UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Instituto de Estadística

Como reconstruir el INSE en una encuesta sanitaria poblacional

Ramón Alvarez - Andrés Castrillejo
Febrero-2015

Documentos de Trabajo

Serie DT (15 / 01) - ISSN : 1688-6453

Como reconstruir el INSE en una encuesta sanitaria poblacional

Ramón Alvarez [1]
Instituto de Estadística - Facultad de Ciencias Económicas y de Administración - UdelaR.

Andrés Castrillejo [2]
Instituto de Estadística - Facultad de Ciencias Económicas y de Administración - UdelaR.

RESUMEN

En las encuestas sanitarias de base poblacional es importante poder estratificar las variables a estudiar de acuerdo a características socioeconómicas y demográficas. El nivel socioeconómico no es una variable observada, sino una construcción a partir de una metodología validada para Uruguay, conocida como INSE y considerada de referencia a nivel nacional en todo tipo de estudios poblacionales fundamentalmente en las disciplinas sociales.

En el año 2009 en Uruguay se llevó a cabo la Encuesta Nacional de Tabaquismo en Adultos (ENTA) por parte del Programa de Tabaco del Ministerio de Salud Pública a través de una muestra de hogares con diseño complejo. En esta encuesta algunas de las variables utilizadas para la construcción del INSE no se relevaron, por lo cual se procura la construcción de un indicador para aproximarlo.

Se propone una construcción alternativa que permite asignar un INSE a los hogares a través de métodos de clustering; como alternativa se busca una solución planteándolo como un problema clásico de clasificación a partir de datos de las Encuestas de Hogares (ECH) del INE.

Por otra parte se intenta otra solución a través del uso de medidas de disimilaridad aplicadas sobre variables binarias, algunas de las cuales tienen en cuenta la asimetría de la distribución.

Se presentan los resultados comparando entre sí las diferentes aproximaciones.
Palabras claves: Clustering, Encuestas sanitarias, Indice de nivel socioeconómico, Medidas de disimilaridad.

[1] ramon@iesta.edu.uy
[2] andres@iesta.edu.uy

1. Introducción

La metodología y los resultados que se presentan en este documento de trabajo, fueron presentados como un primer avance en las jornadas académicas de 2011. Con posterioridad se continuó trabajando en el problema, lo que permite mostrar lo que se consigna a continuación.

En el año 2009 en Uruguay se llevó a cabo la Encuesta Nacional de Tabaquismo en Adultos (ENTA) por parte del Programa de Tabaco del Ministerio de Salud Pública a través de una muestra de hogares con diseño complejo.

En las encuestas sanitarias de base poblacional es importante poder estratificar las variables a estudiar de acuerdo a características socioeconómicas y demográficas. Una forma aproximada y sintética de hacerlo es a través del uso de una variable resumen o índice que mida el " nivel socioeconómico" de las personas. El nivel socioeconómico no es una variable observada, sino una construcción a partir de una metodología validada para Uruguay, conocida como Indice de Nivel Socioeconómico, INSE, y considerada de referencia a nivel nacional en estudios poblacionales.

En la ENTA algunas de las variables utilizadas para la construcción del INSE no se relevaron y este trabajo procura una reconstrucción que aproxime adecuadamente al INSE.

En la siguentes subsecciones se plantea el problema concreto y se describe la encuesta ENTA y el índice INSE. Luego en la sección 2 se describen las aproximaciones y algunos de los métodos utilizados en la resolución del problema. En la sección 3 se presentan los resultados de las aplicaciones y finalmente en la sección 4 se establecen algunas conclusiones y recomendaciones.

1.1. Encuesta Nacional de Tabaquismo en Adultos

El consumo de tabaco es la principal causa prevenible de muerte prematura y enfermedad. Es responsable de más de 5 millones de muertes por año en el mundo (Ministerio de Salud Pública, 2010).

Desde la Organización Mundial de la Salud (OMS) se alienta a los países a adherir a un convenio para el control del tabaco, estableciendo mecanismos de monitoreo de la epidemia.
La Encuesta Mundial de Tabaco en Adultos (*Global Adult Tobacco Survey*, GATS) es una encuesta de hogares que se incluyó en el año 2007 como un nuevo componente del Sistema Mundial de Vigilancia de Tabaco (Global Tobacco Surveillance System, GTSS). La

encuesta GATS permite a los países recolectar datos clave para establecer medidas para el control del tabaco en toda la población adulta. Sus resultados darán apoyo a los países en la formulación e implementación de intervenciones efectivas para el control del tabaco (Tobacco Free Initiative (TFI), 2010).

