

Análisis Multivariado I

Cluster y Discriminante

En el presente trabajo se emplean las técnicas de Análisis de Cluster y Análisis Discriminante cubiertas en el curso. En una primera etapa se buscan formar grupos de individuos a partir de un pseudo panel de la Encuesta Continua de Hogares para los años 1995 a 2004 en función de la situación de empleo de los mismos. En la segunda etapa se busca determinar si una selección de variables adicionales (relevantes para el empleo) contribuyen a explicar la formación de grupos. Se realiza un ejercicio de seguimiento de los individuos en el tiempo para visualizar los efectos del a debacle económico de principios de siglo.

Integrantes:

- Coudet, Lucía CI: 4.545.399-9
- Czarniewicz, Daniel CI: 4.744.781-9
- Talvi, Ramón CI: 4.423.883-5

Contenido

Introducción	4
Marco metodológico	5
REGLAS DE DETENCIÓN	6
Presentación y análisis de los resultados	9
CLUSTERING POR MÉTODO DE WARD (DISTANCIA DE MAHALANOBIS)	9
CARACTERIZACIÓN DE GRUPOS	10
ANÁLISIS DISCRIMINANTE	12
CROSS VALIDATION	13
Conclusiones	13
Bibliografía y referencias	14
Anexos	14
ANEXO 1: DESCRIPCIÓN DE LAS VARIABLES RELEVADAS	14
ANEXO 2: TABLA DE CORRELACIONES	15
ANEXO 3: MATRIZ DE DISTANCIAS (MÉTRICA: MAHALANOBIS)	15
ANEXO 4: PROMEDIO DE LOS COEFICIENTES DE PERTENENCIA POR GRUPO	16
ANEXO 5: COMPOSICIÓN Y CARACTERÍSTICAS DE LOS GRUPOS	16
ANEXO 6: SEGUIMIENTO DE LAS UNIDADES EN EL TIEMPO	20
ANEXO 7: SALIDAS DEL TEST DE MARDIA	24
7.A - TESTEO DE NORMALIDAD MULTIVARIADA PARA TODA LA POBLACIÓN	24
7.B - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 1	24
7.C - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 2	25
7.D - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 3	25
7.E - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 4	26
7.F - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 5	26
7.G - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 6	27
ANEXO 8: COEFICIENTES ESTIMADOS DEL MODELO MULTILOGIT (GRUPO DE REFERENCIA: GRUPO 1) ...	27
ANEXO 9: FUNCIÓN DE CROSS-VALIDATION	28

Introducción

Para el presente trabajo se utilizó un pseudo-panel de la Encuesta Continua de Hogares (ECH) correspondiente a los años 1995 y 2004. La misma fue construida definiendo 36 individuo representativos en base a 4 componentes:

- Localidad: Interior (I), Montevideo (M)
- Sexo: Hombres (H), Mujeres (M)
- Tramo de edad: entre 20 y 29 años (1), entre 30 y 49 años (2), 50 años o más (3)
- Nivel educativo: Primario (1), Secundario¹ (2), Terciario² (3)

Las variables contenidas en la base de datos refieren a 5 dimensiones de la realidad laboral de los individuos³:

- Variables referentes al hogar:
 - Jefe
 - Tamaño
 - Ingresos
- Situación de empleo:
 - Desempleo
 - Empleo a tiempo parcial
 - Multiempleo
 - Subempleo⁴
 - Precariedad⁵
- Categoría de ocupación:
 - Empleo privado
 - Empleo público
 - Cuenta propistas con local
 - Cuenta propistas sin local
- Condición de ocupación
 - Profesionales y técnicos
 - Empleados de oficina
 - Empleados manuales
- Rama del establecimiento
 - Industria
 - Comercio
 - Servicios financieros, y a empresas
 - Servicios personales

¹ El mismo engloba tanto a la educación secundaria como a UTU.

² El mismo corresponde a todos los estudios de nivel terciario, incluyendo el Magisterio.

³ Ver Anexo 1 para una explicación más detallada de cada una de ellas.

⁴ Según criterio INE. Ver anexo 1.

⁵ Según criterio INE. Ver anexo 1.

En una primera etapa de nuestro análisis estudiamos el agrupamiento de las observaciones en función de las variables de situación de empleo. Esto lo realizamos mediante la aplicación de las técnicas de Clustering estudiadas en el curso.

En una segunda etapa, buscamos evaluar si el resto de las variables de nuestro set de datos, contribuyen a explicar la pertenencia a los grupos definidos en la etapa 1. Para ello, aplicamos análisis discriminante.

Comenzamos entonces el análisis con estadísticas descriptivas de todas las variables, a fin de tener una primera aproximación de la estructura de datos en nuestra base.

Posteriormente, concentramos nuestra atención en las variables referentes a situación de empleo, y navegamos por todas las técnicas trabajadas en clase, para luego poder tomar una decisión acerca cuáles utilizar y ahondar en ellas.

Una vez acotadas las variables a utilizar, realizamos cluster jerárquico agregativo para los algoritmos presentados en Rencher: vecino más cercano, vecino más lejano, centroide, mediana, y average. A su vez, también complementamos el análisis con clusters no jerárquicos, específicamente: k-medoides, k-medias, Fuzzy Sets, y Clusters basados en modelos.

Dado que en nuestra base todas las variables están definidas en porcentaje (excepto el ingreso), todas ellas recorren un mismo rango de valores. Por lo tanto, entendimos que no era necesario trabajar con datos estandarizados. Igualmente, habiendo aplicado cada metodología con los datos originales, procedimos a realizar lo propio utilizando datos estandarizados, y logramos corroborar que los resultados no varían. Es importante destacar que la variable ingreso no está incluida en nuestro set de variables referentes a la situación de empleo, por lo que no se invalida nuestra apreciación respecto de los rangos de las variables.

Como medidas de disimilaridad, utilizamos la distancia eucídea, la distancia de Mahattan, y la distancia de Mahalanobis, prefiriendo la última. Esto se debe a la presencia de correlaciones entre las variables (ver anexo 2). Consideramos que la correlación entre dos variables es alta cuando la misma se encuentre entre 0.5 y 0.8 (en valor absoluto), y crítica cuando excedía 0.8 (en valor absoluto).

Marco metodológico

En virtud de que contamos con una base no demasiado extensa (360 observaciones), no tuvimos impedimentos para construir clusters jerárquico agregativo. La misma es una técnica de clasificación supervisada ya que los grupos no están definidos de forma a priori. En cada paso de un método jerárquico, los dos clusters/individuo más cercanos (individuo-individuo, individuo-cluster, o cluster-cluster), dada la medida de disimilaridad y el algoritmo utilizados, se unen en un solo grupo. Es por esto que se le conoce como un método de particiones encajadas: una vez que dos clusters se unen, seguirán juntos hasta el final.

Una de las características interesantes del método es que, por su construcción, cuenta con una representación gráfica (dendograma) que da una primera aproximación a la estructura de grupos subyacente en los datos.

