MATEMÁTICA
ESPOL - FCNM JOURNAL
2020, Vol. 18, No. 2 Julio - Diciembre

Construcción de un índice de riqueza utilizando Componentes Principales

Construction of a wealth index using Principal Components Analysis

Eva María Mera, Mario Solórzano, Katherine Loor

Recepción: 23/07/2020 **Aceptación:** 30/09/2020 **Publicación:** 20/10/2020

Abstract The measurement of wealth is a matter of study of Econometrics, within this discipline there are various proposals for the construction of the wealth index. In this work a review is made on the construction method from the income approach, using proxy variables and applying the statistical technique of principal components. At the end of the article, an illustration of this methodology is made using the data from the ENSANUT National Health and Nutrition Survey of 2018 conducted for the Ecuadorian population by the National Institute of Statistics and Censuses – INEC.
Keywords INEC, principal components, proxy variables, wealth index.

Resumen La medición de la riqueza es materia de estudio de la Econometría, dentro de esta disciplina hay diversas propuestas para la construcción del índice de riqueza. En este trabajo se hace una revisión sobre el método de construcción desde el enfoque del ingreso, utilizando variables proxy y aplicando la técnica estadística de componentes principales. Al final del artículo se realiza una ilustración de esta metodología utilizando los datos de la Encuesta Nacional de Salud y Nutrición ENSANUT del año 2018 realizada para la población ecuatoriana por parte del Instituto

Eva María Mera Intriago, M.Sc.
Docente, ESPOL Polytechnic University, Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ciencias Naturales y Matemáticas (FCNM), Campus Gustavo Galindo Km. 30.5 Vía Perimetral P.O. Box 09-01-5863, Guayaquil, Ecuador, e-mail: emera@espol.edu.ec,
https://orcid.org/0000-0003-4276-4714.

Mario David Solórzano Carvajal, M.Sc.
Investigador, ESPOL Polytechnic University, Escuela Superior Politécnica del Litoral, ESPOL, Centro de Estudios e Investigaciones Estadísticas (CEIE), Campus Gustavo Galindo Km. 30.5 Vía Perimetral P.O. Box 09-01-5863, Guayaquil, Ecuador, e-mail: mdsolorz@espol.edu.ec,
https://orcid.org/0000-0001-8682-6652.

Katherine Andreina Loor Valeriano, M.Sc.
Estudiante de doctorado, Universidad Estatal de Campinas, UNICAMP, Campinas, Brasil, e-mail: kataloor@espol.edu.ec, https://orcid.org/0000-0001-6388-4753.

Nacional de Estadística y Censos INEC.
Palabras Claves componentes principales, índice de riqueza, INEC, variables proxy.

1 Introducción

Conocer el nivel socioeconómico de las personas y hogares que conforman una población en particular permite medir el efecto que éste tiene sobre otras variables como el acceso a la educación, estado de salud, fecundidad, hábitos alimenticios, entre otros. Usualmente el nivel socioeconómico de un hogar se construye en función de los ingresos económicos o hábitos de consumo de sus miembros aunque no siempre se tiene acceso a estos datos, o en otras ocasiones no son confiables especialmente para países en desarrollo donde los ingresos pueden venir de diferentes actividades algunas de ellas realizadas de manera esporádica o en algún período específico dentro del año, además está la renuencia por parte del entrevistado a divulgar datos de sus ingresos o consumos (Howe, Hargreaves, y Huttly, 2008).

Este artículo fue elaborado por motivo de la conmemoración de los 25 años de creación de la carrera Ingeniería Estadística de la Escuela Superior Politécnica del Litoral. Este documento está compuesto por seis secciones, la segunda sección corresponde a la definición de las variables proxy a utilizar para la construcción de un índice de riqueza, en la tercera se explica la técnica multivariada de componentes principales haciendo para ello uso del Algebra lineal, la cuarta sección contiene la metodología para la construcción del índice de riqueza ya sea partiendo de la matriz de varianzas y covarianzas o de la matriz de correlaciones muestral, en la quinta sección se ilustra la construcción del índice mediante un ejercicio con datos de la encuesta de salud y nutrición realizada en 2018 para Ecuador, finalmente la última sección está reservada para los comentarios finales de este trabajo y algunas observaciones.