Inicialmente la encuesta GATS fue diseñada para ser implementada en los países con mayor número absoluto de fumadores del mundo: Bangladesh, Brasil, China, Egipto, Filipinas, India, México, Polonia, Federación Rusa, Tailandia, Turquía, Ucrania y Vietnam. Por sus características demográficas Uruguay no integraba inicialmente el grupo de países seleccionados, pero dado que demostró un fuerte compromiso con el Control del Tabaco y ha hecho progresos significativos desde la ratificación del Convenio Marco de la OMS para el Control del Tabaco (CMCT)(WHO, 2004) en setiembre del 2004 fue invitado a participar. Se definió que el Ministerio de Salud Pública (MSP) sería la agencia coordinadora, y el Instituto Nacional de Estadística (INE) la agencia implementadora de la encuesta que se denominó ENTA 2009. Para realizar un seguimiento de las diferentes etapas de la implementación del estudio se integró un Comité Coordinador compuesto por representantes de diferentes instituciones que trabajan en el control del tabaco: Ministerio de Salud Pública (Departamento de Vigilancia en Salud y Programa Nacional para Control del Tabaco), Instituto Nacional de Estadística, Organización Panamericana de la Salud (OPS), Facultad de Medicina de la Universidad de la República, Hospital Universitario y Comisión Honoraria de Lucha Contra el Cáncer (CHLCC).
Los objetivos de la encuesta ENTA eran:

- Monitorear sistemáticamente el consumo de tabaco (fumado y sin humo) en población de 15 años o más y ciertos indicadores clave, en una muestra representativa a nivel nacional.

- Realizar el seguimiento de la implementación de las políticas de control de tabaco recomendadas en el CMCT y delineadas en el paquete MPOWER (Organización Mundial de la Salud, 2008).

Para ese objetivo se trabajó con un diseño muestral multietápico complejo en 4 etapas que permitía tener estimaciones con niveles de desagregación por región (urbano /rural) y por género. La población objetivo eran todos las personas de 15 o más años viviendo en hogares que consideran a Uruguay como su primer país de residencia.
El marco muestral usado se construyó a partir del Censo Nacional de Población Fase 1 (CF1) (Instituto Nacional de Estadística, 2005b) donde se manejaron listados de viviendas, hogares y personas que residen habitualmente en Uruguay. Con esta información se crearon las unidades primarias y secundarias de muestreo (PSU) y (SSU) para llegar a unidades censales a nivel geográfico más detallado, llamado zona censal y que eran las unidades terciarias de muestreo (TSU). Las TSU estaban compuestas por las viviendas,

donde finalmente se determinaban en forma aleatoria las personas de 15 años o más a ser encuestadas. Se trabajó en este diseño con una estratificación en 10 estratos (9 urbanos y 1 rural). En las diferentes etapas se consideraron probabilidades de inclusión generadas con $\pi - ps$ (probabilides proporcionales al tamaño del número de unidades para cada nivel de desagregación). El tamaño de muestra alcanzado fue de 5500 personas, de un total de 6600 originalmente seleccionadas. Tal como estaba preestablecido en el diseño de muestreo, los datos debieron ser calibrados por el método *rake* (Särndal et al., 1992), (Särndal y Lundström, 2005), (Lumley, 2010), usando variables de tipo sociodemográficas y usando los totales poblacionales que surgían de la ECH ampliada 2006.

Descripción del método de calibrado *rake*
Este método equivale a efectuar una post-estratificación incompleta, o lo que es equivalente, a calibrar sobre las marginales conocidas de una tabla poblacional tomada como referencia. Se considera para eso una tabla de dos dimensiones con las cantidades observadas en las celdas, n_{ij}, las cantidades poblacionales desconocidas de las celdas N_{ij}, y sus estimadores $\hat{N}_{ij}$. Las marginales $\sum_j N_{ij} = N_{i+}$ y $\sum_i N_{ij} = N_{+j}$ son conocidas. El procedimiento del *rake* se aplica a las cantidades individuales n_{ij} para iterativamente calcular estimaciones que satisfagan las restricciones marginales $N^*_{i+} = \sum_j N^*_{ij} = N_{i+}$ y $N^*_{+j} = \sum_i N^*_{ij} = N_{+j}$ utilizando una serie de constantes multiplicativas de las filas, a_i, y de las columnas, b_j tal que $N^*_{ij} = a_i b_j n_{ij}$.
El ajuste iterativo proporcional es utilizado para ajustar las celdas al total de las marginales. Como un primer paso del procedimiento, los estimadores son calculados como $N^{(1)}_{ij} = n_{ij} N_{i+}/n_{i+}$. Esto hace que las marginales de las filas estimadas se ajusten exactamente al verdadero valor, pero no sucede lo mismo con las marginales de las columnas. La siguiente iteración ajusta las celdas individuales a las marginales de las columnas $N^{(2)}_{ij} = N^{(1)}_{ij} N_{+j}/N^{(1)}_{ij}$. Y luego las marginales de las filas se ajustan por $N^{(3)}_{ij} = N^{(2)}_{ij} N_{+j}/N^{(2)}_{ij}$. La iteración entre filas y columnas continúa hasta que se llega a la convergencia, esta última se define como $\left|N^*_{i+} - N_{i+}\right| < \epsilon$ y $\left|N^*_{+j} - N_{+j}\right| < \epsilon$ para algún valor pequeño de ϵ.