Para el caso de los **clusters jerárquicos agregativos** se parte de una situación en la que cada individuo forma su propio cluster, y se culmina con todos los individuos en un mismo cluster. Por el contrario, en los métodos **jerárquicos divisivos** se parte de un grupo inicial que contiene todos los individuos y se culmina cuando cada individuo conforma su propio cluster.

Para el presente trabajo se utilizó una amplia gama de combinaciones técnica⁶- métrica⁷-método⁸ para el caso de los métodos jerárquico agregativos. Atendiendo tanto a las distintas reglas de detención formales (R cuadrado, Pseudo F, y Pseudo T), así como a las informales (inspección visual del dendrograma), optamos trabajar con los grupos definidos por la rutina jerárquico agregativa, con distancia de Mahalanobis, y método de Ward.

En el **método de Ward**, la unión de dos grupos se realiza de forma tal que se minimice la variación dentro de los grupos en la nueva partición. Es decir, al descomponer la variación total en variación intra-grupo (within), y variación entre-grupos (between), el método juntará aquellos grupos que produzcan el efecto de hacer mínima la variación within en la nueva partición. El algoritmo tiende a juntar a grupos con pequeño número de observaciones, y es por ende, sesgado hacia la formación de grupos esféricos.

La métrica seleccionada (distancia de Mahalanobis) fue considerada apropiada por contemplar la correlación entre las variables para definir la distancia. La misma se calcula como:

$$d_{ij}^2 = (X_{ik} - X_{jk})' \Sigma^{-1} (X_{ik} - X_{jk})$$

donde Σ es la matriz de varianzas y covarianzas

REGLAS DE DETENCIÓN

Análisis visual del dendrograma: dado que las barras verticales del dendrograma indican el nivel al que se unen los individuos/grupos, analizar visualmente la longitud de dichas barras da una idea de la estructura de grupos que presentan los datos. Este método es muy informal y solo permite una primera aproximación. En caso de querer obtener una cantidad previamente especificada de grupos, esto determinará el nivel en el cual cortar el dendrograma.

Criterio del R^2 : este indicador relaciona la variación explicada y la variación total, dónde la variación explicada está representada por la estructura de grupos que se encuentra en cada nivel. Por lo tanto, el valor del R^2 se encuentra en el intervalo [0,1], valiendo 0 cuando todas las observaciones se encuentran en un mismo grupo (es decir, la variación explicada por la estructura de grupos es 0 ya que todos los individuos están en un mismo grupo), y valiendo 1 cuando cada individuo es un grupo (es decir, cuando la variación explicada es igual a la variación total).

En términos matemáticos:

⁶ Jerárquicos agregativos, jerárquicos divisivos, y no jerárquicos.

⁷ Euclídea, Manhattan, y Mahalanobis.

⁸ Single Linkage, Complete Linkage, Average Linkage, y Ward.

$$R^2 = 1 - \frac{\sum_{k=1}^g \sum_{i=1}^{n_k} \sum_{j=1}^p (x_{ijk} - \bar{x}_{jk})^2}{\sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \bar{x}_j)^2}$$

con p variables, g grupos, n observaciones, n_k observaciones en el grupo k

Criterio pseudo-F: relaciona la suma de variaciones entre los grupos (variación explicada) con la suma de variaciones en los grupos (variación residual):

$$Pseudo F = \frac{Variación\ entre\ grupos / (k - 1)}{Variación\ en\ los\ grupos / (n - k)} \quad k\ grupos, n\ observaciones$$

También puede calcularse como relación entre trazas de matrices:

$$Pseudo F = \frac{tr(B)/(k - 1)}{tr(W)/(n - k)}$$

O como cociente del R^2 :

$$Pseudo F = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)}$$

Reglas empíricas de utilización:

- Si el indicador crece de forma monótona al crecer el número de grupos, no se puede determinar una estructura de grupos.
- Si disminuye de forma monótona al crecer el número de grupos, no se puede determinar una estructura de grupos, pero se puede decir que existe jerarquía.
- Cuando se halla un máximo, la población presenta una estructura de grupos en ese máximo.

Criterio pseudo-t: este indicador busca determinar la significación de juntar dos grupos. Es decir, busca determinar en cada paso, si la disminución de la SCR (variación intra-grupos) como resultado de pasar de k a $k+1$ grupos es significativa o no.

Algebraicamente, este indicador también se basa en las trazas de las matrices de varianza en el grupo (W), antes y después de unir dos grupos:

$$pseudo - t^2 = \frac{tr(W_{GL}) - [tr(W_G) + tr(W_L)]}{[tr(W_G) + tr(W_L)] / (n_G + n_L - 2)}$$

n_i son las observaciones en el grupo $i = G, L$

$tr(W_i)$ es la traza en el grupo $i = G, L, GL$

Si se analiza el vector de valores de pseudo- t^2 desde abajo (1 grupo con n observaciones), hacia arriba (n grupos con 1 observación cada uno), si en $k+1$ grupos presenta una caída “fuerte” respecto de k , entonces nos quedamos con $k+1$ grupos.

Los métodos de agrupación no jerárquicos son técnicas de clasificación no supervisada en lo que requiere definirse a priori el número de grupos con el cuál se va trabajar. A lo existir una estructura jerárquica, no existe una representación gráfica.

En el presente trabajo se utilizaron las técnicas no jerárquicas: k-medoides, k-medias, Fuzzy Sets, y Clusters basados en modelos.

K- medoides requiere definir a priori k grupos y sus respectivos centros de gravedad (individuos representativos). Al comienzo, los centros de gravedad son elegidos por el investigador ("semillas"). Estos pueden ser elegidos aleatoriamente, o se pueden elegir los primeros k items de la base de datos que cumplan con determinada distancia mínima establecida. Luego de elegir las semillas, se agrupan las observaciones y se redefinen los centros de gravedad en cada grupo, permitiendo la relocalización de observaciones. Esto se repite hasta que se llega a la convergencia o al número de iteraciones máximo preestablecido.

El método de **k-medias** se distingue en que los centros de gravedad utilizados corresponden a los vectores de medias de los grupos.

Fuzzy Sets, a diferencia de cualquier otro método, busca calcular coeficientes de pertenencia de cada observación para los distintos grupos. Es decir, no es una clasificación rígida, sino que por el contrario, es una clasificación difusa. En el anexo 4 se presenta una tabla con el coeficiente de pertenencia promedio para cada grupo.

Las técnicas de Clusters basados en modelos en modelos se pueden dividir entre aquellas en las que se asume una distribución de probabilidad en los grupos, y las técnicas en las que no se asume una distribución, sino que se busca estimar esta, por ejemplo mediante kernels. Una explicación detallada de estos métodos pueden encontrarse en Rencher (2002). No nos detenemos aquí en esta, dado que los resultados de la aplicación de estas técnicas para nuestro set de datos fueron desechados⁹.