2 Variables proxy

Una alternativa para la medición del nivel socioeconómico de un hogar que no involucre ingresos o consumos, es la construcción de un índice de riqueza basado en variables fáciles de observar que tengan una alta correlación con la riqueza en el hogar. A estas variables se las conoce como proxy. La posesión de bienes duraderos (refrigerador, televisor, computadora, carro, etc.) y condiciones de vida (acceso a agua potable, acceso a energía eléctrica, acceso a internet, tipo de servicio higiénico, material del piso de la vivienda, etc.) son variables utilizadas comúnmente como proxys de la riqueza de un hogar tal como se utilizan en Córdova (2008), siendo estas del tipo cualitativa casi siempre dicotómicas (Sí/No) o con pocas categorías de respuesta. Al ser el nivel socioeconómico, y, una variable latente, se busca que sea explicada a través de una combinación lineal de variables proxy, siendo $\mathbf{x}^\top =$

$(x_1, x_2, \ldots, x_p) \in \mathbb{R}^p$ el vector que contiene a las p variables proxy utilizadas para la construcción del indicador, el nivel socioeconómico y sería:

$$y = \mathbf{w}^\top \mathbf{x},$$

donde $\mathbf{w} \in \mathbb{R}^p$ es el vector de coeficientes o de ponderaciones que se asignan a cada variable proxy para la explicación del nivel socioeconómico, la asignación de estas ponderaciones se las podría realizar en base a ciertas consideraciones, por ejemplo se podría tomar el valor monetario del bien como ponderación o simplemente que $\mathbf{w}$ sea un vector de unos. Desde el punto de vista estadístico, en éste documento se realizará la construcción del indicador de riqueza a través del método de componentes principales implementado por Filmer y Pritchett (2001), el cual busca reemplazar un vector de variables correlacionadas $\mathbf{x} \in \mathbb{R}^p$ por un vector $\mathbf{y} \in \mathbb{R}^p$ de variables no correlacionadas, denominadas componentes principales, que es combinación lineal del vector original $\mathbf{x}$. Lo que se busca al utilizar esta técnica es la reducción de dimensión del espacio de variables pudiendo con un cierto número k ($k < p$) de variables, explicar la mayor variabilidad posible de $\mathbf{x}$. Para la construcción del índice se utilizará la primera componente principal, la cual como se verá más adelante contiene la variabilidad más alta explicada del vector $\mathbf{x}$.

3 Componentes Principales

Para el vector $\mathbf{x}$ que contiene las variables proxy a utilizar para construir el indicador, con vector de medias $\boldsymbol{\mu} \in \mathbb{R}^p$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, definimos la varianza total de $\mathbf{x}$ como:

$$E[(\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu})] = \sum_{i=1}^{p} \sigma_{ii} = \text{traza}\boldsymbol{\Sigma}. \tag{1}$$

Una propiedad de la matriz $\boldsymbol{\Sigma}$ es que es simétrica por lo que existe una matriz ortogonal $\mathbf{H}$ tal que $\mathbf{H}^\top \boldsymbol{\Sigma} \mathbf{H} = \mathbf{D}$, al ser $\mathbf{H}$ ortogonal sus vectores columnas $\mathbf{H} = (\mathbf{h}_1 \quad \mathbf{h}_2 \ldots \mathbf{h}_p)$ forman una base ortonormal donde cada uno de los $\mathbf{h}_i$ se denominan vectores propios, mientras que $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ contiene en su diagonal los valores propios de $\boldsymbol{\Sigma}$ asociados a los vectores propios, adicionalmente la matriz de varianzas y covarianzas es semidefinida positiva por lo que sus valores propios son mayores o iguales a cero. Definiendo el vector

$$\mathbf{y} = \mathbf{H}^\top \mathbf{x} = \begin{pmatrix} \mathbf{h}_1^\top \mathbf{x} \\ \mathbf{h}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{h}_p^\top \mathbf{x} \end{pmatrix}$$

como combinación lineal del vector $\mathbf{x}$, se tiene que la $Var(\mathbf{y}) = Var(\mathbf{H}^\top \mathbf{x}) = \mathbf{H}^\top \boldsymbol{\Sigma} \mathbf{H} = \mathbf{D}$, siendo la varianza total de $\mathbf{y}$ igual a $\sum_{i=1}^{p} \lambda_i$. Realizando las operaciones matriciales correspondientes se determina que $\mathbf{x}$ y $\mathbf{y}$ tienen igual varianza

total. Las variables $y_i = \mathbf{h}_i^\top \mathbf{x}$ son por definición las componentes principales de $\mathbf{x}$, destacando además que las variables y_i son no correlacionadas (Bilodeau y Brenner, 1999).