Como toda encuesta sanitaria de base poblacional era imprescindible tener alguna forma de caracterización socioeconómica de las personas encuestadas, para poder profundizar en el análisis epidemiológico de la información. Lamentablemente en esta encuesta en la aplicación del cuestionario sociodemográfico solo se relevaron algunas de las variables necesarias para el cálculo de ambas versiones del INSE, por lo cual se debió construir un *INSE* ad-hoc con la información disponible.

1.2. Indice de Nivel Socio-Económico

El INSE es un índice creado a través de un modelo de respuesta discreta *logit*, donde se explica el ingreso de los hogares de un microcenso, a partir de un conjunto de variables explicativas que se detallan a continuación (Fernandez y Perera, 2003),(CAINSE, 2007).

- Ocupación del jefe del hogar - Cantidad de Perceptores de ingreso - Nivel educativo del jefe del hogar
- Tenencia de tarjeta de créditos
- Número de baños de la vivienda
- Tenencia de electrodomésticos (TV, DVD, Refrigerador, Aire Acondicionado...)

Para el INSE existen 2 variantes:

- $INSE = \sum_1^{18} \alpha_i X_i$ que toma en cuenta 18 variables, donde los α_i son los ponderadores que surgen del modelo logit antes mencionado.
- $INSEred = \sum_1^{9} \beta_i X_i$ que consiste en un índice reducido donde solo se toman en cuenta 9 variables.

1.3. El problema

El conjunto de preguntas de la ENTA no incluye todas la que permiten construir el INSE (ni el INSE reducido). La matriz de datos de tipo sociodemográfico con la que se cuenta para la encuesta ENTA tiene solamente la información que corresponde a *Nivel educativo de la persona encuestada* (4 tramos) y 17 variables binarias de tenencia de bienes que aparecen en la tabla que sigue:

1	Baño con cisterna	10	No Calentador instantáneo de agua
2	Teléfono	11	Conexión a tv por abonados
3	Televisión	12	Reproductor de DVD
4	Radio	13	Lavavajillas
5	Refrigerador	14	Horno microondas
6	Automóvil	15	Equipo de aire acondicionado
7	Lavadora automática	16	Tiene Computadora Plan Ceibal
8	Secador de ropa	17	Conexión a internet
9	Calefón o termofón		

Tabla 1: Tabla

La matriz de datos sobre la que se debe trabajar es de la forma

```
 > head(Gats)
   E X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17
1  1  1  0  1  1  0  0  0  1  1   0   1   0   0   0   0   0   1
2  3  1  1  1  1  0  1  0  1  1   1   0   0   1   0   0   0   1
.....................................................................
4  2  1  0  0  1  0  1  0  1  1   0   1   0   0   0   1   1   1
6  4  0  1  1  1  0  1  0  1  1   1   1   0   1   0   0   0   1
```

2. Aproximaciones para la Reconstrucción

A continuación se presentan las alternativas metodológicas que se exploraron para la resolución del problema planteado en la sección 1.3.

2.1. Análisis Factorial

Sobre las 17 variables relativas al confort del hogar o tenencia de bienes se aplicó un Análisis Factorial de Correspondencias Múltiples (ACM). El ACM es una técnica descriptiva multivariante que mediante un proceso algebraico logra una simplificación del problema al crear nuevas variables, que se denominan factores y que son combinaciones lineales de las variables relevadas, adjudicando a cada variable un peso o importancia a través de un coeficiente estimado (coordenada de la variable en el subespacio vectorial donde se proyectan cada una de ellas). Los factores finalmente considerados pueden ser representados gráficamente en diagramas de dispersión que se llaman planos factoriales. Como resultado se consideran, luego de usar el ajuste de Benzecri ,2 ejes factoriales (Blanco, 2006).

2.2. Regresión Logística Multinomial

Se trató de construir sobre la Encuesta Continua de Hogares (ECH) (Instituto Nacional de Estadística, 2005a),(Instituto Nacional de Estadística, 2009), un modelo alternativo al INSE tratando de explicar el ingreso categorizado en tramos a partir de las 17 variables de confort que estaban disponibles en la ENTA y en la ECH usando para eso regresión logística politómica o multinomial.

