ISR), Italia (ITA), Costa de Marfil (IVO), Jamaica (JAM), Japón (JAP), República de Corea (KOR), Luxemburgo (LUX), Islas Mauricio (MAU), Países Bajos (NET), Nueva Zelanda (NEW), Noruega (NOR), Panamá (PAN), Portugal (POR), Sierra Leona (SIE), Suiza (SWI), Siria (SYR), Taiwán (TAI), Tailandia (THAI), Turquía (TUR), Reino Unido (U.K.), Estados Unidos (U.S.A.), Venezuela (VEN) y Yugoslavia (YUG). Los datos fueron procesados con una planilla de cálculo de Excel diseñada por Ricardo Camina y también se utilizó intensamente el XLSTAT 4.0 para el procesamiento estadístico multivariado.

El criterio para seleccionar las variables privilegió, en primer lugar, la existencia de series completas para cada una de las variables composicionales. En segundo lugar, se optó por descartar aquellos países que poseyeran alguna de sus series con valor cero. Si bien existen técnicas sugeridas para lidiar con este problema [Martín-Fernández y otros, 2000] se decidió no emplearlos dado el carácter exploratorio e inicial del análisis.

Las variables se presentaban originalmente como porcentajes respecto al stock de capital total per capita de la economía de Estados Unidos referenciado a un año base 1980 y ello hacía que la suma total de componentes no coincidiera con el cien por ciento. Se procedió entonces a delimitar la composición, y clausurarla cada vector composicional de año en año. Así se obtuvo, para cada año, la participación de cada vector composicional en el cien por cien de explicación del stock de capital per capita de cada economía. Luego se obtuvo el promedio geométrico de cada vector para todo el periodo de análisis. Se clausuró nuevamente, dado que la media geométrica de las variables era menor al cien por cien, para así obtener la participación promedio de cada variable en la composición durante todo el periodo. Con ello se obtuvo para cada país cinco valores, uno por cada vector composicional, promedio de todo el período. La razón de este procedimiento reside en que de este modo toman valores representativos de largo plazo de cada economía, los cuales están ajenos a distorsiones producidas por posibles fluctuaciones a corto plazo.

### 3.1 Análisis de los datos

Para el procesamiento estadístico a los cinco vectores se les aplicó la transformación *clr* (*centered logratio transformation*) para luego realizar un análisis de componentes principales (ACP) a fin de detectar patrones dentro de la estructura de datos. Estos patrones brindarán información preliminar sobre las relaciones entre vectores composicionales y subcomposiciones para su visualización en el espacio simplex. Ello dada la carencia al momento de redacción de este trabajo de información previa cualificada sobre las posibles relaciones a testear.

La transformación *clr* comprende la operación es una transformación uno a uno de  $x \in S_c^5$  a  $z \in \mathbb{R}^5$  definida por

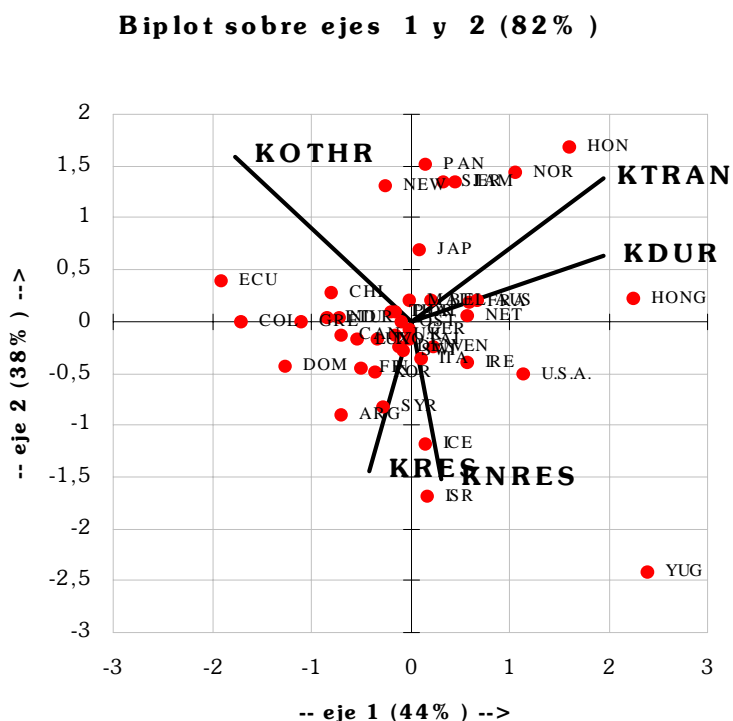
$$\text{clr}(x) = \left( \ln \frac{x_1}{g(x)}, \ln \frac{x_2}{g(x)}, \ln \frac{x_3}{g(x)}, \ln \frac{x_4}{g(x)}, \ln \frac{x_5}{g(x)} \right) = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5) = \varepsilon$$

