CIECS CONICET U N C

ASTROLABIO ISSN 1668-7515
Número 12, 2014. NUEVA ÉPOCA

CONSTRUCCIÓN DE ESTRATOS SOCIOECONÓMICOS EN PARAGUAY: APLICACIÓN CON ANÁLISIS DE COMPONENTES PRINCIPALES EN R.

CONSTRUCTING SOCIO-ECONOMIC STRATA IN PARAGUAY: IMPLEMENTATION WITH PRINCIPAL COMPONENTS ANALYSIS IN R

Pablo Sebastián Gómez
CIECS (CONICET Y UNC)
enclavepablo@yahoo.com.ar

Resumen

Este artículo presenta la aplicación de la técnica Análisis de Componentes Principales (ACP) en la construcción de estratos socioeconómicos. Generalmente, las medidas de bienestar de los hogares pueden captarse a través de información sobre el ingreso, el consumo o el gasto. Sin embargo, la información detallada sobre estas variables resulta problemática y la utilización de indicadores sobre los activos que tiene el hogar constituye una opción en la medición de la desigualdad. Se discute cómo los índices socioeconómicos son construidos, cómo pueden utilizarse y sus limitaciones. Se tratan de manera específica temas relativos a la elección de las variables y la preparación de los datos.

Abstract

This paper presents the application of Principal Component Analysis (PCA) technique in the construction of socio-economic indexes. Commonly, household welfare measures are constructed considering the information on income, consumptions or expenditure. However, the details on these variables are

problematic and the use of indicators on assets of households is an option on the measurement of inequality. We present the uses and limitation of socioeconomic indexes. It specifically addressed issues related to the choice of variables and data preparation.

Palabras clave: Metodología, Estratificación, Análisis de Componentes Principales, Niveles Socioeconómicos, Bienes.

Key words: Methodology, Stratification, Principal Component Analysis, Socioeconomic levels, Assets.

Introducción

En la investigación sobre la inequidad social y la diferenciación socioeconómica de los hogares, uno de los puntos que más atención ha suscitado lo constituye la discusión sobre el ingreso como variable de segmentación. Sin embargo, esto no está exento de problemas. En efecto, los indicadores basados en esta información para un determinado periodo de referencia presentan un conjunto de inconvenientes significativos tanto en su relevamiento en campo como en su análisis posterior (Minujin y Bang, 2002). Éstos problemas son: el ocultamiento de las ganancias reales, las dificultades para captar las diferentes fuentes (trabajo informal o ingreso por rentas) y el carácter puntual de la medición ignorando las variaciones estacionales del ingreso (Minujin y Bang, 2002). Es decir, muchas veces inclusive cuando el ingreso es captado de forma correcta, hay dificultades para utilizar este dato. Además, esta información no considera el hecho de que muchos hogares pueden tener ingresos (sobre todo en áreas rurales) en especies; si bien la alternativa de relevar información proveniente del consumo o el gasto es más confiable y fácil de recolectar en términos operativos, su relevamiento en campo es mucho más costosa. En función de tales dificultades otros métodos para captar niveles socioeconómicos han sido desarrollados. Este artículo se sitúa en ese contexto y utiliza la información sobre la propiedad de bienes durables y las características materiales del hogar para

capturar diferentes niveles de vida. Esta alternativa de estratificación social aporta dos dimensiones: a) conocer la distribución de los hogares tomando una medida estructural de condiciones de vida que escape a las fluctuaciones provocadas por la economía coyuntural; b) la necesidad de estratificar grupos sociales tomando indicadores que reflejen las condiciones de vida acumuladas en el tiempo. En este trabajo se presenta la utilización del Análisis de Componentes Principales (ACP) para lograr este objetivo. La aplicación se realiza con el software R (R Core Team, 2013) y paquetes específicos (Milborrow, 2009; Wickham, 2009; Vu, 2011).

Si bien la alternativa de utilizar esta estrategía es creciente, existen puntos en los cuales no hay consenso (Vyas y Kumaranayake, 2006). En primer lugar, esta medida refleja las condiciones de vida del hogar que han sido acumuladas en el tiempo, y no captura de forma adecuada la situación actual. Esta implica que si la variable de interes está asociada con los recursos actuales, el indice no es la opción más apropiada. En segundo lugar, la propiedad de bienes del hogar no siempre captura la calidad de los mismos. Es decir, la información sobre la propiedad de un televisor no distingue entre los diferentes tipos del mismo. Por último, la dificultad es cómo agregar gran cantidad de variables en una medida unidimensional de nivel socioeconómico. Este punto es crucial, porque cada variable individualmente no es suficiente para diferenciar a los hogares.

Construcción del índice de estratos socioeconómicos (ESE)

La pregunta es ¿cómo agregamos una serie de indicadores de propiedad de bienes específicos en una sola variable que pueda servir como proxy del nivel socioeconómico? En la literatura (McKenzie, 2005; Moser y Felton, 2007) se sugieren tres alternativas:

1) Un método relativamente simple es sumar el número de activos del hogar. Este método tiene la virtud de la simplicidad, pero también tiene la limitación de asignar la misma importancia a la propiedad de diferentes activos. Además de ser totalmente arbitrario.

2) Otra forma de asignar ponderaciones a los activos es a través del precio de los mismos en el mercado. Sin embargo, este enfoque es problemático por los mismos problemas que tienen los datos del ingreso. Los datos de precios pueden ser difíciles de obtener en algunos contextos, especialmente en las economías que tienen altos niveles de trueque o informalidad.
3) La tercera opción es la utilización de la técnica Análisis de Componentes Principales para determinar la ponderación de cada variable en el índice. Caroline Moser (2009) señala que los economistas del desarrollo han seguido la recomendación de Filmer y Pritchett (2001) en la utilización del análisis de componentes principales (ACP) para agregar varias variables binarias de propiedad de activos en una sola dimensión. El ACP es relativamente fácil de calcular, entender y proporciona los pesos de cada activo de una forma más precisa que la simple suma. Existen diversas técnicas aplicadas para la construcción del índice. Sahn y Stifel (2003) emplean el análisis factorial que es utilizado más para la exploración de datos que para la reducción de dimensiones. Booysen, Van der Berg, Burger, Maltitz y Rand, (2008) utilizan el análisis de correspondencias múltiples que en su trabajo es promovido como el más adecuado cuando se trabaja con variables categóricas. Finalmente en trabajos de Kolenikov y Angeles (2009) se describe una nueva técnica, Análisis de Componentes Principales Policórico.[1]

Se describe a continuación el procedimiento de Análisis de Componentes Principales.