- Modelo *logit* con variable de respuesta Ingreso (RI_3) en 3 tramos, considerando el primer quintil como categoría mas baja, los 3 siguientes quintiles como categoría de ingreso medio y el último quintil correspondiente a los ingresos altos. El modelo tuvo una performance de un 70 % de acierto.

Con los modelos ajustados mediante las diferentes técnicas se hicieron predicciones que luego se cruzaron con los clusters creados mediante el métodos jerárquico de Ward sobre los factores creados con el ACM.

2.3. Random Forests

Alternativamente al análisis de cluster se planteó el problema como un problema clásico de clasificación a partir de datos de las Encuestas de Hogares del INE.
A través de una *muestra de entrenamiento* ($\mathcal{L}$) que era en este caso la ECH para el segundo semestre de 2009, con las 17 variables de confort relevadas en la encuesta ENTA y como variable de respuesta el Ingreso con diferentes estratificaciones se aplicó *Random Forests* (Breiman, 2001) que es un método de agregación de árboles de clasificación.
Random Forests consiste en la agregación de muchos árboles de decisión aleatorizados (en cada nodo se particiona a partir de un número pequeño de variables seleccionadas aleatoriamente), construidos sobre remuestras *bootstrap* del conjunto de entrenamiento ($\mathcal{L}$). El clasificador resultante se obtiene por voto mayoritario en el conjunto de árboles.

2.4. Clustering

Se propuso una construcción alternativa que permitía asignar un INSE a los hogares a través de métodos de clustering, poniendo especial atención en el uso de medidas de disimilaridad aplicadas sobre variables binarias que habitualmente se usan en ecología (Legendre, P. & Legendre(1998),Gower (1986)) para el estudio de especies.(Jurasinski, 2007)
Se usaron disimilaridades para variables binarias para las cuales se arma la siguiente tabla tetracórica al comparar 2 de ellas

a	b	**a+b**
c	d	**c+d**
a+c	**b+d**	**a+b+c+d**

Tabla 2: Tabla

donde:

- a = número de individuos que comparten ambos atributos
- b = número de individuos que no tienen el primer atributo pero si el segundo
- c = número de individuos que tienen el primer atributo pero no el segundo
- d = número de individuos que no tienen ninguno de los atributos
- N = a+b+c+d

Originalmente en ecología las unidades o individuos son puntos de muestreo o localizaciones y los atributos son las especies de las que se busca abundancia o riqueza.

A partir de esta tabla se pueden armar muchas distancias de tipo simétricas y asimétricas, para este trabajo se consideran 2:

$$IJ = \frac{a+d}{a+b+c+d} \tag{1}$$

Simplematching es una distancia ampliamente usada en ecología, aunque ya existen aplicaciones en el campo de la salud y de la economía (Alvarez y Riaño, 2010). Pese a ser sencilla de interpretar, esta distancia posee la desventaja de que dos pares de individuos con distintas configuraciones de ceros y unos pueden estar a la misma distancia.
Sobre esta matriz de distancia es que luego se aplica el algoritmo de *Ward* (Kaufman y Rousseeuw, 1990) de carácter agregativo que busca optimizar, en cada etapa, la dispersión de las clases de la partición obtenida por agregación de individuos o grupos.

La otra distancia binaria que se manejó es la de Jaccard donde se considera la proporción de individuos que tienen ambos atributos sobre el total que tiene al menos uno de ellos.

$$SM = \frac{a}{a+b+c} \tag{2}$$

3. Resultados

Para la tabla de datos original formada por las 17 variables binarias se tiene 2 espacios de proyección, uno relativo a las variables y sus modalidades (es el que aparece representado en la figura 1). A su vez, para la nube de individuos que corresponde en este caso a los 5581 hogares existe un espacio vectorial donde las distancias que se usan luego de esta transformación son las euclídeas, lo que permite encontrar proximidades entre individuos, en este caso los hogares. Con estos se puede crear una clasificación en un número reducido

de grupos, a través de la técnica de Análisis de Clusters o conglomerados (AC), de acuerdo a una distancia previamente establecida. Para el caso particular de la ENTA se usó el método jerárquico de Ward, usando la distancia euclidiana, creando una tipología de los hogares en 3 o 4 grupos de acuerdo al confort (Ministerio de Salud Pública, 2010)