con  $g(x) = (x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5)^{1/5}$ , es decir, la media geométrica de la muestra. Dado que no afecta la isometría de los datos esta transformación es la más útil para realizar análisis estadístico multivariado.

El ACP fue llevado a cabo tomando en cuenta la covarianza entre las variables. Este procedimiento al reducir la dimensionalidad de los datos permitió visualizar posibles subcomposiciones factibles de análisis dentro del simplex. Los resultados se publican en el Gráfico 1.

Del ACP inicial se podría establecer la existencia de tres subcomposiciones dentro de las cinco variables que mantienen una baja correlación entre sí. Por un lado, KRES y KNRES, las dos variables que identifican a la construcción de infraestructura residencial y no residencial, muestran una correlación del mismo orden y magnitud con respecto al resto de los vectores. De la misma manera se agrupan KTRAN y KDUR, las dos formas de capital basada en infraestructura de transporte y equipos y maquinarias de producción, manteniendo una dirección distante de los dos primeros el capital destinado a otro tipo de construcción, KOTHR. Es decir, podría formarse a partir de esta información tres subcomposiciones:  $X1 = \text{KNRES} + \text{KRES}$ ;  $X2 = \text{KTRAN} + \text{KDUR}$ ;  $X3 = \text{KOTHR}$  para observar más claramente la existencia de estas relaciones.

**Gráfico 1. Biplot de 5 vectores composicionales**



Asimismo, de la misma observación del Gráfico 1, puede percibirse una cierta recta

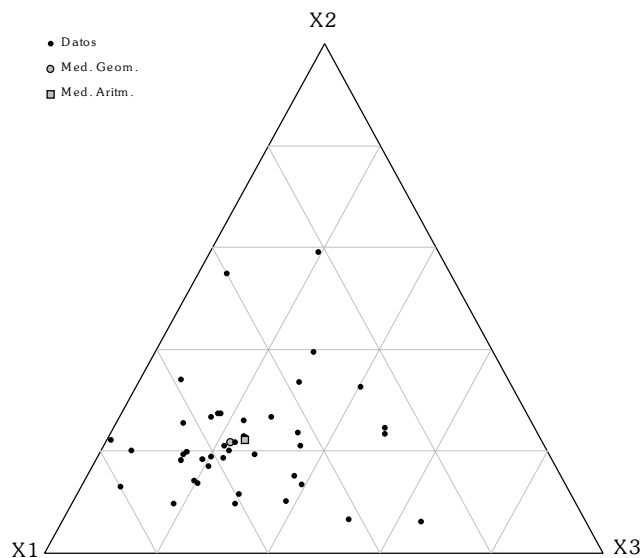
composicional entre KTRAN, KDUR y KRES si bien, por otro lado, esta forma dista de ser clara y concluyente. Entonces, basados en esta información preliminar se decidió realizar un análisis en el simplex de las subcomposiciones X1, X2 y X3 y luego investigar si existe efectivamente una relación entre los vectores KTRAN, KDUR y KRES.

### 3.1.1 Análisis de subcomposiciones

Para cada subcomposición se utilizó la suma de los vectores composicionales que se observan en la descripción anterior. Los subcomposiciones fueron clausuradas y sus resultados expuestos en el simplex, como se observa a continuación en el Gráfico 2.

Se observa una alta dispersión entre los valores antes y después de centrado de los mismos, tal como sugería el ACP, dado que las subcomposiciones mantenían un alto grado de falta de correlación. A partir de estos resultados se intentó hacer un análisis de conglomerados (*cluster*) para delimitar la posible existencia de dos ó más grupos homogéneos dentro de la muestra. Utilizando el método de Ward, de manera análoga al estudiado por Martín-Fernández y otros (1998) y al empleado por Pawlowsky-Glahn y Buccianti (2000) para un análisis geológico, se detectaron 4 grupos homogéneos en términos de los vectores composicionales  $(x_1, \dots, x_5)$ . Los resultados de los agrupamientos se observan en el Anexo 1 al final de este trabajo.