Las componentes principales pueden ser vistas como una rotación de los ejes de las variables originales a nuevos ejes ortogonales entre sí, de tal manera que la primera componente defina la dirección del eje con la máxima varianza de los datos, la segunda componente defina un nuevo eje ortogonal al de la primera componente maximizando la varianza de los datos en esa dirección, así hasta obtener las p componentes. La dirección de estos nuevos ejes la dan los vectores propios asociados a $\boldsymbol{\Sigma}$, y la distancia que hay entre el punto medio del vector $\mathbf{x}$ y la proyección ortogonal de las observaciones sobre los ejes definidos por los vectores propios, dan lugar a los vectores de observaciones denominados scores. La varianza de cada variable asociada al vector de componentes principales $\mathbf{y}$ viene siendo la diagonal principal de la matriz $\mathbf{D}$ que contiene los valores propios de $\boldsymbol{\Sigma}$, de manera que la primera componente y_1 tiene como varianza al valor propio de más alto valor λ_1, la varianza de la segunda componente y_2 es λ_2 y así para cada componente principal y_i, al final vamos a tener que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. De tal manera que si seleccionamos las k primeras componentes principales y calculamos la razón $\frac{\sum_{i=1}^{k} \lambda_i}{\text{traza}\boldsymbol{\Sigma}}$, si esta es cercana a uno las k primeras componentes pueden reemplazar a $\mathbf{x}$ sin perder mucho en términos de la variabilidad total explicada.

Para obtener los valores propios de la matriz $\boldsymbol{\Sigma}$ se debe satisfacer que $\boldsymbol{\Sigma}\mathbf{H} = \mathbf{HD}$, lo que es equivalente a resolver $|\boldsymbol{\Sigma} - \lambda\mathbf{I}| = 0$ con lo que se obtendrá una ecuación polinomial de grado p, siendo los valores propios las raíces de esa ecuación. La matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ casi siempre es desconocida por lo que las componentes principales son determinadas a partir de una muestra de la que se calcula la matriz de varianzas y covarianzas $\mathbf{S}$, dada una muestra de tamaño n, dicha matriz se define como:

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top. \tag{2}$$

De la matriz $\mathbf{S}$ obtenemos los valores y vectores propios como estimadores de los valores y vectores propios de $\boldsymbol{\Sigma}$, quedando $\widehat{\mathbf{H}}^\top \mathbf{S}\widehat{\mathbf{H}} = \widehat{\mathbf{D}} = \text{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_p)$ donde la matriz $\widehat{\mathbf{H}} = (\widehat{\mathbf{h}}_1 \quad \widehat{\mathbf{h}}_2 \ldots \widehat{\mathbf{h}}_p)$. Las componentes principales muestrales de $\mathbf{x}$ se definen como $\widehat{y_i} = \widehat{\mathbf{h}_i}^\top \mathbf{x}$.

4 Índice de riqueza

Para el índice de riqueza según Filmer y Pritchett (2001) se utiliza la matriz de varianzas y covarianzas de las variables proxy que podrían estar más correlacionadas con la riqueza del hogar, la primera componente principal calculada se la utiliza como la variable que representará a la variable latente riqueza del hogar, quedando

de la siguiente manera:

$$\widehat{y_1} = \widehat{h}_1 x_1 + \widehat{h}_2 x_2 + \ldots \widehat{h}_p x_p. \tag{3}$$

Tal como se expresa en la ecuación (3), las ponderaciones para las variables proxy que conforman el índice de riqueza vienen dadas por los elementos del primer vector propio de la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$, tomando en cuenta que el primer vector propio corresponde al valor propio más alto, aquellas variables proxy que tengan la mayor variabilidad tendrán un peso mayor aportando al indicador, de aquí que si las variables proxy se encuentran en diferentes escalas esto ocasionará que los mayores pesos sean para aquellas variables cuyo rango de valores sea mayor, una solución a este problema es utilizar todas las variables en una misma escala o realizar el proceso de estandarización de variables restando a cada observación la media y dividiendo para la desviación estándar, estandarizar las variables es equivalente a calcular las componentes principales a partir de la matriz de correlaciones $\mathbf{R}$ de los datos muestrales quedando definida como:

$$\mathbf{R} = \mathbf{S}_0^{-1/2} \mathbf{S} \mathbf{S}_0^{-1/2},$$

donde $\mathbf{S}$ es la matriz de varianzas y covarianzas muestral y $\mathbf{S}_0 = \text{diag}(s_{11}, s_{22}, \ldots, s_{pp})$, al ser $\mathbf{R}$ diagonalizable es posible descomponerla en $\mathbf{R} = \widehat{\mathbf{G}}\text{diag}(f_1, f_2, \ldots, f_p)\widehat{\mathbf{G}}^\top$, donde f_i son los valores propios de la matriz de correlaciones, definiendo las componentes principales para las variables estandarizadas $\mathbf{z} = \mathbf{S}_0^{-1/2}(\mathbf{x} - \bar{\mathbf{x}})$ como $\widehat{y_i} = \widehat{\mathbf{g}}_i^\top \mathbf{z}$, la primera componente principal queda:

$$\widehat{y_1} = \widehat{g}_1 \left(\frac{x_1 - \bar{x}_1}{s_1} \right) + \widehat{g}_2 \left(\frac{x_2 - \bar{x}_2}{s2} \right) + \ldots + \widehat{g}_p \left(\frac{x_p - \bar{x}_p}{s_p} \right). \tag{4}$$

Como se observa en la ecuación (4), al determinar las componentes basadas en la matriz de correlación de los datos, las variables proxy son estandarizadas con lo que se tiene a todas en una misma escala.

El mejor uso de componentes principales se da cuando los datos son continuos, sin embargo, para el caso del indicador de Filmer y Pritchett (2001) las variables proxy son categóricas y para una fácil interpretación de los resultados las variables proxy se las define como dicotómicas, para el caso de posesión de bienes se toma como 0 el hecho de no poseer el bien y 1 el hecho de poseerlo, de tal manera que si un hogar posee dicho bien el índice de riqueza se incrementará de acuerdo al peso asignado por la primera componente principal, ahora, si la variable proxy tiene más de dos categorías como el material del piso de la vivienda o la fuente de suministro de agua para la vivienda, Filmer y Pritchett (2001) dieron como solución la construcción de variables dummy, es decir si una variable tiene r categorías se construyen $r - 1$ variables dicotómicas denominadas dummy, de tal manera que al final la construcción del índice se realiza con todas las variables de tipo dicotómico. Al crear variables dummys existe un inconveniente de colinealidad entre las variables dummys correspondientes a una misma variable categórica, puesto que si una

de las variables dummys toma el valor de 1 automáticamente las otras tomarían el valor de 0, generando una alta correlación entre ellas, se sabe que al construir las componentes principales la primera componente es la que recoge la mayor variabilidad de los datos originales, esto quiere decir que los pesos asignados a las variables dummys sean altos debido a que la correlación es alta entre ellas. Una ilustración de este análisis lo puede encontrar en Ward (2014); Oyedepo, Lasabi, y Adekanmbi (2019).

Para evitar la colinealidad de variables dummys se define a la variable categórica como una variable ordinal, asignando el orden de cada categoría de tal manera que valores altos de la variable signifiquen mayor riqueza del hogar y valores bajos, menor riqueza. De esta manera se realizaría el cálculo de componentes principales con variables ordinales. Adicional a esto, el uso de variables discretas en técnicas donde se espera se utilicen datos continuos viola los supuestos de la distribución de los datos (Kolenikov y Angeles, 2009), al trabajar la construcción del índice con variables ordinales, asumimos que la distancia existente entre categorías de una variable es la misma y la matriz de correlación se calcula en base a estos datos. Pearson en 1922 desarrolló el concepto de correlación policórica como estimador de máxima verosimilitud de dos variables continuas no observadas distribuidas como una normal basada en la observación de la versión discretizada de las mismas, de esta manera se tendría una matriz de correlaciones policóricas de variables ordinales asumiendo que son variables discretizadas de variables no observadas normales. Mayor detalle sobre correlación policórica se encuentra en Pearson y Pearson (1922). El índice de riqueza se lo suele utilizar como variable de explicación de otras variables como la prevalencia de enfermedades crónicas de acuerdo al nivel de riqueza de la población (Vukovic, Bjegovic, y Vukovic, 2008).