Posterior al análisis de cluster, realizamos **Análisis Discriminante (AD)**. El mismo es una técnica de clasificación supervisada, ya que existe previamente una categorización de las observaciones. Es decir, sabemos que las observaciones pertenecen a un determinado grupo. Lo que buscamos es encontrar una regla de clasificación asociada a las variables que se están midiendo, que permita asignar cada individuo a un grupo determinado con el menor error posible. Esta técnica encuentra su mayor utilidad a la hora de clasificar nuevas observaciones.

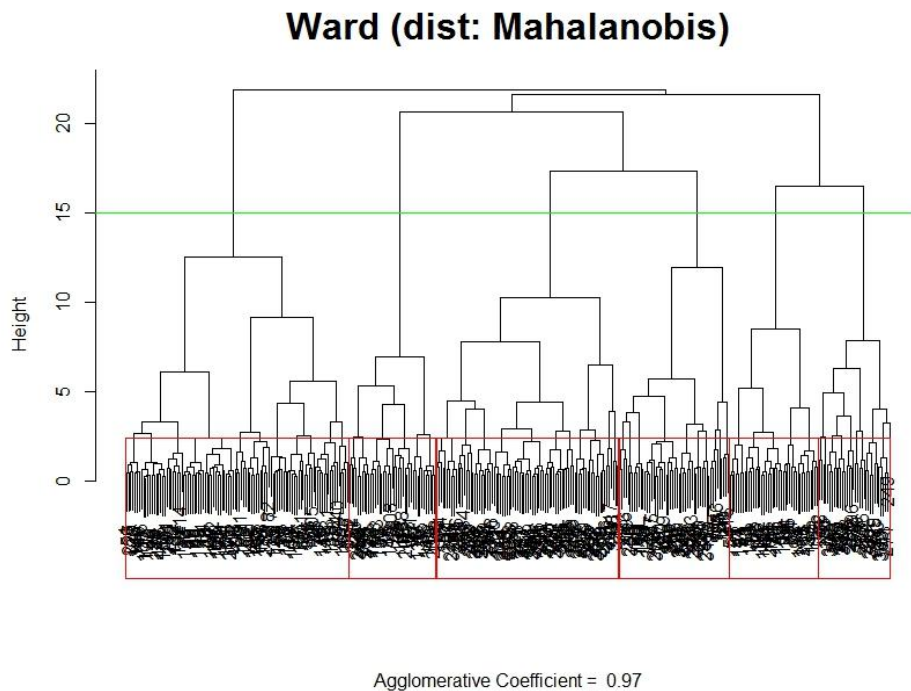
El primer paso en el análisis discriminante es evaluar, en este caso, a través del Test de Mardia, si los datos tienen una distribución normal multivariante en cada grupo. Para nuestros datos, rechazamos esta hipótesis al 5% de significación (ver anexo 7). Por ende, no nos fue posible realizar ni AD probabilístico lineal, ni AD probabilístico cuadrático (ambos basados en la normalidad de los grupos). Optamos por estimar un discriminante multilogístico.

⁹ Es importante estacar aquí que la cantidad de observaciones requeridas para realizar buenas estimaciones, y poder aplicar correctamente estas técnicas, crece exponencialmente con el número de variables. El solo contar con 360 observaciones para 19 dimensiones constituye el principal argumento por el cual los resultados de la aplicación de estas técnicas fueron desechados por el equipo.

Presentación y análisis de los resultados

CLUSTERING POR MÉTODO DE WARD (DISTANCIA DE MAHALANOBIS¹⁰)

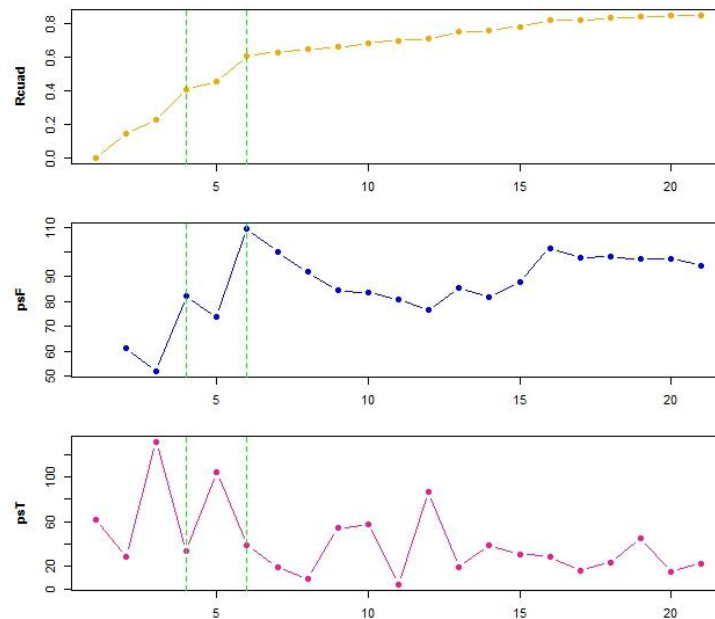
En el dendograma, claramente podemos apreciar que se forman grupos esféricos y consideramos el más apropiado un $k=6$, siendo k el número de grupos.



Según los indicadores, una estructura de 4 o 6 grupos serían las más adecuadas para nuestros datos. Según el R cuadrado, no elegimos $k=4$ porque este aún no se encuentra estabilizado. Sin embargo, a partir de $k=6$, el crecimiento marginal deja de ser significativo. El pseudo F presenta máximos locales tanto en $k=4$ como en $k=6$, por lo no nos permite discernir entre uno y otro. El pseudo t presenta caídas relevantes para ambos casos.

¹⁰ En el anexo 3 se incluye la matriz de distancias calculadas con métrica de Mahalanobis.

Ward (dist: Mahalanobis)



CARACTERIZACIÓN DE GRUPOS¹¹

Grupo 1	Grupo de observaciones heterogéneas.
Grupo 2	Predominan universitarios de mayores de 29 años.
Grupo 3	Mujeres. Predominan personas de 50 o más años. Todas las observaciones tienen un nivel de formación terciario (con una única excepción).
Grupo 4	Grupo de observaciones heterogéneas.
Grupo 5	Está compuesto por personas de ambos sexos. De un total de 52, solamente 4 alcanzan un nivel de formación terciario. A su vez, hay escasas observaciones con 50 o más años.
Grupo 6	Es un grupo con predominancia de mujeres de nivel educativo terciario y rango de edad de 20 a 29 años.

¹¹ El anexo 5 incluye una caracterización de la composición de cada grupo.

Medias por grupo					
	Desempleo	T. Parcial	Multiempleo	Subempleo	Precariedad
1	10.687	14.678	8.075	7.646	15.898
2	3.166	18.176	29.258	5.638	6.159
3	5.588	37.516	14.425	5.204	10.773
4	11.793	33.910	9.905	15.218	34.467
5	15.391	26.386	5.846	10.330	53.808
6	27.721	64.471	7.545	15.477	49.353

Si nos disponemos a caracterizar los grupos en función del set de variables seleccionado para nuestro análisis (referentes a la situación de empleo de los individuos), vemos claramente que el grupo 2, seguido del 3, se destacan por presentar situaciones más deseables en lo que respecta la calidad de la inserción laboral (bajos niveles de desempleo, a la vez que bajos niveles de subempleo, y de precariedad), mientras que los grupos 4, 5, y 6 (especialmente el 6) se destacan por presentar una situación menos deseable en calidad de inserción laboral.