**Gráfico 2. Datos de las subcomposiciones X1, X2 y X3 en el diagrama ternario**



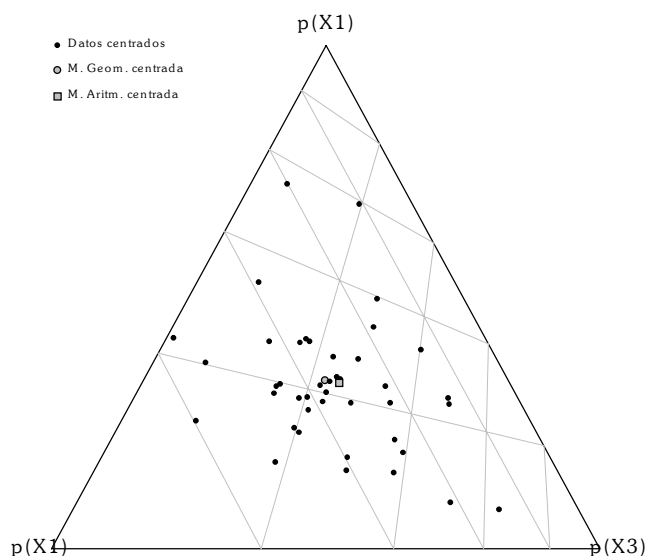
El Grupo 1 obtuvo 27 países, un número de observaciones considerado suficiente como para aislarlo y trabajar sobre él. Por ello se decidió proseguir el estudio de estas economías que mantenían un grado de dispersión menor. La homogeneidad del Grupo 1 puede sostenerse, a parte de la evidencia estadística, por ciertas características a las que se hará mención a continuación.

1) El 60% por cierto de los países de la muestra poseen de niveles de mediana a alta industrialización.

2) El 56% de los países de la muestra son europeos. De los pertenecientes al continente asiático (22%), sólo Siria (SYR) no es un país industrializado ó en vías de industrialización.

3) Las economías pertenecientes al continente americano, casi 15%, incluyen a los países más industrializados del continente, Canadá (CAN) y Estados Unidos (U.S.A.)