5 Ilustración

Para ilustrar la construcción del índice de riqueza se realizará un cálculo con datos de la Encuesta Nacional de Salud y Nutrición ENSANUT del año 2018 realizada para la población ecuatoriana por parte del Instituto Nacional de Estadística y Censos (INEC). La cobertura de esta encuesta es nacional y es dirigida a hogares, el tamaño de la muestra es de 46638 viviendas (INEC, 2018).

Esta encuesta contiene datos de características de la vivienda y de la posesión de activos en el hogar, para el indicador de riqueza se tomarán en cuenta 18 variables divididas en 13 variables correspondiente a la posesión de bienes o servicios en el hogar, 2 variables de acceso a servicios básicos en el hogar, y 3 variables sobre características de la estructura de la vivienda. Las variables relacionadas a la posesión de bienes o servicios son dicotómicas, es decir, 0 si el hogar no posee el bien o servicio y 1 si lo posee; para las características del hogar y el acceso a servicios básicos se las define como variables ordinales teniendo entre tres y cuatro categorías cada una. El análisis de componentes principales será realizado considerando la matriz de correlación de este vector de variables.

El listado de las variables, así como la proporción de cada categoría presente en la muestra se presenta en la tabla 1. Cada una de las variables están en una escala ordinal, de tal manera que el material del piso de la vivienda de categoría más baja es tierra luego caña o tabla, cemento o ladrillo y la más alta es cerámica o baldosa; el material del techo de la vivienda tiene como categoría más baja techo de palma o paja, luego zinc, asbesto y la más alta teja, hormigón o losa. Para el tipo de servicio higiénico la categoría más baja es no tener servicio; luego están los hogares que poseen letrina, seguido de excusado con pozo séptico o ciego y la categoría más alta aquellos que lo tienen, pero con red de alcantarillado; el número de cuartos para dormir en el hogar se codifican en tres categorías la más baja es tener un cuarto, la siguiente dos cuartos y la más alta tener tres o más cuartos.

Tabla 1: Proporción de Bienes o Servicios

Bien o servicio	**Proporción**			
	No	**Sí**		
Refrigeradora	0.157	0.843		
Computadora	0.604	0.396		
Lavadora	0.444	0.556		
Licuadora	0.198	0.802		
Horno, Microondas	0.703	0.297		
Plancha	0.255	0.745		
Televisor	0.100	0.900		
DVD	0.638	0.362		
Calefón	0.890	0.110		
Línea telefónica	0.669	0.331		
Automóvil	0.769	0.231		
Acceso a Internet	0.583	0.417		
Acceso a Televisión por cable	0.648	0.352		
Acceso a servicio de agua por tubería	0.082	0.918		
	Tierra	**Caña/ tabla**	**Cemento/ ladrillo**	**Cerámica/ baldosa**
Material del piso de la vivienda	0.034	0.112	0.366	0.488
	Palma/ Paja	**Zinc**	**Asbesto**	**Teja/losa/ hormigón**
Material del techo de la vivienda	0.006	0.437	0.120	0.436
	No tiene	**Letrina**	**Pozo**	**Alcantarillado**
Tipo de servicio higiénico	0.026	0.014	0.304	0.656
	Un cuarto	**Dos cuartos**	**Tres o más cuartos**	
Número de cuartos para dormir	0.084	0.201	0.715	

Fuente: Elaboración propia

Para estas variables se calcula la matriz de correlación, la figura 1 presenta un gráfico de la correlación por colores entre cada par de variables, la escala va desde el color azul oscuro cuando la correlación es negativa y rojo oscuro cuando es

positiva, un color cercano al blanco indica una baja correlación. En la gráfica todas las correlaciones son positivas, entre las correlaciones altas se encuentran el tener una computadora y tener acceso a internet, también el tener una línea telefónica en el hogar con tener acceso a internet, otra relación positiva se da en el hecho de que mientras mejor sea el piso de la vivienda, mejor será el tipo de servicio higiénico que tenga el hogar.