A continuación se construyó un ranking de situación de empleo de los grupos. Para ellos se tuvieron en cuenta las variables: desempleo, subempleo, y precariedad. Las variables empleo a tiempo parcial, y multiempleo no fueron tenidas en cuenta dado que responden a decisiones voluntarias de los individuos. Podría argumentarse que la variable multiempleo refleja una necesidad (para al menos algunos de los individuos), y no una decisión de las personas. Si bien esto puede ser cierto para algunas observaciones de la ECH¹², la alta correlación con la variable profytec (% de profesionales y técnicos), invita a pensar que en la mayoría de los casos esto no es cierto, sino que se trata de individuos con formación superior que eligen realizar varios empleos (por ejemplo: consultorías privadas, docencia, etc.).

Para la construcción del ranking se les asignaron las posiciones 1 a 6 en cada dimensión de forma tal que se otorga el menor valor, al grupo con menor media en cada respectiva dimensión, y se procedió a realizar la suma horizontal de las posiciones, obteniéndose las siguientes posiciones:

Ranking					
Grupos	Desempleo	Subempleo	Precariedad	Suma	Posición
1	3	3	3	9	3
2	1	2	1	4	1
3	2	1	2	5	2
4	4	5	4	13	4
5	5	4	6	15	5
6	6	6	5	17	6

El equipo no es ajeno al hecho de que la construcción del ranking con los lineamientos antes descritos es sumamente cuestionable. Por ejemplo, podría objetarse el hecho de que las

¹² Recuérdese que aquí contamos con un pseudo panel de la ECH, y no con las encuestas completas.

posiciones en el ranking no dependen de valores absolutos, sino que son relativas a la muestra obtenida. Por otra parte, debe tenerse en cuenta que las variables utilizadas para su construcción responden únicamente a disponibilidad de datos. El objetivo de este ejercicio es el seguimiento de las unidades en el tiempo (siendo este uno de los posibles análisis a realizar con la estructura de grupos seleccionada).

En dicho seguimiento, puede observarse claramente que todos los individuos (a excepción de IM33, MH23, y MM21) se deterioraron (perdieron posiciones) a partir del año 2000, fenómeno probablemente asociado al declive económico de principios del siglo XXI. No se observa una predominancia de afección ni por nivel educativo, ni por edad, ni por sexo, sino que el efecto fue generalizado. Sin embargo, parecería haber una tendencia a que, a menor nivel educativo, la afección sea mayor¹³.

ANÁLISIS DISCRIMINANTE

Como se mencionó anteriormente, se rechaza la hipótesis nula de distribución normal multivariada para toda la población, y por grupos, por lo que, no resulta certero la implementación de un AD probabilístico normal. (Ver salidas del Test de Mardia en el anexo 7).

Entendimos por lo tanto que, utilizando los resultados obtenidos en Clustering como variable de respuesta, AD multilogístico es lo más apropiado. Utilizamos el grupo 1 como grupo de referencia

Las estimaciones de los coeficientes del modelo multilogit puede verse en el anexo 8.

El modelo presenta una precisión de aproximadamente 87.2%, lo cual es la proporción de valores predichos correctamente.

		Valores Predichos						Total
		1	2	3	4	5	6	
Grupos	1	98	0	0	4	2	1	105
	2	1	37	1	2	0	0	41
	3	1	1	39	1	0	0	42
	4	4	2	1	68	10	1	86
	5	4	0	0	8	40	0	52
	6	0	0	0	1	1	32	34
Total		108	40	41	84	53	34	360

Asumiendo igualdad de costos por error de clasificación en todos los grupos, construimos la tabla de contingencia, la cual nos permite apreciar inmediatamente que los porcentajes de predicciones correctas para cada grupo superan ampliamente el 70%. Inclusive,

¹³ En el anexo 6 se presentan los gráficos de seguimiento de las unidades en el tiempo para las 36 unidades.

para 3 de los grupos, los porcentajes se encuentran por encima del 90%, con un total del 87.76%.

		Valores Predichos						Total
		1	2	3	4	5	6	
Grupos	1	93.33%	0.00%	0.00%	3.81%	1.90%	0.95%	
	2	2.44%	90.24%	2.44%	4.88%	0.00%	0.00%	
	3	2.38%	2.38%	92.86%	2.38%	0.00%	0.00%	
	4	4.65%	2.33%	1.16%	79.07%	11.63%	1.16%	
	5	7.69%	0.00%	0.00%	15.38%	76.92%	0.00%	
	6	0.00%	0.00%	0.00%	2.94%	2.94%	94.12%	
Total								87.76%

CROSS VALIDATION

Utilizando la función programada por Manuel Amunategui (ver anexo 9), implementamos la técnica con 360 particiones. Esta arrojó un nivel de precisión del 79,7%, el cual es alto, y cercano al valor obtenido anteriormente al aplicar discriminante multilogístico.

Conclusiones

Del análisis cluster se desprende que los grupos conformados disciernen particularmente en función del nivel educativo alcanzado (si bien varios de ellos resultaron bastante heterogéneos). En particular, se separan las observaciones con nivel educativo terciario (grupos 2 y 6), y primario (grupo 5) del resto de las observaciones. Asimismo, vemos que el sexo es otra característica de nuestros individuos representativos que permite discernir entre un grupo y otro. En particular, el grupo 3 está conformado en su totalidad por mujeres, y el grupo 5 tiene una alta predominancia de estas.

Si seguimos a los individuos representativos a lo largo del tiempo, vemos que la mayoría de ellos, al comenzar la crisis de principios del siglo XXI pasan a grupos con propiedades menos deseables en materia de calidad de inserción laboral.

En lo que respecta a AD, la función multilogística nos permite arribar a una clasificación de observaciones con un grado alto de precisión. Esto es, las variables no incluidas en la categoría situación de empleo, contribuyen a explicar la formación de grupos.