Análisis de Componentes Principales

El análisis de componentes principales (Jolliffe, 2002; Perez Lopez, 2004) es una técnica de análisis estadístico multivariado que puede ser clasificada dentro de los métodos de simplificación o reducción de dimensiones. La idea central del análisis de componentes principales (ACP) es reducir la

dimensionalidad de un conjunto de datos con un gran número de variables interrelacionadas entre si y reteniendo la mayor cantidad de información posible. El objetivo es obtener un menor número de variables resultado de la combinación lineal de las primitivas. Esto se logra transformando los datos a un nuevo grupo de variables, los *componentes principales (CP)*, que no están relacionados entre sí y que están ordenados de tal manera que el primer componente retiene la mayor variación presente en todas las variables originales. Se trata de una técnica para el análisis de la interdependencia (Sharma, 1996), es decir, el objetivo no es trabajar con un grupo de variables como independientes y otras como dependientes. Por el contrario, el objetivo es entender e identificar por qué y cómo las variables están relacionadas entre ellas. No es necesario que el investigador haya establecido previamente las jerarquías entre las variables, ni se necesita comprobar la normalidad en sus distribuciones.

En términos matemáticos, Vyas y Kumaranayake (2006) señalan que de un conjunto inicial de *n* variables relacionadas, el análisis de componentes principales crea índices o componentes no relacionados entre sí, donde cada componente es una combinación ponderada lineal de las variables iniciales. Por ejemplo: de un conjunto de variables X_1 a X_n,

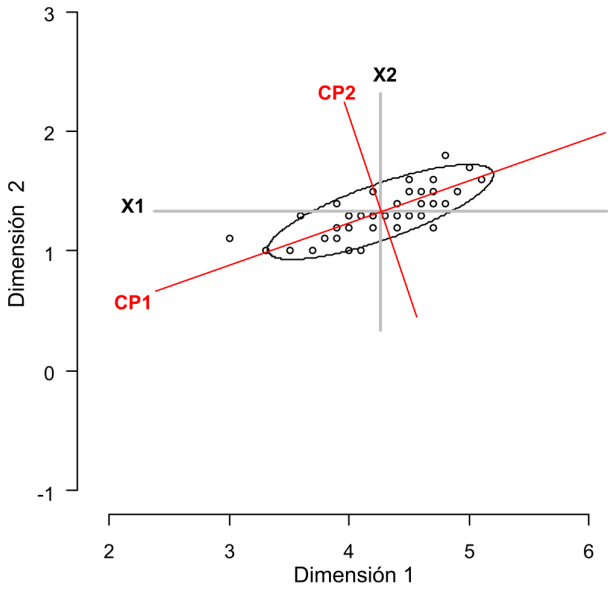
$$PC_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n$$

$$\vdots$$

$$PC_m = a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mn}X_n$$

Donde a_{mn} representa la ponderación por el *m*ésimo componente principal y la *nésima variable.* Gráficamente el concepto de ACP puede observarse en la figura 1. La propiedad de no relación entre los componentes se verifica por las líneas perpendiculares, lo cual indica que son índices que están midiendo diferentes dimensiones en los datos. La correlación entre los componentes es cero.

FIGURA 1: Representación de dos componentes en Análisis de Componentes Principales



Fuente: elaboración propia en base a Vu (2011).

La fundamentación algebraica puede resumirse de la siguiente manera (Vyas y Kumaranayake, 2006; Sharma, 1996):

El peso de cada componente principal está dado por los "vectores propios" ("eigenvectors") de la matriz de correlaciones o, si los datos originales están estandarizados por la matriz de covarianza.

La varianza (λ) para cada componente principal está dada por los "valores propios" ("eigenvalues") de los correspondientes vectores propios.[2] Los componentes son ordenados de tal manera que el primer componente (PC_1) explica la mayor cantidad de variabilidad en los datos originales, sujeto a la restricción de que la suma de los pesos al cuadrado ($a_{11}^2 + a_{12}^2 + \cdots + a_{1n}^2$) es igual a uno. Como la suma de los valores propios es igual al número de variables en el conjunto inicial de datos, la proporción del total de variación en los datos originales representada por cada componente principal está dada por λ_i/n. El segundo componente (PC2) no tiene absolutamente ninguna relación con el primer componente y explica variabilidad adicional (pero en menor cantidad) que el primer componente (sujeto

también a las mismas restricciones). De manera subsecuente los componentes siguientes no están relacionados con los anteriores y cada uno captura una dimension adicional de los datos, explicando cantidades menores de la proporción de la variabilidad en los datos originales. Debemos tener en cuenta que mientras mayor es el grado de correlacion entre las variables originales, menor es la cantidad de componentes requeridos para capurar toda la información presente.

Construcción de estratos socioeconómicos con Análisis de Componentes Principales

Inicialmente la construcción de niveles socioeconómicos a través de un índice generado por el análisis de componentes principales fue aplicado a países con problemas para captar la variable ingreso. Como señala McKenzie (2005), relevar datos sobre indicadores de activos del hogar pero no sobre ingresos o consumo es una característica general de las encuestas en países con niveles relativamente bajos de industrialización. Por ejemplo, la encuesta sobre la Agenda Pública del Sur Este de Europa, llevada a cabo por IDEA internacional, tiene una serie de preguntas detalladas sobre actitudes hacia la democracia, la participación política, la confianza en las instituciones nacionales e internacionales y el rol de los medios, y también sobre indicadores de activos para nueve países del sur este de Europa[3] (pero no sobre el ingreso). El Banco Mundial a través de su serie “Diferencias socioeconómicas en salud, nutrición y población” (Gwatkin, Rustein, Johnson, Suliman, Wagstaff y Amouzou, 2007), construyó índices de bienes basados en el ACP. La encuesta demográfica sobre salud (Demographics Health Survey)[4] es la base de datos que se toma como referencia y a través de la cual se realizaron los primeros trabajos sobre estratificación social utilizando el índice de niveles socioeconómicos a través del método de componentes principales. La encuesta demográfica sobre salud releva información sobre la propiedad de bienes durables, el acceso a los servicios públicos y de infraestructura (por ejemplo acceso a agua potable), y las características de la vivienda (número de habitaciones, material de construcción, etc.).