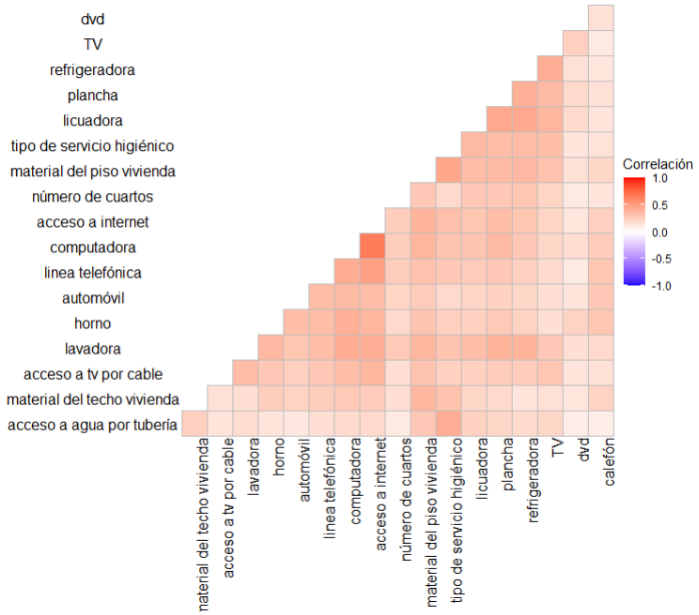

Fig. 1: Gráfico de matriz de correlaciones
Fuente: Elaboración propia

A partir de esta matriz de correlaciones calculamos las componentes principales para las 18 variables proxy utilizando el software R (Everitt y Hothorn, 2011), en la tabla 2 se muestran los coeficientes de la primera componente principal o pesos de las variables proxy, dada esta matriz, el índice de riqueza quedaría de la siguiente manera:

$$IR = \widehat{g}_1\left(\frac{x_1 - \bar{x}_1}{s_1}\right) + \widehat{g}_2\left(\frac{x_2 - \bar{x}_2}{s_2}\right) + \ldots + \widehat{g}_{18}\left(\frac{x_{18} - \bar{x}_{18}}{s_{18}}\right). \tag{5}$$

Dado que las variables proxy son ordinales, y dicotómicas en algunos casos, la interpretación del índice se realiza en función de la presencia o ausencia del bien en el hogar o el incremento del tipo de material del piso, techo de la vivienda o el número de cuartos. Debido a que las componentes principales se determinaron con la matriz de correlaciones, para calcular el incremento en el índice al poseer un bien x_i en el hogar se tiene que considerar la razón entre el coeficiente y la desviación

estándar de la variable proxy $\frac{\widehat{g_i}}{s_i}$, de esta manera de acuerdo a los datos de la tabla 2, el tener refrigeradora en el hogar aumenta en 0.619 el índice de riqueza del hogar frente a uno que no lo tenga, si se tiene computadora el índice incrementa en 0.606 y si adicionalmente a tener computadora se cuenta con acceso a Internet en el hogar el índice incrementaría en 1.21; para las variables con más de dos categorías el incremento se da en relación al paso de una categoría a otra, así para el material del piso de la vivienda un hogar que tenga cerámica o baldosa en el piso de la vivienda incrementa su índice de riqueza en 0.999 frente a un hogar con piso de tierra en la vivienda, caso similar el tipo de servicio higiénico donde contar con excusado y alcantarillado en el hogar incrementa en 1.005 el índice de riqueza frente a un hogar que no tiene servicio higiénico dentro de la vivienda.

Tabla 2: Coeficientes primera componente principal

Variable $\mathbf{x}_i$	**Coeficiente** $\widehat{g_i}$	**Desviación estándar** s_i	$\widehat{g_i}/s_i$
Refrigeradora	0.249	0.402	0.619
Computadora	0.288	0.476	0.606
Lavadora	0.272	0.499	0.545
Licuadora	0.252	0.421	0.599
Horno, Microondas	0.241	0.420	0.575
Plancha	0.263	0.463	0.567
Televisor	0.220	0.345	0.637
DVD	0.139	0.471	0.295
Calefón	0.168	0.282	0.597
Línea telefónica	0.253	0.449	0.563
Automóvil	0.218	0.401	0.544
Acceso a Internet	0.291	0.481	0.604
Acceso a Televisión por cable	0.223	0.480	0.465
Acceso a servicio de agua por tubería	0.169	0.299	0.565
Material del piso de la vivienda	0.280	0.839	0.333
Material del techo de la vivienda	0.195	0.943	0.207
Tipo de servicio higiénico	0.253	0.757	0.335
Número de cuartos para dormir	0.197	0.663	0.298