Bibliografía y referencias

- Blanco (2006) - Introducción al análisis multivariado - 1era edición
- Everitt & Hothorn (2014) - A handbook of statistical analyses using R - 3rd edition
- Peña (2002) - Análisis de datos multivariantes
- Rencher (2002) - Methods of multivariate analysis - 2nd edition

Anexos

ANEXO 1: DESCRIPCIÓN DE LAS VARIABLES RELEVADAS

	Variable	Clase	Descripción
1	jefe	Variables referentes al hogar	% de jefes de hogar
2	size		% de personas que trabajan en establecimientos con menos de 5 personas
3	ingreso		Mediana del ingreso total proveniente de remuneraciones del trabajo
4	desemp	Situación de empleo	% de personas desempleadas
5	tparcial		% de personas que trabajan menos de 35 horas por semana y que no buscan otro trabajo
6	multiemp		% de personas con dos o más empleos
7	subemp*		% de trabajadores subempleados
8	preacario**	Categoría de ocupación	% de trabajadores precarios
9	privado		% de trabajadores del sector privado
10	publico		% de trabajadores en el sector público
11	cpsl		% de trabajadores por cuenta propia sin local
12	cpcl	Condición de ocupación	% de por cuenta propia con local
13	profytec		% de profesionales y técnicos
14	oficina		% de empleados de oficina
15	manual	Rama del establecimiento	% de empleados manuales
16	indust		% de personas que trabajan en la industria
17	comercio		% de personas que trabajan en el comercio
18	sfinan		% de personas que trabajan en servicios financieros y a eporesas
19	sperson		% de personas que trabajan en servicios personales

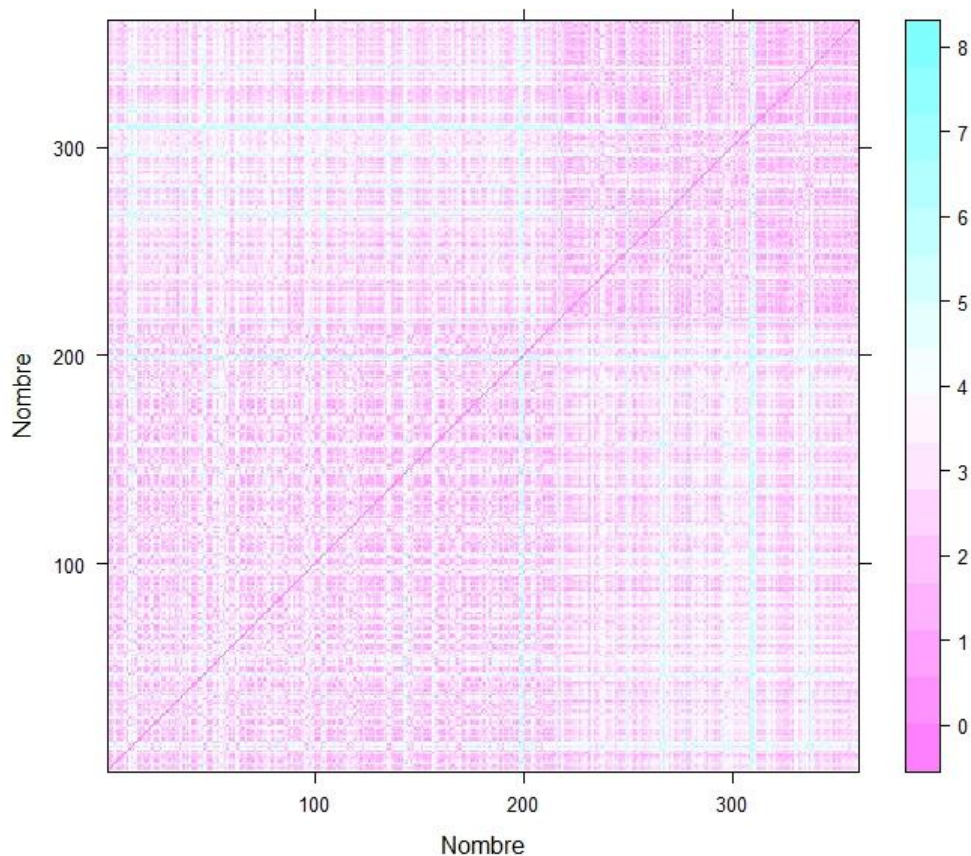
* Según criterio INE. Corresponde a personas que desempeñan su actividad a tiempo parcial de forma involuntaria. Incluye subempleo por insuficiencia de horas trabajadas, y subempleo por insuficiencia de volumen de trabajo. Se considera trabajador a tiempo parcial a aquel que trabaja menos de 40 horas semanales.

** Según criterio INE. Incluye a los asalariados en el sector privado que no están protegidos por el sistema de seguridad social, así como a las personas que se encuentran buscando otro trabajo para sustituir al actual en razón de que el mismo es poco estable o porque son trabajadores familiares no remunerados

ANEXO 2: TABLA DE CORRELACIONES

	jefe	desemp	size	tparcial	multiemp	privado	publico	cpsl	cpcl	profytec	oficina	manual	indust	comercio	sfinan	sperson	subemp	precario
jefe	1.00	-0.69	-0.03	-0.54	0.22	-0.61	0.17	0.44	0.55	-0.03	-0.32	0.45	0.19	-0.20	0.10	-0.33	-0.43	-0.45
desemp	-0.69	1.00	0.27	0.65	-0.48	0.77	-0.53	-0.09	-0.61	-0.31	0.10	-0.17	0.02	0.40	-0.25	0.20	0.65	0.76
size	-0.03	0.27	1.00	0.30	-0.47	0.23	-0.57	0.30	0.31	-0.61	-0.51	0.11	0.23	0.18	-0.59	0.41	0.36	0.59
tparcial	-0.54	0.65	0.30	1.00	-0.11	0.24	-0.08	-0.15	-0.18	0.10	0.04	-0.45	-0.34	0.00	-0.25	0.02	0.65	0.48
multiemp	0.22	-0.48	-0.47	-0.11	1.00	-0.54	0.66	-0.47	0.23	0.83	0.12	-0.50	-0.52	-0.64	0.62	-0.22	-0.29	-0.58
privado	-0.61	0.77	0.23	0.24	-0.54	1.00	-0.80	-0.05	-0.70	-0.56	0.06	0.06	0.38	0.53	-0.20	0.37	0.44	0.68
publico	0.17	-0.53	-0.57	-0.08	0.66	-0.80	1.00	-0.30	0.27	0.84	0.23	-0.34	-0.64	-0.62	0.36	-0.37	-0.33	-0.66
cpsl	0.44	-0.09	0.30	-0.15	-0.47	-0.05	-0.30	1.00	0.04	-0.50	-0.52	0.80	0.39	0.15	-0.45	-0.29	0.13	0.23
cpcl	0.55	-0.61	0.31	-0.18	0.23	-0.70	0.27	0.04	1.00	0.09	-0.20	-0.02	-0.06	-0.23	-0.06	0.07	-0.31	-0.40
profytec	-0.03	-0.31	-0.61	0.10	0.83	-0.56	0.84	-0.50	0.09	1.00	0.29	-0.57	-0.74	-0.63	0.55	-0.30	-0.18	-0.55
oficina	-0.32	0.10	-0.51	0.04	0.12	0.06	0.23	-0.52	-0.20	0.29	1.00	-0.48	-0.22	0.17	0.48	-0.12	0.01	-0.34
manual	0.45	-0.17	0.11	-0.45	-0.50	0.06	-0.34	0.80	-0.02	-0.57	-0.48	1.00	0.65	0.28	-0.43	-0.20	-0.12	0.15
indust	0.19	0.02	0.23	-0.34	-0.52	0.38	-0.64	0.39	-0.06	-0.74	-0.22	0.65	1.00	0.52	-0.34	0.17	-0.08	0.20
comercio	-0.20	0.40	0.18	0.00	-0.64	0.53	-0.62	0.15	-0.23	-0.63	0.17	0.28	0.52	1.00	-0.24	0.12	0.10	0.36
sfinan	0.10	-0.25	-0.59	-0.25	0.62	-0.20	0.36	-0.45	-0.06	0.55	0.48	-0.43	-0.34	-0.24	1.00	-0.29	-0.36	-0.61
sperson	-0.33	0.20	0.41	0.02	-0.22	0.37	-0.37	-0.29	0.07	-0.30	-0.12	-0.20	0.17	0.12	-0.29	1.00	0.02	0.27
subemp	-0.43	0.65	0.36	0.65	-0.29	0.44	-0.33	0.13	-0.31	-0.18	0.01	-0.12	-0.08	0.10	-0.36	0.02	1.00	0.63
precario	-0.45	0.76	0.59	0.48	-0.58	0.68	-0.66	0.23	-0.40	-0.55	-0.34	0.15	0.20	0.36	-0.61	0.27	0.63	1.00
ingreso	0.44	-0.56	-0.54	-0.28	0.83	-0.56	0.57	-0.33	0.23	0.69	0.15	-0.34	-0.33	-0.52	0.71	-0.31	-0.44	-0.70