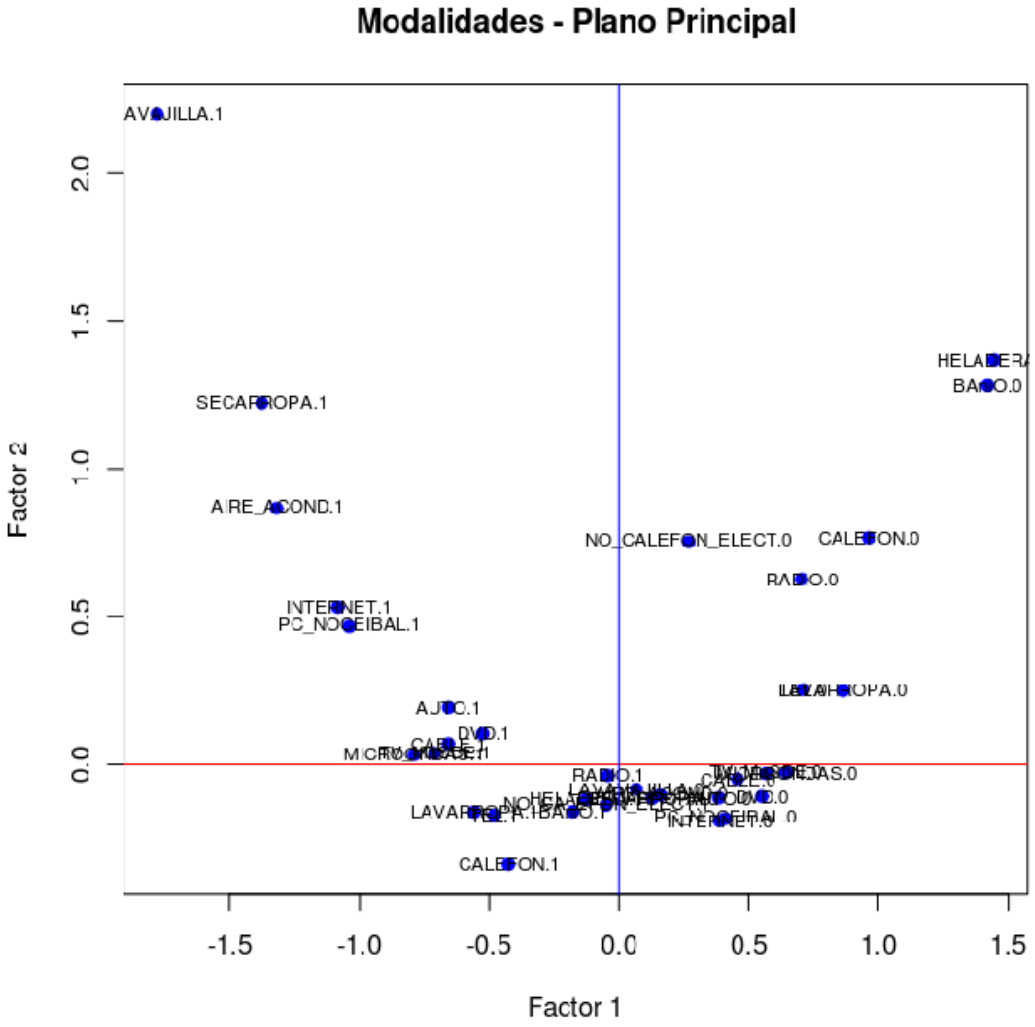

Figura 1: Plano factorial principal considerando las modalidades de las 17 variables binarias activas