### Gráfico 3. Diagrama ternario de los valores de las subcomposiciones centradas

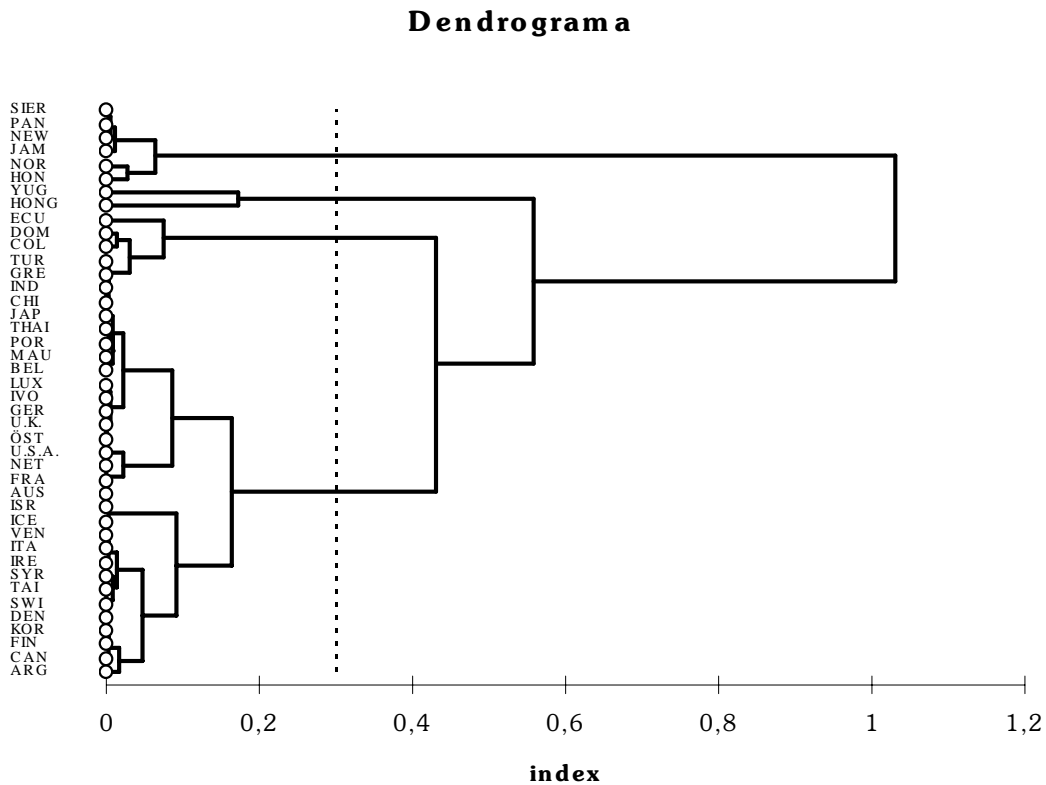


De esta breve descripción queda expuesto que el agrupamiento seleccionado corresponde a uno con muy fuerte preeminencia de países industrializados, por lo que las conclusiones posteriores deberán asociarse a esta caracterización de muestra. Al mismo tiempo no debe olvidarse el tema de la calidad de los datos. Una de las razones de este sesgo puede residir en que estos países poseen series estadísticas de su desempeño económico de mayor calidad que las de los países menos desarrollados. Luego se recalcularon el ACP para las mismas. Este análisis se describe a continuación.

#### 3.1.2 Análisis dentro del Grupo 1

Nuevamente se repetirá el análisis realizado para el caso de 2.1 aunque ahora concentrado a los datos del Grupo 1. Éstos, siempre transformados por  $\text{clr}$ , fueron reducidos por ACP y se presenta el biplot de los dos primeros componentes en el Gráfico 5. Se vislumbra claramente dos rectas composicionales, al contrario de la falta de claridad a este respecto observado en el Gráfico 1. Aquí puede existir una relación lineal entre KTRAN, KDUR y KNRES así como, por otro lado, entre KOTH y KRES.

**Gráfico 4. Dendrograma del análisis de cluster por el método de Ward**

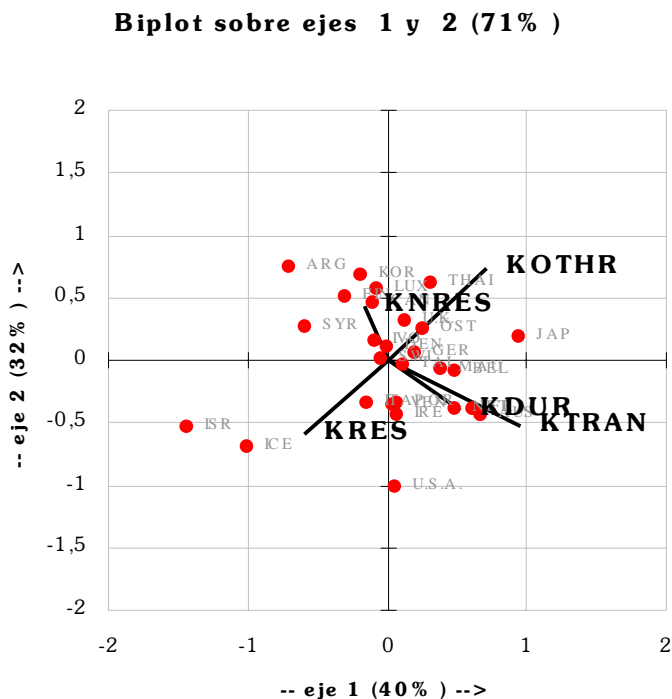


Los datos de la recta composicional KTRAN (X1), KDUR (X2) y KNRES (X3) en el diagrama ternario se aprecian en los Gráficos 6 (datos no centrados) y 7 (datos centrados). Se observa claramente la baja dispersión relativa en comparación con los Gráficos 2 y 3, al trabajar con un grupo de datos más homogéneos. Los datos del Gráfico 6 muestran una relación más perceptible de agrupamiento en una recta composicional que en otros diagramas analizados previamente (y en muchos más no publicados en este informe). Asimismo, se observa claramente la preeminencia del porcentaje destinado a XNRES seguido de KDUR. Una vez perturbados por la inversa de la media geométrica clausurada de la composición (centrados, en el Gráfico 7) se puede notar una recta composicional difusa y con destacada variabilidad por encima del baricentro. De todos modos, convenga la insistencia, persisten valores de las variables dispersos.