La construcción de índices de niveles socioeconómicos utilizando el ACP ha sido llevada a cabo por los pioneros trabajos de Filmer y Pritchet (2001) y Minujing y Bang (2002). Filmer y Pritchet (2001), muestran cómo es posible analizar la relación entre riqueza y escolarización sin contar con datos sobre el ingreso o el gasto a través de variables relativas a la posesión de bienes en India, Indonesia, Nepal y Paquistan. Minujing y Bang (2002), aplican la técnica al caso Argentino señalando la pertinencia de su aplicación en la identificación de grupos diferenciados en la distribución del acceso a bienes y servicios sociales. En ese sentido, los autores postulan la idoneidad del instrumento en contextos donde las fuentes de datos no poseen o no pueden incorporar la medición del ingreso de los hogares. Sin embargo, postulan dos puntos a considerar: a) diferencias observadas con respecto a la distribución de los hogares según el ingreso corriente revelan que los fenómenos medidos por cada indicador son distintos; b) problemas derivados de la selección final de los indicadores a ser incluidos. La condición fundamental para la construcción de este tipo de índice es la inclusión de indicadores con capacidad diferenciadora en términos sociales y que el conjunto de indicadores resulte diferenciador para toda la escala social. Sin embargo, este carácter diferencial puede variar sensiblemente en el tiempo. De esta manera, es difícil la comparabilidad en el tiempo, ya que no se podría mantener el mismo conjunto de indicadores (las comparaciones intercensales deberían ser adecuadas a partir del conjunto de indicadores análogos para uno y otro censo)

Nosotros utilizamos datos de la encuesta permanente de hogares de Paraguay correspondiente al año 2010 y se aplica el análisis de componentes principales. El objetivo principal de la Encuesta Permanente de Hogares 2010 fue generar indicadores relacionados con el empleo, el desempleo, los ingresos y otras características sociales y económicas, que permitan conocer la evolución del bienestar de la población paraguaya. La investigación va dirigida a la población que reside habitual o permanentemente en viviendas particulares. Se excluye de la investigación a la población residente en las viviendas colectivas. Esta categoría comprende: los hoteles, pensiones y otras casas de huéspedes; aunque, sí se incluyen a las familias que, formando un grupo independiente residen dentro de estos establecimientos, como puede ocurrir con los directores de los centros,

conserjes, porteros, etc. El tamaño de la muestra es de 5.003 hogares y 20.475 individuos.

Los pasos en la construcción del índice de condiciones socioeconómicas siguen la estructura propuesta por Vyas y Kumaranayake (2006): a) se presenta la selección de los bienes. Se examina la problemática relativa a la elección de los bienes y las variables que comúnmente son utilizadas; b) se aplica el análisis de componentes principales. Se presentan los problemas relativos a la preparación de los datos y la identificación del número de componentes principales a extraer; c) se interpretan los resultados; d) finalmente se clasifican los hogares en grupos socioeconómicos.

Selección de las variables a incluir

En estudios sobre la problemática se han utilizado diferentes criterios en la selección de las variables a incluir (Schellenberg J., Victora, Mushi, De Savigni, Schellemberg, D., Mishinda y Bryce, 2003; Filmer y Pritchett, 2001). Montgomery, Gragnolati, Burke y Paredes (2000) señalan la ausencia de una sola manera de seleccionar variables como proxy de las condiciones de vida y muchas veces el criterio es “ad hoc”. En efecto, no existe un criterio universal para seleccionar las variables ya que estas elecciones dependen de las variables que diferencian socioeconómicamente a los hogares en cada contexto regional. Según Minujín y Bang (2002) para individualizar grupos socialmente diferenciados resultan relevantes aquellos indicadores capaces de caracterizar la situación de los hogares reflejando su riqueza acumulada, reemplazando el ingreso corriente por el permanente. Si bien ambos se encuentran altamente correlacionados, existen disparidades.

Según Vyas y Kumaranayake (2006) el análisis de componentes principales funciona mejor cuando las variables seleccionadas están correlacionadas pero también cuando la distribución de las variables varía entre los casos, o los hogares en nuestra aplicación. Son los activos que están más desigualmente distribuidos entre los hogares los que son más ponderados en el ACP. Por el contrario, variables

con baja desviación estándar serán menos ponderadas en el ACP. Por ejemplo, un bien que todos los hogares tienen o que ningún hogar tiene (desviación estándar cero) no exhibe variabilidad entre los hogares y su ponderación será cero y de poca utilidad para diferenciar estratos socioeconómicos.

El primer paso en la construcción del índice fue la selección de los indicadores que se iban a incluir disponibles en la encuesta permanente de hogares de Paraguay aplicada en el año 2010. El criterio principal, siguiendo a Minujing y Bang (2002) fue el de incluir bienes considerados *a priori* como diferenciales (socialmente diferenciadores). La hipótesis que compartimos con los autores es que las distintas combinaciones de bienes se encuentran asociadas a grupos sociales diferenciados entre sí, cada grupo presenta tendencialmente combinaciones típicas respecto a la posesión de bienes durables.

El segundo paso es el análisis descriptivo de las variables incorporadas (tabla 1). Se seleccionaros variables relativas a la posesión de bienes del hogar y variables relativas a las características materiales de la vivienda.

Las variables ordinales fueron transformadas a variables dummy (posesión o no posesión del bien). Por ejemplo, los indicadores relativos a: material de las paredes; material del piso; material del techo; provisión de agua; disposición de la basura fueron dicotomizadas.

TABLA 1: Estadísticas descriptivas. Indicadores seleccionados

	Paraguay			
Variables	Media[5]	D.E[6]	Coef. Puntuación.	Coef * D.E.[7]
Bienes generales				
Radio	0.815	0.388	0.129	0.05
Televisor	0.875	0.331	0.171	0.057
Heladera	0.779	0.415	0.218	0.091
Cocina	0.708	0.455	0.255	0.116
Máquina de lavar	0.618	0.486	0.216	0.105
Video/DVD	0.412	0.492	0.204	0.101
Termocalefón	0.081	0.272	0.155	0.042
Aire acondicionado	0.22	0.414	0.242	0.1
Antena parabólica	0.059	0.236	0.043	0.01
Tv cable	0.161	0.367	0.206	0.076