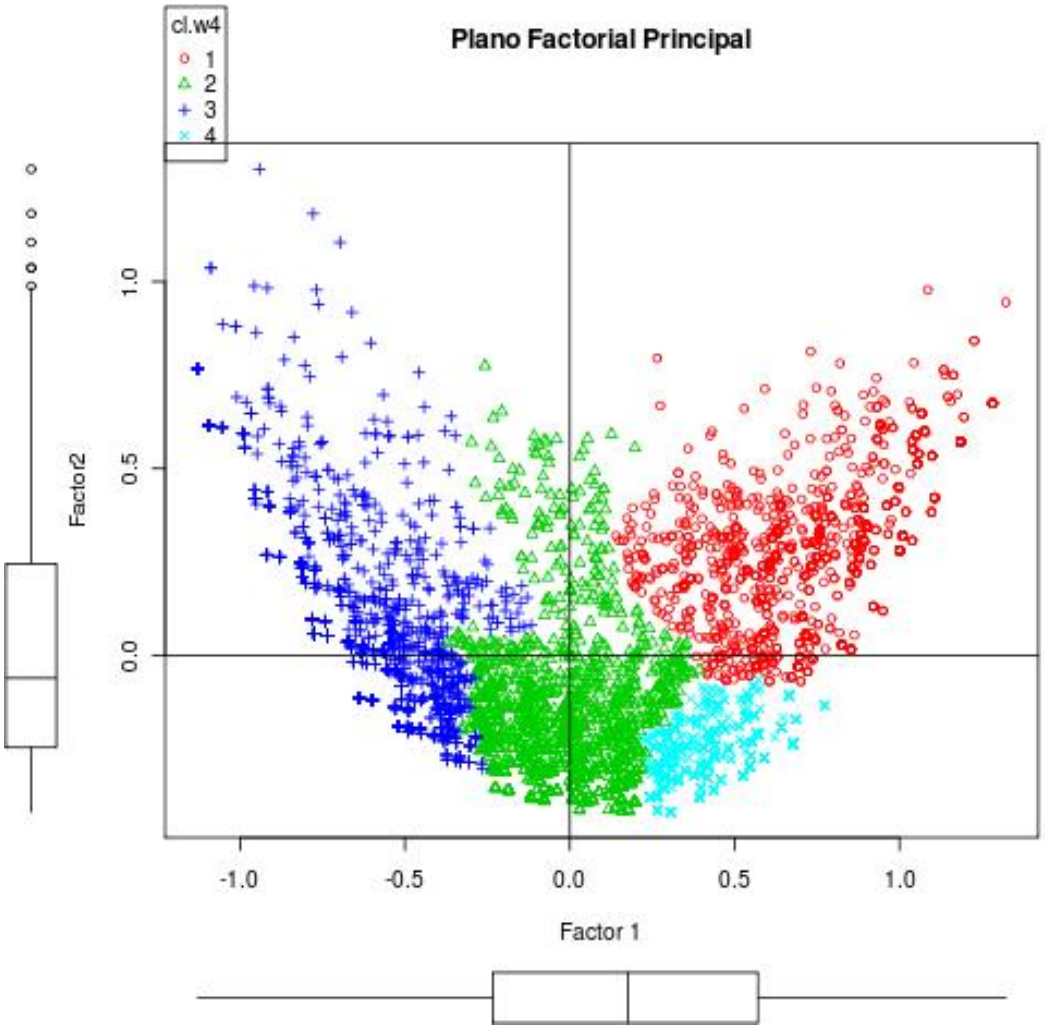

Figura 2: Plano factorial principal considerando la tiplogía de 4 grupos con algoritmo de Ward

Descripción	código	Frecuencia	Color	%
Bajo	3	1633	verde	28,9 %
Medio Bajo	2	833	azul	9,37 %
Medio Alto	1	1467	rojo	20,1 %
Alto	4	1648	celeste	41,6 %
Total		5581		

Tabla 3: Descripción de los grupos creados sobre el ACM

Una forma de ver el efecto que se captó en los grupos generados sobre el (AF) es comparar cómo se distribuye un indicador simple que resume la intensidad de tenencia (no cuales, aspecto que justamente sí capta el (AF) al haber encontrado asociaciones entre variables y modalidades de éstas)

$$INSEg = \sum_{i=1}^{i=17} X_i \tag{3}$$

podría ser este indicador que es un caso particular de $\sum_{i=1}^{i=17} \alpha_i X_i$, donde cada variable pesa lo mismo.

	grupos				
$INSEg = \sum_{i=1}^{i=17} X_i$	1	2	3	4	Total
0		43			43
1		7534			7534
2		39441			39441
3		84038			84038
4		79737	37989		117726
5		20116	114319		134435
6	361		165006		165366
7	19483		163773		183257
8	212525		12894		225419
9	223988		503	3372	227863
10	199950			37811	237761
11	56867			164671	221538
12				223699	223699
13				209350	209350
14				165243	165243
15				137432	137432

16				51413	51413
17				33839	33839
Total	713174	230910	494483	1026830	2465398

Tabla 4: Distribución de los grupos según nivel de INSE

3.1. Performance de los modelos aplicados

El modelo logístico ajustado tuvo una correcta capacidad explicativa con una tasa de acierto del 68 % (es decir que una vez ajustado el modelo solo se logró tener casi un 70 % de los hogares reclasificados en las 3 grupos de ingreso. Ese 30 % de error es el que se trasladaría al aplicarlo sobre los hogares relevados en la ENTA

Modelo estimado	Comparación de Modelos
Modelo 1 *logit* RI_3	70 %
Modelo 2 Random Forest en quintiles	45 %
Modelo 3 Random Forest en 5 intervalos de igual rango	49 %
Modelo 4 Random Forest en 10 intervalos colapsados 12,3,4,5678,910	55 %

Tabla 5: Comparación de modelos

Teniendo en cuenta las 2 distancias presentadas en las ecuaciones (1) y (2) en la sección 2.4, se construyeron las matrices de disimilaridad sobre las cuales luego se intentó construir una tipología mediante cluster jerárquico a través del algoritmo de Ward. En ambos casos no resultó una estructura clara, ya que los indicadores y reglas de detención que se usaron (*pseudo* t^2 o *pseudo* F o R^2) no mostraban una estructura de grupos.
A los efectos de ilustrar como resultaba insatisfactoria esta clasificación se cortó el dendrograma en 3 grupos para ambas distancias binarias y se muestra a continuación el resultado de proyectar en el plano factorial principal las 5581 observaciones considerando la membresía de cada hogar a los 3 grupos.
Ese gráfico muestra que los grupos sobre las distancias binarias no responden al patrón bien definido visto en la figura 2.