Fuente: Elaboración propia

Con el índice de riqueza calculado para cada hogar en la muestra, es posible determinar quintiles que representen el bienestar de los individuos que conforman el hogar, así el primer quintil corresponde a los hogares con el 20 % más bajo de índice de riqueza y el quinto quintil corresponde a los hogares con el 20 % mayor de este índice. Con los quintiles es posible realizar una validación interna con las variables proxy utilizadas, para eso utilizamos cuatro variables: si posee refrigeradora en el hogar, si posee computadora, el tipo de servicio higiénico en la vivienda y el material del piso de la vivienda, en la figura 2 se presenta los gráficos bivariados de cada variable con los quintiles de riqueza determinados con el indicador, en los gráficos se observa que para los hogares pertenecientes al primer quintil poco más del

60 % de los hogares no tienen refrigeradoras en su vivienda, casi el 100 % no tiene computadora, el 10 % no tiene servicio higiénico en la vivienda y más del 50 % de los hogares tienen el piso de su vivienda de caña, tabla o tierra. De manera general se puede observar que a mayor quintil de riqueza el porcentaje de hogares que poseen bienes es mayor, así como será mejor la infraestructura de la vivienda; para el grupo que se encuentra en el quinto quintil casi en su totalidad tienen refrigeradora y computadora en el hogar, así como casi el 100 % tiene excusado y alcantarillado y en más del 90 % el piso de la vivienda es de cerámica o baldosa.

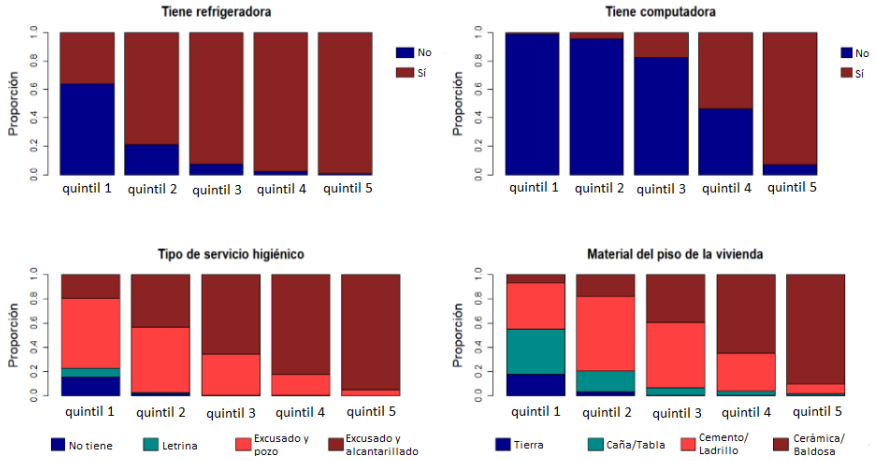

Fig. 2: Gráficos bivariados: Quintiles – Variables proxy
Fuente: Elaboración propia

Con el índice es posible determinar desigualdades en cuanto a la distribución de la riqueza dentro de subpoblaciones de la población utilizada para la construcción del índice, McKenzie (2005) propuso una medida de inequidad en subpoblaciones basada en la desviación estándar de las subpoblaciones y la varianza de la primera componente principal, la medida se define como:

$$I_s = \frac{\sigma_s}{\sqrt{\lambda}}. \tag{6}$$

La desviación estándar de cada subpoblación se estima con la desviación estándar muestral, de esta manera un índice con valor mayor que 1 indicará que la subpoblación s presenta mayor inequidad en cuanto a distribución de la riqueza en comparación con la población. Con está razón es posible también realizar comparaciones entre subpoblaciones. Para ilustrar su uso, se medirá inequidad para las provincias del Ecuador, recordar que los datos para la construcción del índice fueron tomados de la ENSANUT 2018, para el diseño muestral de dicha encuesta se definió que el dominio de estudio sea a nivel provincial de tal manera que es posible inferir

resultados para dichas subpoblaciones (INEC, 2018). En la figura 3, se muestra el gráfico con el índice de inequidad por provincia, de las 25 provincias en 8 la inequidad en distribución de riqueza es mayor que a nivel nacional, 4 son provincias de la región Oriental: Pastaza, Orellana, Napo y Morona Santiago, siendo esta última la de mayor inequidad entre todas; las siguientes 4 pertenecen a la región Sierra: Loja, Cotopaxi, Chimborazo y Bolívar. Entre las provincias con menor inequidad están Galápagos, El Oro y Los Ríos. Si comparamos la provincia Guayas con Pichincha, la última tiene menor inequidad con relación a Guayas.