Debe notarse también que sería deseable que el porcentaje de variabilidad explicado por estos componentes fuese mayor. Asimismo, la información del Gráfico 5 nos permitiría realizar una clasificación dentro de este grupo de países de la tendencia de su asignaciones de capital durante el período bajo estudio.



**Gráfico 5. Biplot del primer y segundo componente principal del Grupo 1**



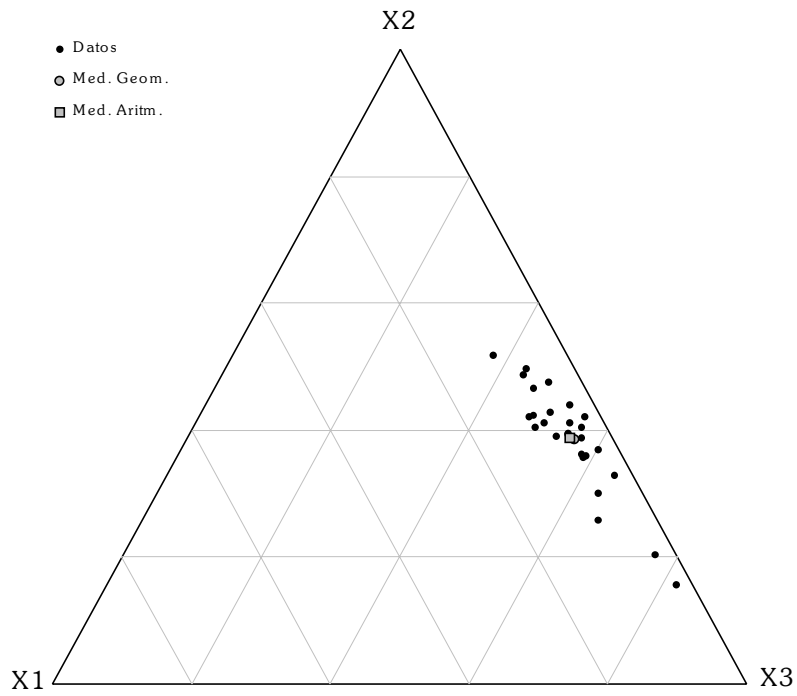
#### 4. Análisis de resultados

En el conglomerado de naciones, con fuerte sesgo industrial, considerado en el análisis la relación entre el porcentaje del capital total destinado a equipos durables de producción y a equipos de transporte mantienen el mismo signo y sentido, es decir son colineales, mientras que el porcentaje de capital destinado a construcción no residencial es colineal a estas dos pero de sentido opuesto.

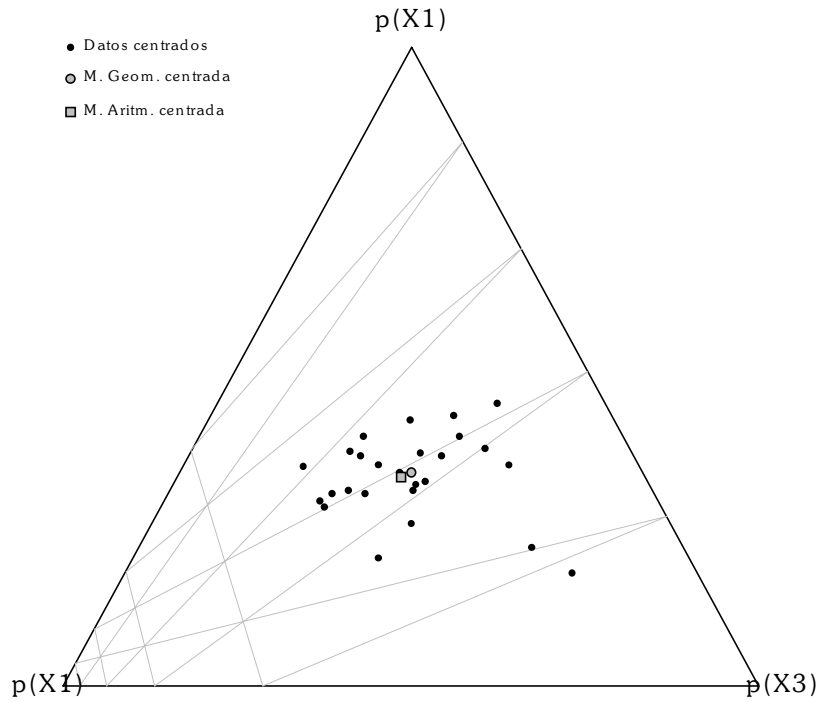
Al mismo tiempo, el porcentaje de capital per capita destinado a construcción residencial y la participación del capital destinado a otras construcciones son colineales pero de sentido opuesto entre ambas. Éstas dos son ortogonales a los tres porcentajes de capital descritos en primer término.