Horno microondas	0.138	0.345	0.195	0.067
Horno eléctrico	0.245	0.43	0.187	0.081
Automóvil/camión	0.237	0.426	0.212	0.09
Motocicleta	0.421	0.494	-0.009	-0.004
Personas/Cuartos para dormir	2.054	1.263	-0.137	-0.173
Material de las paredes				
Estaqueo	0.004	0.063	-0.033	-0.002
Adobe	0.004	0.065	-0.036	-0.002
Madera	0.33	0.47	-0.234	-0.11
Ladrillos	0.651	0.477	0.242	0.115
Bloque de cemento	0.006	0.077	0.023	0.002
Tronco de palma	0.003	0.053	-0.027	-0.001
Cartón o madera	0.001	0.037	-0.023	-0.001
Material de construcción del piso				
Tierra	0.148	0.355	-0.204	-0.072
Madera	0.022	0.148	-0.029	-0.004
Ladrillos	0.087	0.281	-0.056	-0.016
Lecherada	0.296	0.456	-0.089	-0.041
Baldosa común	0.219	0.413	0.13	0.054
Mosaico, cerámica	0.225	0.418	0.188	0.079
Parquet	0.002	0.042	0.011	0
Alfombra	0.001	0.032	0.017	0.001
Material del techo				
Teja	0.569	0.495	0.181	0.09
Paja	0.056	0.229	-0.13	-0.03
Fibrocemento	0.226	0.419	-0.131	-0.055
Provisión de agua	0.101	0.302	-0.069	-0.021
Tablilla de madera	0.003	0.053	-0.002	0
Hormigón armando, loza	0.044	0.205	0.081	0.017
Cartón, madera	0.001	0.035	-0.025	-0.001
Provisión de agua				
ESSAP	0.236	0.425	0.178	0.076
Senasa o junta de saneamiento	0.279	0.449	-0.056	-0.025
Red comunitaria	0.136	0.342	-0.032	-0.011
Red privada	0.077	0.267	0.04	0.011
Pozo artesiano	0.018	0.134	0.024	0.003
Pozo con bomba	0.124	0.329	-0.006	-0.002
Pozo sin bomba	0.114	0.318	-0.143	-0.046
Manantial	0.011	0.105	-0.049	-0.005

Tajamar	0.002	0.049	-0.023	-0.001
Agua de lluvia	0.002	0.042	0.002	0
Disposición de la basura				
Quema	0.493	0.5	-0.235	-0.117
Recoge camión	0.401	0.49	0.262	0.128
Tira al arroyo	0.063	0.243	-0.033	-0.008
Tira al patio, baldío	0.024	0.152	-0.019	-0.003
Tira al vertedero municipal	0.009	0.094	0.019	0.002
Tira en la chacra	0.006	0.078	-0.018	-0.001
Tira en arroyo, rio o laguna	0.004	0.066	-0.008	-0.001

Nota: el porcentaje de varianza explicada por el primer componente es 13%. El primer valor propio es 7.4 y el segundo 2.46.

Fuente: elaboración propia en base a EPH 2009, Paraguay.

Aplicación del análisis de componentes principales

El análisis de componentes principales eliminó aquellas variables cuya varianza era cero. Un punto relevante en este paso es el tema de los valores perdidos donde hay dos criterios: 1) eliminar los hogares con valores perdidos en las variables seleccionadas, como señalan Houweling, Kunst y Mackenbach (2003); 2) reemplazar los valores perdidos por el valor medio de esa variable, como señalan Gwatkin y otros (2007). En nuestro caso y en función de que el número de casos perdidos era insignificante se optó por borrar los casos con valores perdidos. La decisión de un criterio u otro depende de las características de la base y los objetivos. En efecto, borrar casos puede provocar que se disminuya el tamaño de la muestra de manera considerable y sesgar el análisis. En tanto que imputar los valores perdidos puede provocar que haya una reducción de la variabilidad entre los hogares e incrementar los peligros del “truncamiento” y “agrupamiento”. En ambos casos el peligro es más pronunciado con altos valores de casos perdidos.

El número de componentes principales a extraer es definido por el usuario y una regla común es seleccionar aquellos componentes cuyo valor propio asociado es mayor que uno. Sin embargo, para la construcción de niveles socioeconómicos el primer componente principal es el que se asume que captura la mayor variabilidad

posible (Vyas y Kumaranayake, 2006). Estudios como los de McKenzie (2005) o Kolenikov (2009) consideran la utilización de componentes adicionales en la caracterización socioeconómica de los hogares, sin embargo, concluyen que el primer componente es suficiente. Filmer y Pritchett (2001), también consideran la utilización de componentes adicionales en el análisis, sin embargo, concluyen que los coeficientes son difíciles de interpretar.

El valor propio de cada componente principal indica el porcentaje de variabilidad en el total de los datos que es explicado. En los estudios previos, el primer componente explica entre el 12 % y el 27% (McKenzie, 2005), entre el 13% y 16% (Vyas y Kumaranayake, 2006) y 26% (Filmer y Pritchett, 2001). Estos porcentajes no son altos e indican la complejidad de las correlaciones entre las variables.

Los resultados en nuestros análisis son presentados en la Tabla 1 y 2. Para la totalidad de los hogares el primer componente explica el 13 % de la covarianza. En tanto que para hogares urbanos el valor es del 12% y para hogares rurales de 11%.[8]

Interpretación de los resultados

El resultado del análisis de componentes principales se presenta en la tabla en la columna sobre los coeficientes de puntuación. Los puntajes positivos generalmente están asociados a valores altos en el índice de condiciones socioeconómicas y de manera inversa, valores negativos están asociados a puntuaciones bajas en el índice. De acuerdo a la interpretación de McKenzie (2005), como todas las variables (excepto número de habitaciones por cuarto) toman el valor de 0 y 1 la ponderación tiene una interpretación relativamente sencilla: un hogar que tiene cocina tiene un índice 0.116 mayor que uno que no lo tiene. Un hogar que tiene televisor aumenta el índice 0.057 unidades (datos presentados en la cuarta columna, resultado de la múltiplicación entre la desviación estandar y el coeficiente de puntuación del ACP).

En relación a los coeficientes de puntuación, Vyas y Kumaranayake (2006), señalan que en algunos estudios (Gwatkin y otros, 2007; Houweling, Kunst y

Mackenbach, 2003; McKenzie, 2005) se puede observar que la posesión de algunos bienes durables, como bicicletas, contribuyen a una ponderación negativa en el indice Esto implica que un hogar que posee una motocicleta puede tener una puntuación más baja en el indice que un hogar que no la tenga. La razón para este paradójico resultado está en que la posesión de una motocicleta está más asociada con variables que se espera que tengan menor puntuación en el índice (como baja calidad de agua y condiciones sanitarias). Según la revisión bibliográfica, situaciones como esta pueden encontrarse cuando el índice se construye de manera combinada para hogares rurales y urbanos. Como puede observarse en la tabla 1, el coeficiente de la posesión de motocicleta tiene un valor negativo.

Utilizando los coeficientes de puntuación del análisis de componentes principales como ponderadores, se puede construir una variable dependiente para cada hogar (Y_1), media 0 y desviación estandar 1. Esta variable dependiente puede ser considerada como una "puntuación socioeconómica", mientras más alta la puntuación mayor es el nivel socioeconómico del hogar. Un hogar con mayor cantidad de bienes, mejor material del piso, el techo, etc tendrá una puntuación mayor.