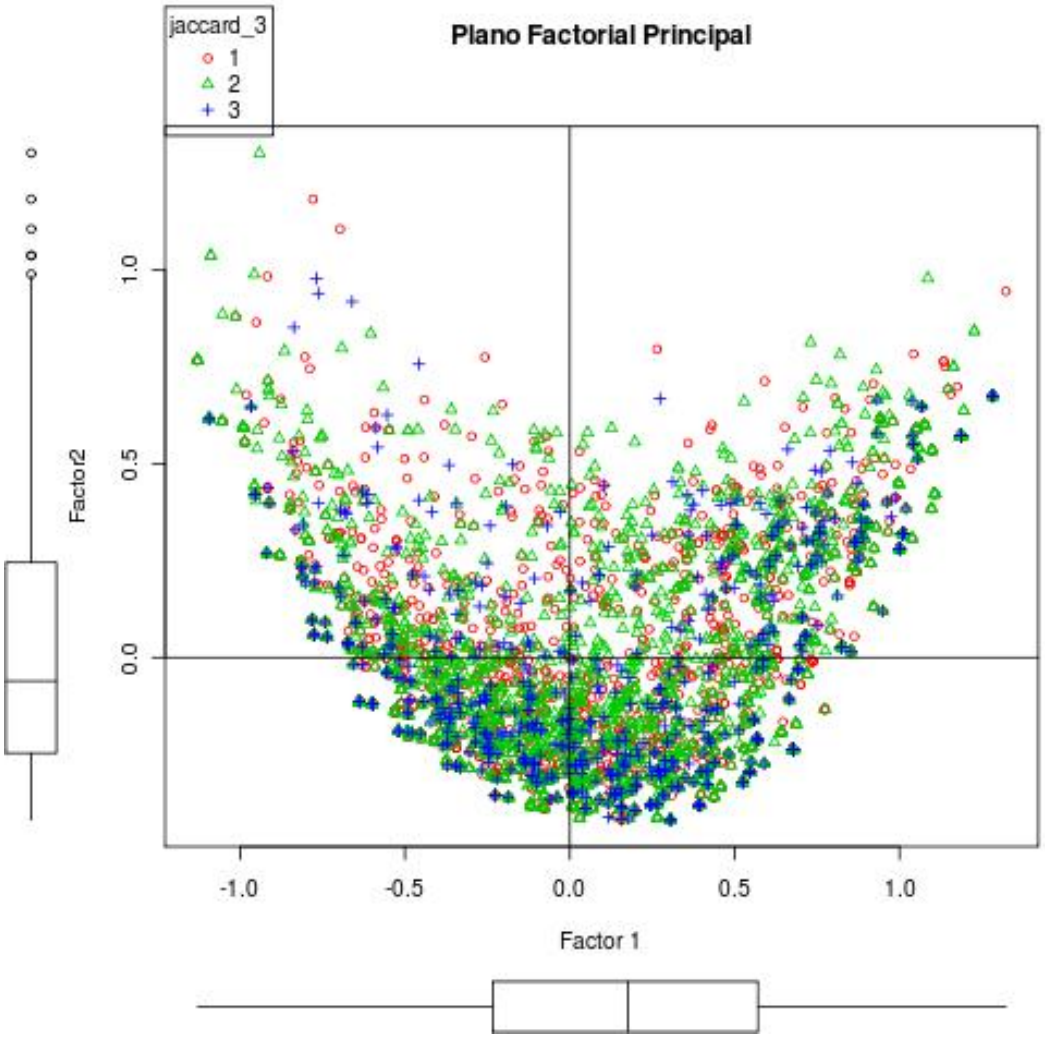

Figura 3: Plano factorial principal considerando 3 grupos creados con distancia de Jaccard

Una interrogante que se planteó era tratar de ver qué pasaba al querer encontrar proximidades entre hogares que pueden compartir parte o la totalidad de los bienes del hogar. En la tabla 6 se presentan algunas situaciones para hogares con configuraciones de tenencia de bienes diferentes y como de alguna manera las 2 distancias elegidas las consideran próximas.