En términos de teoría económica puede observarse aquí un costo de oportunidad del capital dado por su uso alternativo dentro de una misma funcionalidad. Existe una oposición entre asignaciones de capital cuando se destinan a fines productivos. Por ejemplo, cuando existe una asignación marcada de capital hacia construcción no residencial (fábricas, por ejemplo) aparece una disminución en el porcentaje de capital destinado a maquinarias y elementos de transporte (dentro de la misma funcionalidad del capital, es decir destinado a la producción).

**Gráfico 6. Subcomposición KTRAN, KDUR, KNRES en el diagrama ternario**



**Gráfico 7. Subcomposición KDUR, KTRAN, KNRES centrada en el diagrama ternario**



Lo mismo ocurre con la construcción residencial y la destinada a otros fines (por ejemplo, construcción de espacios recreativos), las cuales no comparten la función de producción. Cuando existe un gran destino de fondos hacia uno de estos propósitos, al parecer, se hace en detrimento de otro destino dentro de la misma funcionalidad. Por último, sea que se forme capital para usos destinados a la producción ó destinado para otros usos distintos de la producción, no se valoran ambos destinos como alternativos ó competidores. Es decir ninguno de las dos asignaciones compite por fondos destinados a la otra (ello por la ortogonalidad mencionada anteriormente).

## **5. Resumen y conclusiones preliminares**

El análisis se inició con la aplicación de ACP al conjunto de vectores composicionales transformados por  $\text{clr}$ . La identificación de tres posibles subcomposiciones permitió presentar la información en el diagrama ternario. La presencia de fuerte variabilidad en los datos dió soporte a un análisis de cluster usando el método de Ward con distancias de Aitchinson para identificar posibles grupos homogéneos internos a esta estructura de datos. La obtención de una subpoblacional suficientemente grande permitió disminuir la dispersión de los datos, a los cuales nuevamente se les aplicó ACP.

En este último procedimiento se identificaron, a través del biplot, posibles rectas composicionales entre tres vectores. Observados en el diagrama ternario, se confirmó la menor variabilidad de los datos así como una posible relación lineal entre éstos. Las deducciones expuestas en la sección 4 de análisis de resultados reflejan la identificación de relaciones y posibles respuestas a las mismas.

## **Referencias**

- AITCHINSON, J. (1986), THE STATISTICAL ANALYSIS OF COMPOSITIONAL DATA, Chapman & Hall, London.
- MARTÍN-FERNÁNDEZ, J.A., C. BARCELÓ VIDAL y Vera PAWLOWSKY-GLAHN (1998), "A Critical Approach to Non-Parametric Classification of Compositional Data", Proceedings IFCS.
- 
- (2000), "Zero Replacement in Compositional Data Sets", Proceedings IFCS.

**Anexo 1. Agrupamiento de los datos transformados por clr por método de Ward**

Clustered observ.	/ Clusters size :		
Grupo 1	Grupo 2	Grupo 3	Grupo 4
27	7	6	2
ARG	CHI	HON	HONG
CAN	IND	NOR	YUG
FIN	GRE	JAM	
KOR	TUR	NEW	
DEN	COL	PAN	
SWI	DOM	SIER	
TAI	ECU		
SYR			
IRE			
ITA			
VEN			
ICE			
ISR			
AUS			
FRA			
NET			
U.S.A.			
ÖST			
U.K.			
GER			
IVO			
LUX			
BEL			
MAU			
POR			
THAI			
JAP			