Nota: las celdas con fondo rojo señalan correlaciones mayores a 0.8 (en valor absoluto), mientras que las celdas con fondo anaranjado señalan correlaciones entre 0.5 y 0.8 (en valor absoluto). El color de fuente roja señala correlaciones negativas, mientras que el negro señala correlaciones positivas.

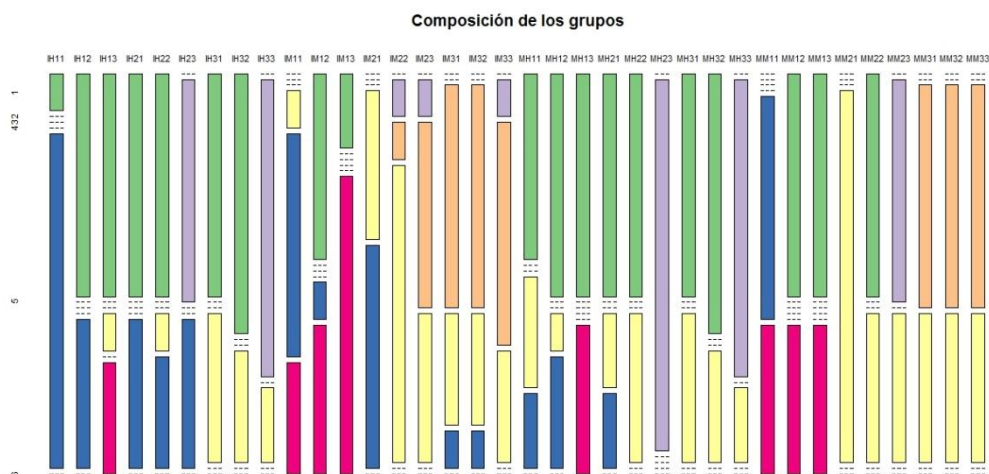
ANEXO 3: MATRIZ DE DISTANCIAS (MÉTRICA: MAHALANOBIS)**Matriz de distancias de Mahalanobis**

ANEXO 4: PROMEDIO DE LOS COEFICIENTES DE PERTENENCIA POR GRUPO

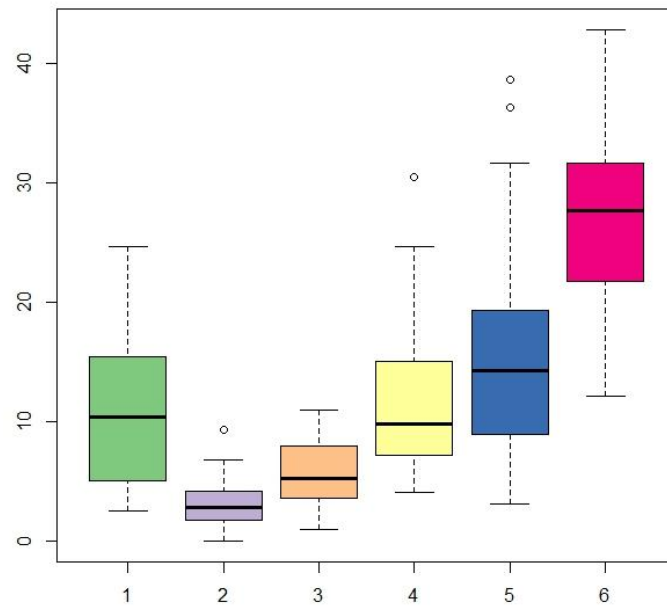
Promedio de coeficientes de pertenencia (Fuzzy Sets) por grupos						
	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
Grupo 1	0.29893845	0.14021227	0.14021233	0.14021232	0.14021232	0.14021231
Grupo 2	0.11319817	0.17736023	0.17736043	0.17736040	0.17736039	0.17736037
Grupo 3	0.17455447	0.16508903	0.16508914	0.16508913	0.16508912	0.16508911
Grupo 4	0.13845374	0.17230940	0.17230918	0.17230921	0.17230922	0.17230925
Grupo 5	0.11535947	0.17692827	0.17692802	0.17692806	0.17692808	0.17692811
Grupo 6	0.11455462	0.17708900	0.17708912	0.17708910	0.17708909	0.17708908