Posibles problemas en la construcción: *truncamiento* y *agrupamiento*

En la construcción de niveles socioeconómicos con analisis de componentes principales, dos problemas son importantes a evitar: el "agrupamiento" y el "truncamiento" (McKenzie, 2005; Vyas y Kumaranayake, 2006). El *agrupamiento* implica que hay un número insuficiente de indicadores y los hogares son agrupados en un número pequeño de grupos (en el ejemplo extremo, con solo un indicador hay un grupo que tiene el bien y otro grupo que no lo tiene). Esta situación limita la información útil que puede obtenerse sobre desigualdad tomando en consideración los activos del hogar. El *truncamiento* implica los problemas derivados de la distribución del índice de activos, que ocurre cuando no hay indicadores que puedan diferenciar entre los hogares pobres y muy pobres o entre los hogares ricos y muy ricos. McKenzie (2005) sugiere graficar un histograma y la función de densidad de la probabilidad del índice para observar si el agrupamiento o el truncamiento aparecen.

La solución para estos casos es incorporar más indicadores en el análisis de componentes principales que puedan capturar desigualdades entre los hogares. En cuanto al número de variables consideradas los estudios van desde 10 (Schellenberg y otros, 2003) hasta 30 (McKenzie, 2005). Otros métodos para incluir variables comprenden la incorporación de variables continuas y una combinación de bienes durables, acceso a infraestructura, características del hogar y otras variables relevantes como indicadores de nivel socioeconómicos. El punto crítico es poder capturar con los indicadores seleccionados el nivel socioeconómico de los hogares.

Para ejemplificar este punto se construyó el mismo indicador pero segmentando por áreas: uno para los hogares urbanos y otros para los hogares rurales (Tabla 2). Como puede observarse en la tabla hay algunos ítems problemáticos: el material de las paredes en los hogares urbanos, en este caso el material estaqueo y adobe no parece ser un indicador que pueda diferenciar estratos, lo mismo ocurre con el bloque de cemento o paredes de tronco y palma. En la mayoría de estos hogares el ladrillo es el material principal de las paredes. Lo mismo ocurre con la provisión de agua, muchos ítems son borrados del análisis de componentes principales porque su varianza es cero.

Como puede observarse en la tabla, existen diferencias marcadas entre los hogares pertenecientes al ámbito urbano y aquellos que pertenecen al espacio rural. Por ejemplo: la mayoría de los hogares tienen cocina en el medio urbano (casi 90%) en tanto que los rurales tan solo 64%. Estas diferencias en relación a los bienes que posee el hogar son profundas en todos los bienes bajo consideración, indicando situaciones más precarias en los espacios rurales. La cantidad de personas por cuarto para dormir (tomado como un indicador de hacinamiento) es mayor en los hogares rurales (2.27 por cuarto) en tanto que en los hogares urbanos la cantidad está alrededor de 1.94 personas por cuarto.

En relación al material de las paredes. En los hogares urbanos el material predominante es ladrillo (80%) en tanto que en los hogares rurales la mayoría tiene paredes de madera (casi 60%). El material de construcción de los pisos también acusa diferencias importantes, la mayoría de los hogares rurales tiene pisos de tierra o lecherada (entre ambos suman casi el 65%) mientras que en los hogares urbanos la mayoría de los pisos son de lecherada, baldosa o mosaico (entre esos

tres ítems suman casi el 85%). La disposición de la basura (como un indicador también de las disparidades en cuanto a los servicios comunitarios). La mayoría de los hogares urbanos dispone la basura a través de un camión que la recoge (62%). En tanto que en los hogares rurales la mayoría de los hogares quema la basura (81%).

TABLA 2: Estadísticas descriptivas. Indicadores seleccionados

	Paraguay Urbano				Paraguay Rural			
Variables	Media	D.E	Coef.punt.	Coef * D.E	Media	D.E	Coef.punt.	Coef * D.E
Bienes generales								
Radio	0.848	0.359	0.139	0.050	0.764	0.425	0.16	0.068
Televisor	0.945	0.228	0.145	0.033	0.765	0.424	0.2041	0.087
Heladera	0.863	0.344	0.217	0.074	0.648	0.478	0.2726	0.130
Cocina	0.863	0.344	0.231	0.079	0.467	0.499	0.2921	0.146
Máquina de lavar	0.728	0.445	0.206	0.092	0.445	0.497	0.2563	0.127
Video/dvd	0.516	0.500	0.195	0.097	0.250	0.433	0.2355	0.102
Termocalefón	0.120	0.325	0.171	0.056	0.018	0.134	0.108	0.015
Aire acondicionado	0.323	0.468	0.265	0.124	0.058	0.234	0.2037	0.048
Antena parabólica	0.045	0.208	0.054	0.011	0.081	0.273	0.1544	0.042
Tv cable	0.254	0.435	0.217	0.095	0.016	0.125	0.1168	0.015
Horno microondas	0.205	0.404	0.212	0.086	0.033	0.178	0.1457	0.026
Horno eléctrico	0.319	0.466	0.185	0.086	0.129	0.335	0.2304	0.077
Automóvil/camión	0.314	0.464	0.230	0.107	0.119	0.323	0.2248	0.073
Motocicleta	0.360	0.480	-0.007	-0.003	0.517	0.500	0.1364	0.068
Personas/hab. Dormir	1.914	1.115	-0.135	-0.151	2.271	1.436	-0.1718	-0.247
Material de las paredes								
Estaqueo	0.001	0.036	-0.017	-0.001	0.008	0.090	-0.0484	-0.004
Adobe	0.001	0.031	-0.032	-0.001	0.009	0.096	-0.0427	-0.004
Madera	0.185	0.388	-0.231	-0.090	0.557	0.497	-0.1978	-0.098
Ladrillos	0.802	0.399	0.229	0.091	0.416	0.493	0.2264	0.112
Bloque de cemento	0.009	0.095	0.013	0.001	0.001	0.032	0.0131	0.000
Tronco de palma	0.001	0.036	-0.018	-0.001	0.005	0.071	-0.0414	-0.003
Cartón o madera	0.001	0.026	-0.021	-0.001	0.003	0.051	-0.0358	-0.002
Material del piso								
Tierra	0.055	0.228	-0.167	-0.038	0.294	0.456	-0.2633	-0.120
Madera	0.010	0.099	-0.043	-0.004	0.042	0.201	0.0441	0.009
Ladrillos	0.055	0.229	-0.081	-0.019	0.135	0.342	0.01	0.003
Lecherada	0.264	0.441	-0.169	-0.075	0.345	0.475	0.0337	0.016