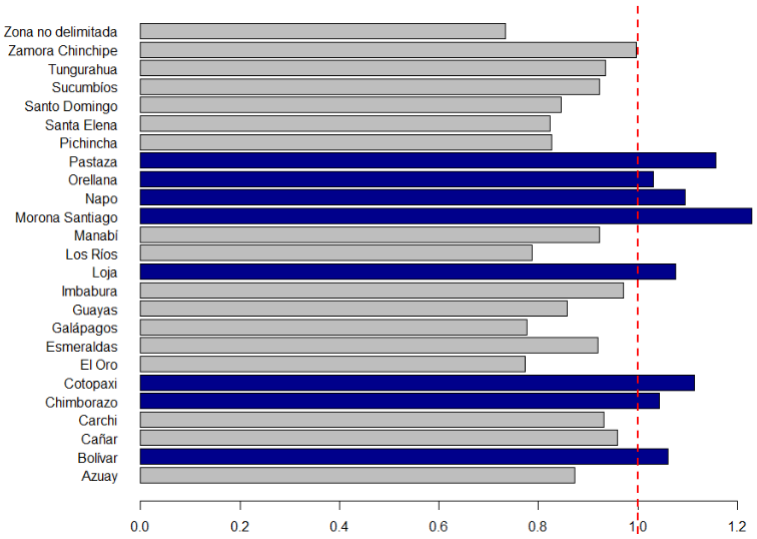

Fig. 3: Índice de Inequidad por provincia
Fuente: Elaboración propia

6 Conclusiones

A manera de conclusiones, ante la ausencia de datos acerca de ingresos económicos o consumos en el hogar o la mala calidad en la recolección de los mismos, una alternativa para medir la riqueza es el uso de variables de fácil medición y que se asocien a la riqueza en el hogar como la posesión de bienes o la infraestructura de la vivienda. Las componentes principales se pueden determinar a partir de la matriz de varianzas y covarianzas de los datos o a partir de la matriz de correlaciones, si se utiliza la primera hay que tomar en cuenta la escala en la que estén definidas las variables. Hay que tomar en cuenta que el índice de riqueza por componentes principales mide la desigualdad entre los hogares de la población y los quintiles se definen

para dicha población, de tal manera que la comparación de índices para dos o más poblaciones no es fiable puesto que estar en el quintil más alto en una población no indica que también pertenecerá al quintil más alto en otra ya que los índices dependerán de la variabilidad de cada población, si se desea realizar un análisis de este tipo se sugiere unificar las poblaciones en una sola de tal manera que los coeficientes de las variables proxy expliquen la variabilidad entre las poblaciones.

7 Bibliografía

Referencias

Bilodeau, M., y Brenner, D. (1999). *Theory of multivariate statistics*. Springer.

Córdova, A. (2008). Nota metodológica: midiendo riqueza relativa utilizando indicadores sobre bienes del hogar. *Perspectivas desde el Barómetro de las Américas*.

Everitt, B., y Hothorn, T. (2011). *An introduction to applied multivariate analysis with r*. Springer Science & Business Media.

Filmer, D., y Pritchett, L. H. (2001). Estimating wealth effects without expenditure data–or tears: an application to educational enrollments in states of India. *Demography*, *38*(1), 115–132.

Howe, L. D., Hargreaves, J. R., y Huttly, S. R. (2008). Issues in the construction of wealth indices for the measurement of socio–economic position in low–income countries. *Emerging themes in epidemiology*, *5*(1), 3.

INEC. (2018). Documento metodológico de la Encuesta Nacional de Salud y Nutrición ENSANUT. Ecuador.

Kolenikov, S., y Angeles, G. (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, D. J. (2005). Measuring inequality with asset indicators. *Journal of Population Economics*, *18*(2), 229–260.

Oyedepo, E., Lasabi, O., y Adekanmbi, A. (2019). Determinants of wealth status among rural and urban households in Nigeria. *Journal of Studies in Social and Humanities*, 68–77.

Pearson, K., y Pearson, E. S. (1922). On polychoric coefficients of correlation. *Biometrika*, 127–156.

Vukovic, D., Bjegovic, V., y Vukovic, G. (2008). Prevalence of chronic diseases according to socioeconomic status measured by wealth index: Health survey in Serbia. *Croatian medical journal*, 832–841.

Ward, P. (2014). Measuring the level and inequality of wealth: An application to China. *Review of Income and Wealth*, 613–635.