ANEXO 5: COMPOSICIÓN Y CARACTERÍSTICAS DE LOS GRUPOS

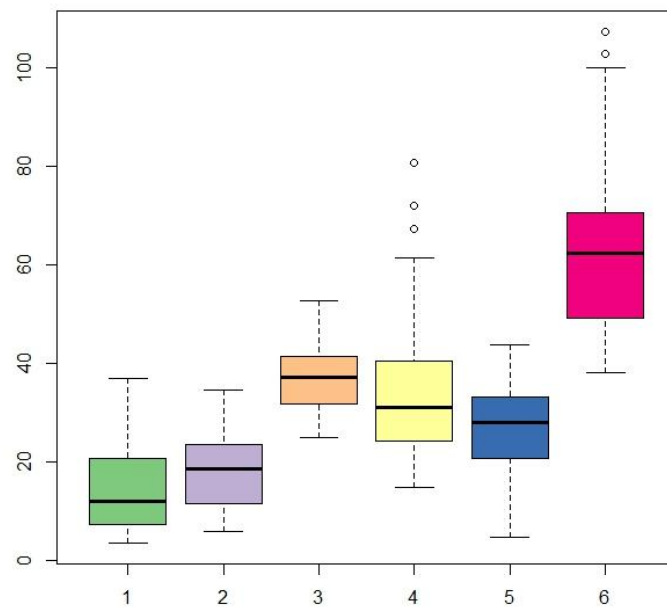
Composición de los grupos										
Grupo	Localidad		Sexo		Rango etario			Nivel educativo		
	Interior	Montevideo	Hombres	Mujeres	20 a 29	30 a 49	50 o más	Primario	Secundario	Terciario
1	42.86%	57.14%	76.19%	23.81%	46.67%	28.57%	24.76%	28.57%	52.38%	19.05%
2	41.46%	58.54%	78.05%	21.95%	0.00%	58.54%	41.46%	0.00%	2.44%	97.56%
3	57.14%	42.86%	0.00%	100.00%	0.00%	14.29%	85.71%	28.57%	30.95%	40.48%
4	43.02%	56.98%	34.88%	65.12%	6.98%	47.67%	45.35%	40.70%	36.05%	23.26%
5	75.00%	25.00%	59.62%	40.38%	59.62%	36.54%	3.85%	69.23%	23.08%	7.69%
6	52.94%	47.06%	20.59%	79.41%	100.00%	0.00%	0.00%	20.59%	23.53%	55.88%

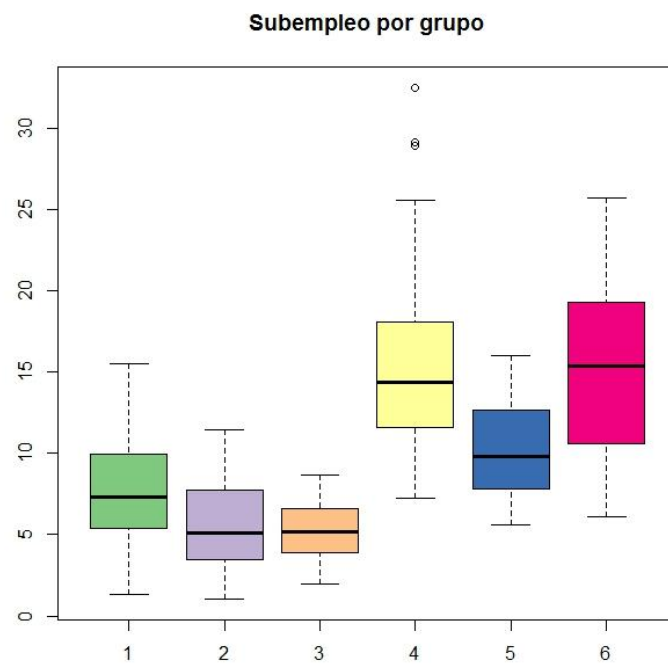
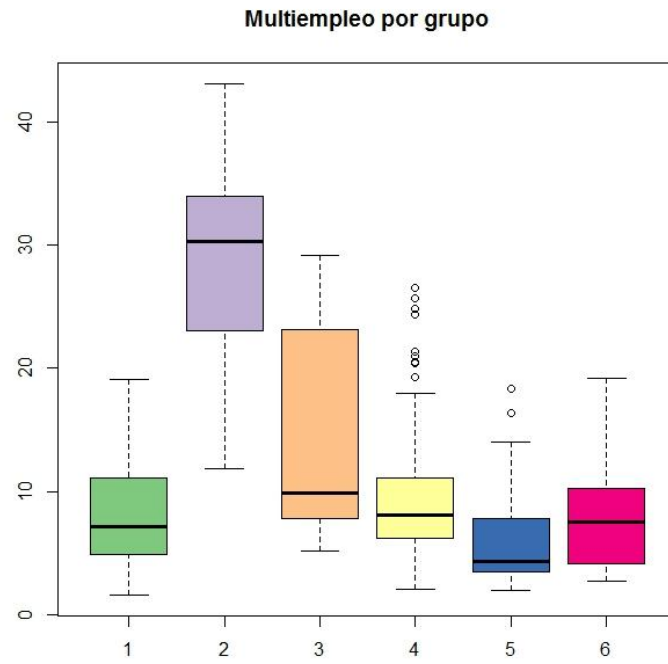


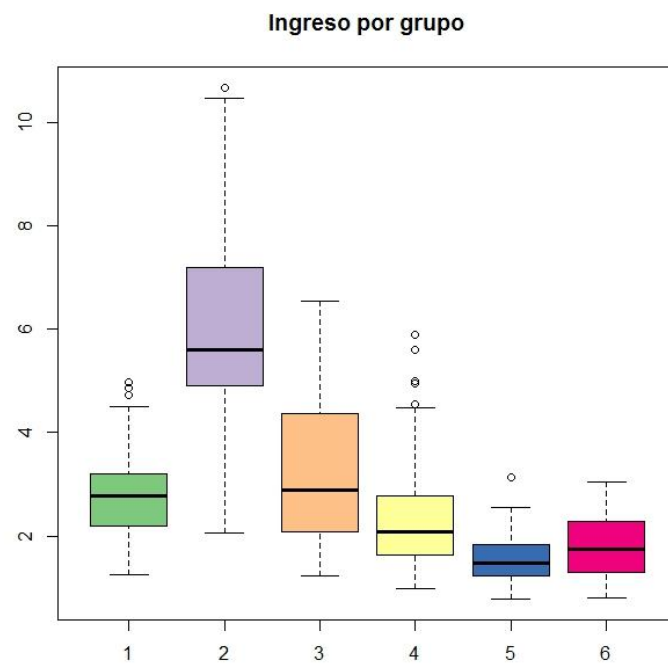
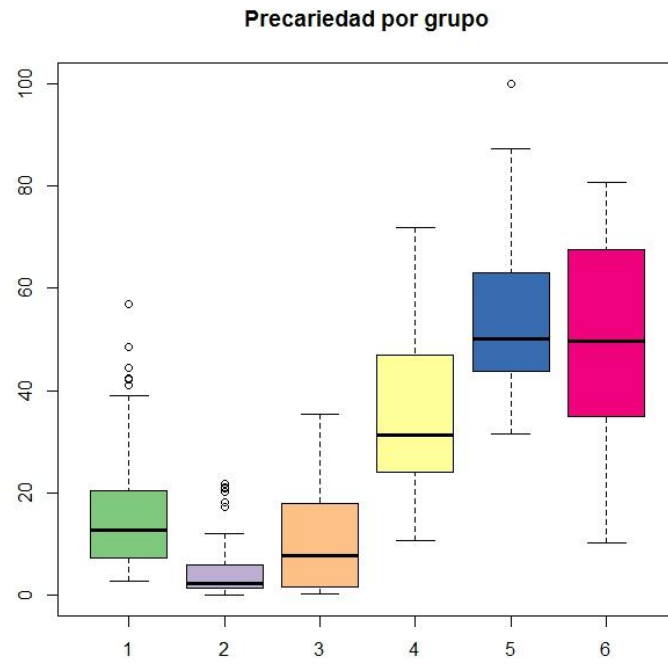
Desempleo por grupo