Baldosa común	0.302	0.459	0.101	0.047	0.088	0.283	0.1199	0.034
Mosaico, cerámica	0.310	0.463	0.190	0.088	0.093	0.290	0.1873	0.054
Parquet	0.001	0.031	-0.004	0.000	0.003	0.055	0.0585	0.003
Alfombra	0.002	0.040	0.018	0.001	0.000	0.000	0.000	0.000
Material del techo								
Teja	0.656	0.475	0.179	0.085	0.433	0.496	0.1881	0.093
Paja	0.010	0.102	-0.065	-0.007	0.126	0.332	-0.1747	-0.058
Fibrocemento	0.190	0.393	-0.181	-0.071	0.283	0.451	-0.0789	-0.036
Chapa Cinc	0.073	0.260	-0.096	-0.025	0.145	0.352	-0.0057	-0.002
Tablilla de madera	0.003	0.051	0.007	0.000	0.003	0.055	-0.0168	-0.001
Hormigón armando, loza	0.067	0.251	0.070	0.018	0.008	0.087	0.0572	0.005
Cartón, madera	0.000	0.018	-0.021	0.000	0.003	0.051	-0.0357	-0.002
Provisión de agua								
ESSAP	0.378	0.485	0.153	0.074	0.015	0.121	0.033	0.004
Senasa o saneamiento	0.227	0.419	-0.079	-0.033	0.361	0.480	0.0377	0.018
Red comunitaria	0.125	0.330	-0.033	-0.011	0.152	0.360	-0.0347	-0.012
Red privada	0.108	0.310	-0.005	-0.001	0.030	0.170	0.0648	0.011
Pozo artesiano	0.020	0.139	0.020	0.003	0.016	0.127	0.0538	0.007
Pozo con bomba	0.091	0.288	-0.010	-0.003	0.174	0.379	0.109	0.041
Pozo sin bomba	0.051	0.219	-0.129	-0.028	0.214	0.410	-0.1508	-0.062
Mantantial	0.001	0.026	-0.008	0.000	0.028	0.164	-0.0497	-0.008
Tajamar	0.000	0.000	0.000	0.000	0.006	0.078	-0.0225	-0.002
Agua de lluvia	0.000	0.000	0.000	0.000	0.005	0.068	0.0414	0.003
Disp. de la basura								
Quema	0.284	0.451	-0.231	-0.104	0.817	0.386	-0.1152	-0.045
Recoge camión	0.626	0.484	0.260	0.126	0.050	0.218	0.1486	0.032
Tira al arroyo	0.044	0.205	-0.057	-0.012	0.092	0.289	0.0319	0.009
Tira al patio, baldío	0.025	0.155	-0.050	-0.008	0.022	0.147	0.003	0.000
Tira al vertedero mun.	0.013	0.115	0.004	0.000	0.002	0.045	0.0346	0.002
Tira en la chacra	0.000	0.000	0.000	0.000	0.016	0.125	0.0089	0.001

Nota: en los hogares urbanos, el porcentaje de varianza explicada por el primer componente es 12%. El primer valor propio es 6.29 y el segundo 2.55. En los hogares rurales, el porcentaje de varianza explicada por el primer componente es 11%. El primer valor propio es 5.75 y el segundo 2.69.
Fuente: elaboración propia en base a EPH 2009, Paraguay.

Se graficaron los histogramas del índice para observar si el agrupamiento o el truncamiento aparecen (gráfico 2).

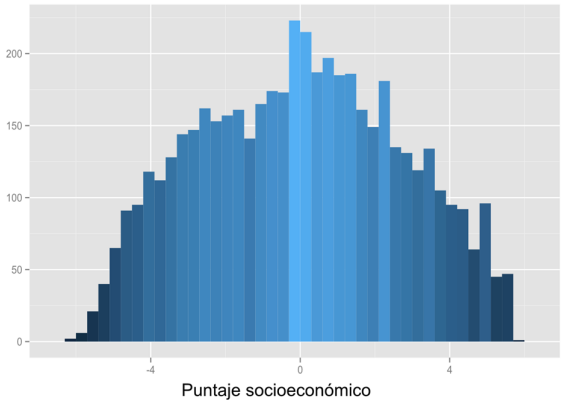
La distribución del índice tiende a seguir la distribución de una curva normal para el total de los hogares. Sin embargo, para los hogares urbanos se puede

observar una asimetría a la izquierda. Para los hogares rurales, una asimetría a la derecha. Estas asimetrías señalan los problemas de truncamiento y agrupamiento, lo que hace difícil diferenciar los grupos socioeconómicos. En el caso de los hogares urbanos, la dificultad está en los estratos socioeconómicos más altos y en los hogares rurales en los estratos socioeconómicos más bajos.

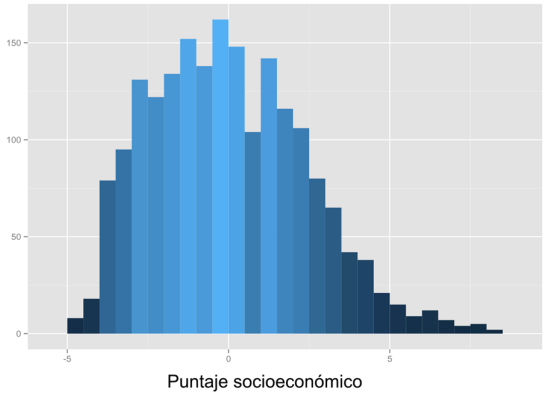
Como señalan Vyas y Kumaranayak (2006) la decisión de construir el índice a nivel nacional o por niveles comunitarios depende de los objetivos del estudio y de la comparación. A nivel nacional existe el riesgo de fallar en capturar las diferencias a nivel de áreas. Sin embargo, construir el índice a nivel comunitario o por áreas trae consigo el problema del truncamiento o agrupamiento, como en este caso puede observarse. La solución en estos casos es incluir bienes o ítems que puedan servir como indicadores que permitan diferenciar entre estratos en específicas comunidades, a través del conocimiento local en profundidad. Esto implica la generación de datos primarios y la necesaria evaluación del costo asociado a esto o la relativa simplicidad de utilizar índices a nivel agregado.