a	b	c	d	IJ	SM
1	0	6	10	0,25	0,65
1	1	5	10	0,25	0,65
1	5	1	10	0,25	0,65
1	6	0	10	0,25	0,65
2	0	12	3	0,25	0,29
2	1	11	3	0,25	0,29
2	2	10	3	0,25	0,29
2	11	1	3	0,25	0,29
2	12	0	3	0,25	0,29

Tabla 6: Comparación de Distancias binarias

Algunas consideraciones que surgen de estas 2 distancias aplicadas:

- Para el primer caso la *distancia de Jaccard (IJ)* vale 0,25 se tiene una situación diferente donde hay 2 hogares que solo tienen por una parte un atributo en común, mientras que tambień lo que tienen en común a su vez es que les falta 10 atributos.
- *IJ* vale lo mismo cuando 2 hogares comparten 2 atributos y son 3 los atributos que no tienen ambos.
- Sin embargo para el primer caso para la *distancia de Simple Matching (SM)* los 2 hogares son más similares que para el segundo caso.

4. Conclusiones y futuros pasos

Debe destacarse que el problema planteado era difícil aunque no poco frecuente en la práctica de las encuestas donde un conjunto de variables que forman parte de un instrumento que ya fue validado se releva mal o en forma incompleta, con la consecuencia de que la información que se logra recolectar no sirve para medir esas variables 'inobservables' a las que se llega por aproximación.
Los enfoques metodológicos seguidos (obviamente no son los únicos) muestran resultados en algunos casos que pueden ser tenidos en cuenta, como el que surge del Clustering hecho sobre el Análisis factorial.

Por otra parte, el comparar las 17 variables contra el ingreso para la ECH muestra para el caso de la Regresión Logística una performance del 70 % que no es excelente pero no debería ser descartada. Al considerar los Bosques Aleatorios RF, a priori se esperaba un mejor comportamiento, y a la luz de los resultados se puede ver que lo que separa a los hogares en términos de ingreso depende de muchas otras cosas que no estan captadas solamente en la tenencia de bienes.

Por último, el ensayo hecho a través de las distancias binarias muestra resultados malos al no poder construir una tipología de hogares clara. Tal vez los resultados dependen mucho de las distancias binarias seleccionadas por lo cual es importante seguir estudiando las características de las otras distancias para datos binarios, considerando la asimetría que pueda existir (muchos hogares que concentran mucha cantidad de bienes o casi ningún bien).

Por lo tanto algunos de los pasos a seguir que se plantean son:

- Probar el clustering sobre el Análisis Factorial ACM para la ECH para poder cruzar efectivamente ambas clasificaciones: una que tiene que ver con el *confort* y otra con el ingreso.
- Probar con otras distancias para datos binarios que tengan en cuenta otros aspectos de asimetría.

Referencias

Alvarez, R. y Riaño, E. (2010). Creación de un índice macroeconómico. Reporte técnico, MSP-INE, Montevideo, Uruguay. (No publicado).

Blanco, J. (2006). *Introducción al Análisis Multivariado*. Instituto de Estadística, FCEA.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

CAINSE (2006-2007). Indice Nacional de Nivel Socio Económico. Reporte técnico, Comisión Agrupada del Índice de Nivel Socio-Económico.

Fernandez, A. y Perera, M. (2003). Indice de niveles socio-económicos(inse). Reporte técnico, Comisión Agrupada del Índice de Nivel Socio-Económico,CPA-Ferrere.

Instituto Nacional de Estadística (2005a). Diseño de la muestra para una encuesta de hogares ampliada. Reporte técnico, Instituto Nacional de Estadística.

Instituto Nacional de Estadística (2005b). Resultados del Censo Fase I: (Datos Definitivos revisados al 25/04/05).

Instituto Nacional de Estadística (2009). Principales resultados 2009 encuesta continua de hogares. Reporte técnico, Instituto Nacional de Estadística.

Jurasinski, G. (2007). *simba: A Collection of functions for similarity calculation of binary data*. R package version 0.2-5.

Kaufman, L. y Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Lumley, T. (2010). *Complex surveys : a guide to analysis using R*. John Wiley.

Ministerio de Salud Pública (2010). Global adult tobacco survey uruguay country report-2009. Reporte técnico, A, Montevideo, Uruguay.

Organización Mundial de la Salud (2008). *MPOWER un plan de medidas para hacer retroceder la epidemia de tabaquismo*. Organización Mundial de la Salud.

Särndal, C.-E. y Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons.

Särndal, C.-E., Swensson, B., y Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer series in Statistics. Springer.

Tobacco Free Initiative (TFI) (2010). Global adult tobacco survey (gats) fact sheet uruguay 2009.

WHO (2004). Convenio Marco de la OMS para el Control del Tabaco.