FIGURA 2: Distribución de los puntajes socioeconómicos

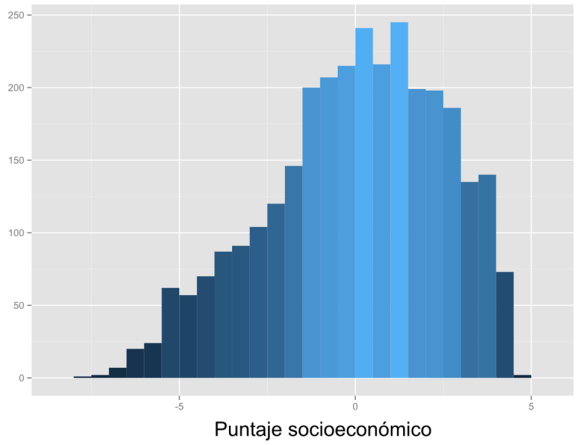
Total de hogares paraguayos



Total de hogares rurales



Total de hogares urbanos



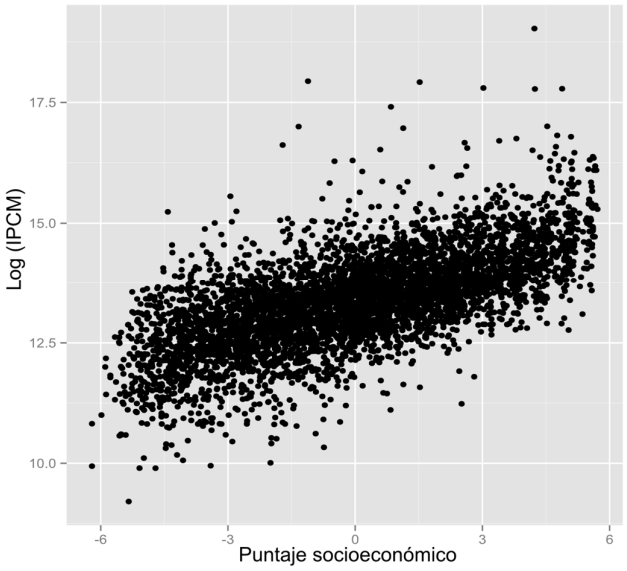
Fuente: elaboración propia en base a EPH 2010, Paraguay.

Clasificación de los hogares en grupos socioeconómicos

Después de la construcción del índice socioeconómico (Y_1), se puede utilizar el mismo como variable independiente continua en un modelo de regresión (aunque las interpretaciones de los coeficientes no son sencillas). Una construcción de este tipo sirve como input para análisis de desigualdad y pobreza. Al ser un índice sintético cuantitativo puede ser utilizado en procedimientos/análisis estadísticos en los distintos tipos de regresiones, o también puede ser "cerrado" en categorías cualitativas a partir de su agrupación en distintos tipos de intervalos.

La mayoría de los estudios construyen quintiles para diferenciar categorías socioeconómicas. A fines analíticos, clasificamos los hogares en quintiles según el índice de Estratos Socioeconómicos (ESE) y también a través de la variable Ingreso per cápita mensual (resultado de la sumatoria de todos los ingresos de los hogares dividido por la cantidad de miembros del hogar)[9]. Como puede observarse en la figura 3 hay una relación significativa entre las dos variables, *r*= .6635., p <.05.

FIGURA 3: Ingreso e Índice de Estratos Socioeconómicos



Fuente: elaboración propia en base a EPH 2010.

A fines de ejemplificar el procedimiento se tabulan ambas variables para los hogares con migrantes recientes (en los últimos cinco años) y para hogares sin migrantes recientes. Como puede observarse en la Tabla 3, si bien existen disparidades entre los dos indicadores construidos, el análisis es convergente.

TABLA 3: Estratificación según el ingreso e índice de estratos socioeconómicos

		Quintiles				
Tipo de Hogar	*Variable de estratificación*	1	2	3	4	5
Migrantes recientes	IPCM (Ingreso per cápita mensual)	22.25	26.23	19.91	18.03	13.58
	INDICE ESE con ACP	22.01	23.42	24.59	17.8	12.18
Sin migrantes recientes	IPCM (Ingreso per cápita mensual)	19.81	19.43	20	20.25	20.51
	INDICE ESE con ACP	19.83	19.7	19.54	20.25	20.68

Fuente: elaboración propia en base a EPH 2009, Paraguay.

Conclusiones

Este artículo exploró las posibilidades de construir estratos socioeconómicos considerando una alternativa a las variables sobre el ingreso, el consumo o el gasto de los hogares. Para lograr este objetivo se utilizó la técnica analisis de componentes principales y se consideraron los bienes y las características del hogar. Una de las principales utilidades de este método es poder solucionar los problemas que existen en las encuestas cuando se mide el ingreso, el consumo o el gasto de los hogares. Si bien existen muchas alternativas para estratificar a los hogares, el ACP es una técnica relativamente sencilla y la mayoría de los paquetes estadísticos pueden realizarla. Los datos utilizados son los que se relevan en la mayoría de las encuestas y se utilizan todas las variables para reducir su dimensionalidad. Los debates sobre el uso de componentes principales, análisis factorial o análisis de correspondencias múltiples reflejan el hecho de que es un indice artificialmente construido. En efecto, si bien este tipo de índice es una medida relativa del nivel socioeconómico y es muy útil para medir la desigualdad entre los hogares, no lo es para detallar información sobre los niveles absolutos de pobreza. Como pudo observarse, cuando se construyeron estratos socioeconómicos su consistencia con la estratificación en base al ingreso es alta (aunque no necesariamente miden lo mismo). En tal sentido, resulta un instrumento apropiado para el estudio de la estratificación social, especialmente en contextos donde es difícil y problemático el relevamiento de la variable ingreso.

Las perspectivas de su aplicación son amplias, debido a la facilidad del relevamiento de los indicadores utilizados como proxy del ingreso que perciben los hogares. Desde un punto de vista metodológico, una alternativa a explorar es la operacionalización de todos los activos que cuenta el hogar, ampliando la selección de considerar solamente a los bienes durables o las características del hogar. Una aproximación de este tipo debería incorporar variables relativas al capital social de los hogares y una serie de dimensiones más amplias que permitan capturar la mayor complejidad involucrada en el concepto de "activos del hogar", en el marco de estrategias de vida.

Referencias bibliográficas

BATTISTIN, Erich; BLUNDELL, Richard y LEWBE, Arthur. (2009). "Why Is Consumption More Log Normal than Income? Gibrat's Law Revisited". *Journal of Political Economy* 117 (6), 1140-54

BOOYSEN, Frikkie; VAN DER BERG, Servaas; BURGER, Ronelle; VON MALTITZ, Michael y DU RAND, Gideon. (2008). "Using an Asset Index to Assess Trends in Poverty in Seven Sub-Saharan African Countries." *World Development*, 36 (6), 1113–1130.

FILMER, Deon y PRITCHETT, Lant H.(2001). "Estimating Wealth Effects without Expenditure Data-or Tears: An Application to Educational Enrollments in States of India." *Demography*, 38 (1), 115–132.

GWATKIN, Davidson; RUTSTEIN, Shea; JOHNSON, Kiersten; SULIMAN, Eldaw; WAGSTAFF, Adam y AMOUZOU, Agbessi. (2007). *Socio-Economic Differences in Health, Nutrition, and Population in Bolivia.* Washington, D.C. : The World Bank.

HOUWELING, Tanja AJ; KUNST, Anton y MACKENBACH, Johan (2003). "Measuring Health Inequality among Children in Developing Countries: Does the Choice of the Indicator of Economic Status Matter?" *International Journal for Equity in Health*,2 (1),1-12.

JOLLIFFE, Ian. (2002). *Principal Component Analysis*. New York: Springer.

KOLENIKOV, Stanislav y ANGELES, Gustavo. (2009). "Socioeconomic Status Measurement with Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer?". *Review of Income and Wealth,* 55 (1), 128–165.

MCKENZIE, David J. (2005). "Measuring Inequality with Asset Indicators." *Journal of Population Economics,* 18 (2), 229–260.

MILBORROW, Stephen. (2009). *plotpc: Plot principal component histograms around a scatter plot*. R package version 1.0-2. http://CRAN.R-project.org/package=plotpc Última consulta en junio de 2014.

MINUJIN, Alberto y BANG, Joon Hee. (2002). "Indicadores de Inequidad Social Acerca Del Uso Del 'Índice de Bienes' Para La Distribución de Hogares." *Desarrollo Económico,* (42), 129–46.

MONTGOMERY, Mark R.; GRAGNOLATI, Michele; BURKE, Kathleen A. y PAREDES, Edmundo.(2000). "Measuring Living Standards with Proxy Variables". *Demography*, 37 (2), 155-174.

MOSER, Caroline y FELTON, Andrew. (2007). "The Construction of an Asset Index Measuring Asset Accumulation in Ecuador". *Development, s/d.* (CPRC working paper).

MOSER, Caroline. (2009). *Ordinary Families, Extraordinary Lives: Assets and Poverty Reduction in Guayaquil, 1978-2004*. Brookings Institution Press.

PEREZ LOPEZ, Cesar. (2004). *Técnicas de Análisis Multivariante de Datos*. Madrid. España: Pearson. Prentice Hall.

R CORE TEAM. (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/ Última consulta en junio de 2014.

SAHN, David y STIFEL, David. (2003). "Exploring Alternative Measures of Welfare in the Absence of Expenditure Data." *Review of Income and Wealth,* 49 (4), 463–489.

SCHELLENBERG, Joanna Armstrong; VICTORA, Cesar G; MUSHI, Adiel; DE SAVIGNY, Don; SCHELLENBERG, David; MSHINDA, Hassan y BRYCE, Jennifer.(2003). "Inequities among the Very Poor: Health Care for Children in Rural Southern Tanzania." *The Lancet,* 361 (9357), 561–566.

SHARMA, Subhash. (1996). *Applied Multivariate Techniques*. Hoboken, NJ: John Wiley & Sons.

UEBERSAX, John S. (2006). "The Tetrachoric and Polychoric Correlation Coefficients. Statistical Methods for Rater Agreement." *Statistical Methods for Rater Agreement*. Recuperado de: http://www.john-uebersax.com/stat/tetra.htm Última consulta en junio de 2014.

VYAS, Seema y KUMARANAYAKE, Lilani. (2006). "Constructing Socio-Economic Status Indices: How to Use Principal Components Analysis." *Health Policy and Planning*, 21 (6), 459–468.

VU, Vincent Q. (2011). *ggbiplot: A ggplot2 based biplot*. R package version 0.55. http://github.com/vqv/ggbiplot Última consulta en junio de 2014.

WICKHAM, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

Anexos

TABLA 4: Estadísticos descriptivos. Índice de Estratos Socioeconómicos

Observaciones	5003
Media	0,000
Desviación Estándar	2,72
Varianza	7,41
Asimetría	0,015
Curtosis	-0.86
Mínimo	-5,71
Máximo	6,21
Rango: máximo-mínimo	11,92
Q1	2,16
Q2, media	-0.03
Q3	2,09
Rango intercuartílico: Q3-Q1	-0,07

TABLA 5: Estadísticos descriptivos. Índice de Estratos Socioeconómicos. Hogares Rurales

Observaciones	1955
Media	0,000
Desviación Estándar	2,40
Varianza	6
Asimetría	0,445
Curtosis	2,84
Mínimo	-5,00
Máximo	8,21
Rango: máximo-mínimo	13,21
Q1	-1,87
Q2, media	-0,187
Q3	1,68
Rango intercuartílico: Q3-Q1	3,55

TABLA 6: Estadísticos descriptivos. Índice de Estratos Socioeconómicos. Hogares Urbanos

Observaciones	3048
Media	0,000
Desviación Estándar	2,51
Varianza	6
Asimetría	0,419
Curtosis	2,55
Mínimo	-7,83
Máximo	4,61
Rango: máximo-mínimo	12,44
Q1	-1,60
Q2, media	0,215
Q3	1,93
Rango intercuartílico: Q3-Q1	3,52

Notas

[1] El término en inglés es “Polychoric Principle Components Analysis”.
La mayoría de los trabajos, sin embargo, utilizan el ACP estándar. El ACP Policórico se basa en correlaciones policóricas. La correlación policóricas, supone que cada variable ordinal fue obtenida categorizando una variable subyacente con distribución normal. Si una de las variables ordinales tiene solo dos categorías la correlación entre dos variables es tetracórica. Un concepto cercano es el de correlaciones poliseriales, que se dan cuando una variable es continua (con normalidad asumida) y otra ordinal. Si la variable ordinal en este caso tiene dos categorías la correlación tiene el nombre de biserial. (Si la variable ordinal tiene más de 10 categorías el ACP Policórico las trata como continuas utilizando la correlación de Pearson). Para más detalle ver Uebersax (2006) y Kolenikov y Angeles (2009).
[2] Para más detalle ver Sharma (1996) y Jolliffe (2002).
[3] Los países relevados fueron: Bulgaria, Macedonia, Kosovo, Rumania, República Srpska, Bosnia y Herzegovina, Serbia, Montenegro y Croacia. Véase www.idea.int/europe_cis/balkans
[4] Disponible en http://www.measuredhs.com/
[5] Cada variable toma el valor de 1 cuando se posee el bien y 0 cuando no (a excepción de la variable personas por cuarto).
[6] Desviación Estándar.
[7] Coeficiente de puntuación * Desviación estándar.
[8] En el anexo se presentan los estadísticos descriptivos de los índices construidos
[9] El lector debe tener en consideración que una transformación logit de la variable ingreso es preferida ya que tiene propiedades que la hacen más deseable. Para más detalle sobre este punto ver el trabajo de Battistin, Blundell y Lewbel (2009).

Fecha de recepción: 28 de noviembre de 2013. Fecha de aceptación: 24 de mayo de 2014.