Trabajo a tiempo parcial por grupo

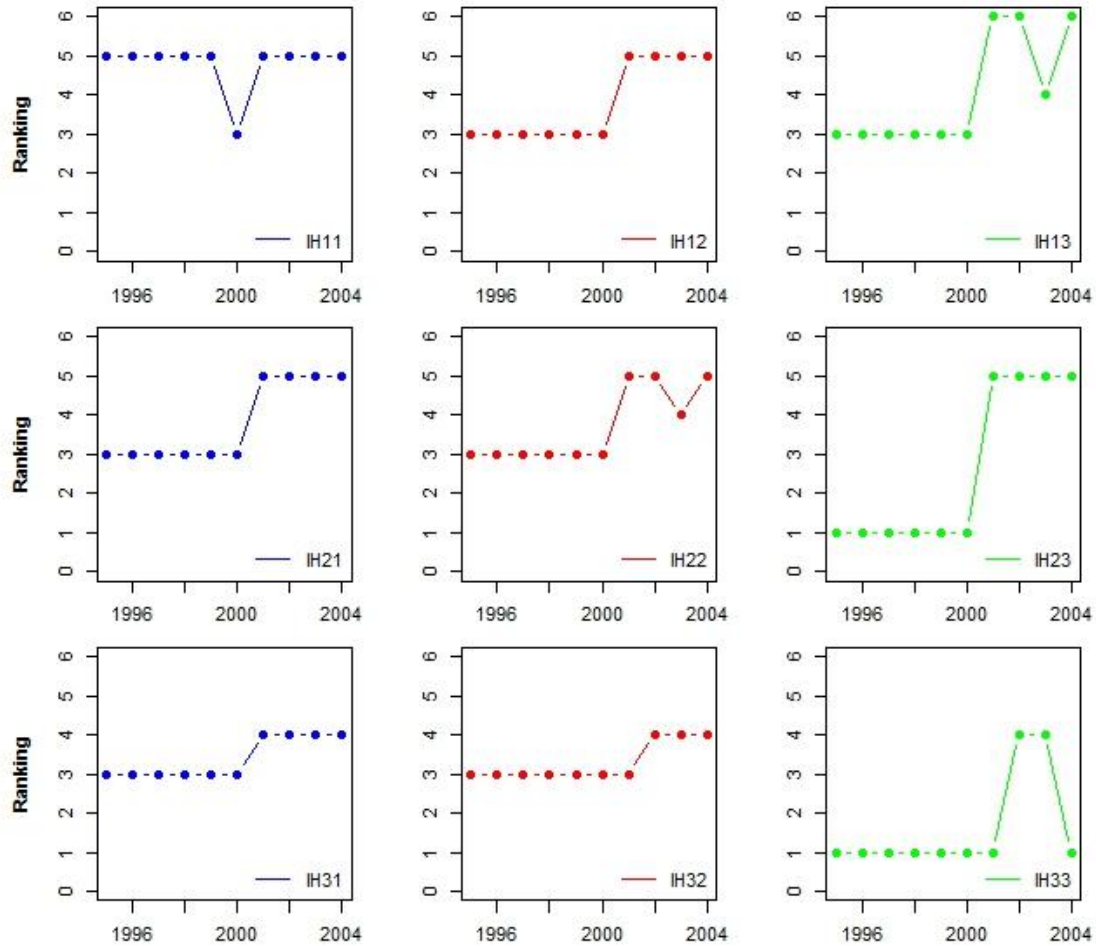




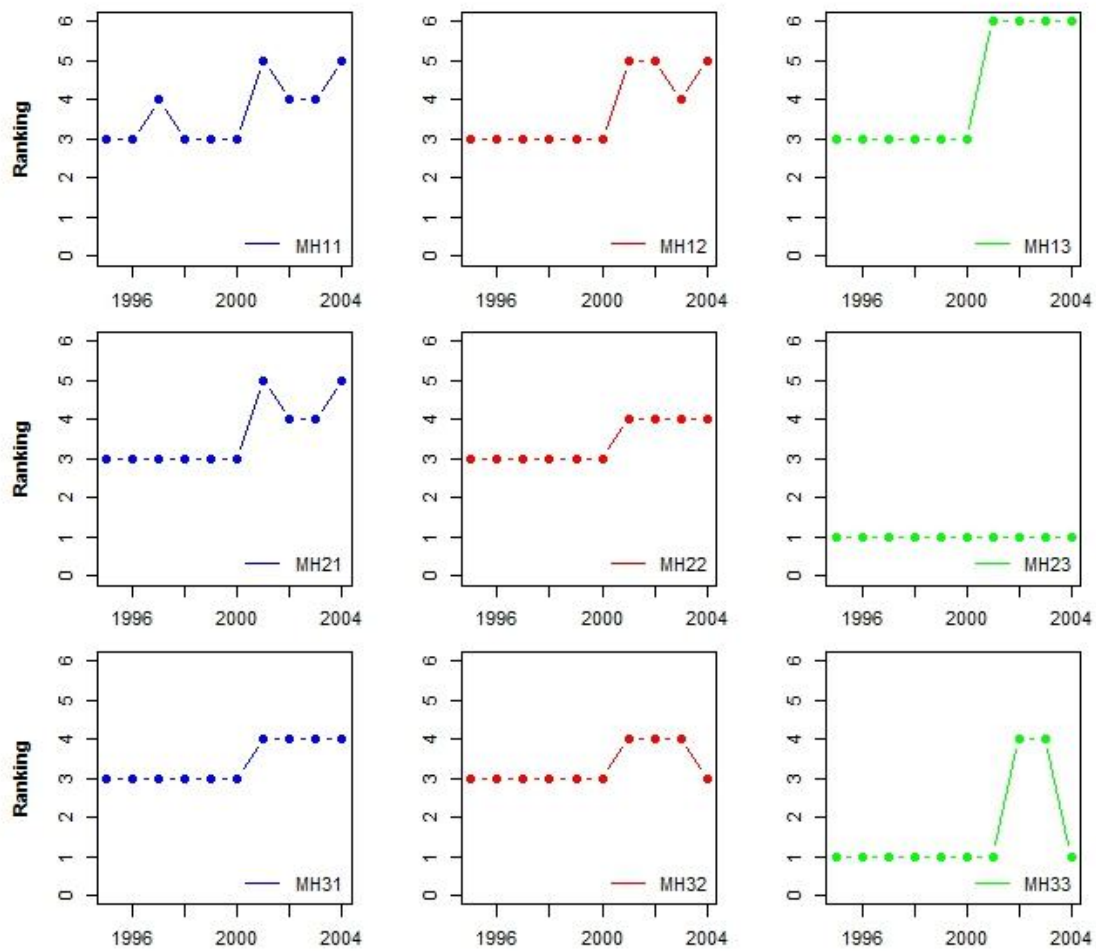


ANEXO 6: SEGUIMIENTO DE LAS UNIDADES EN EL TIEMPO

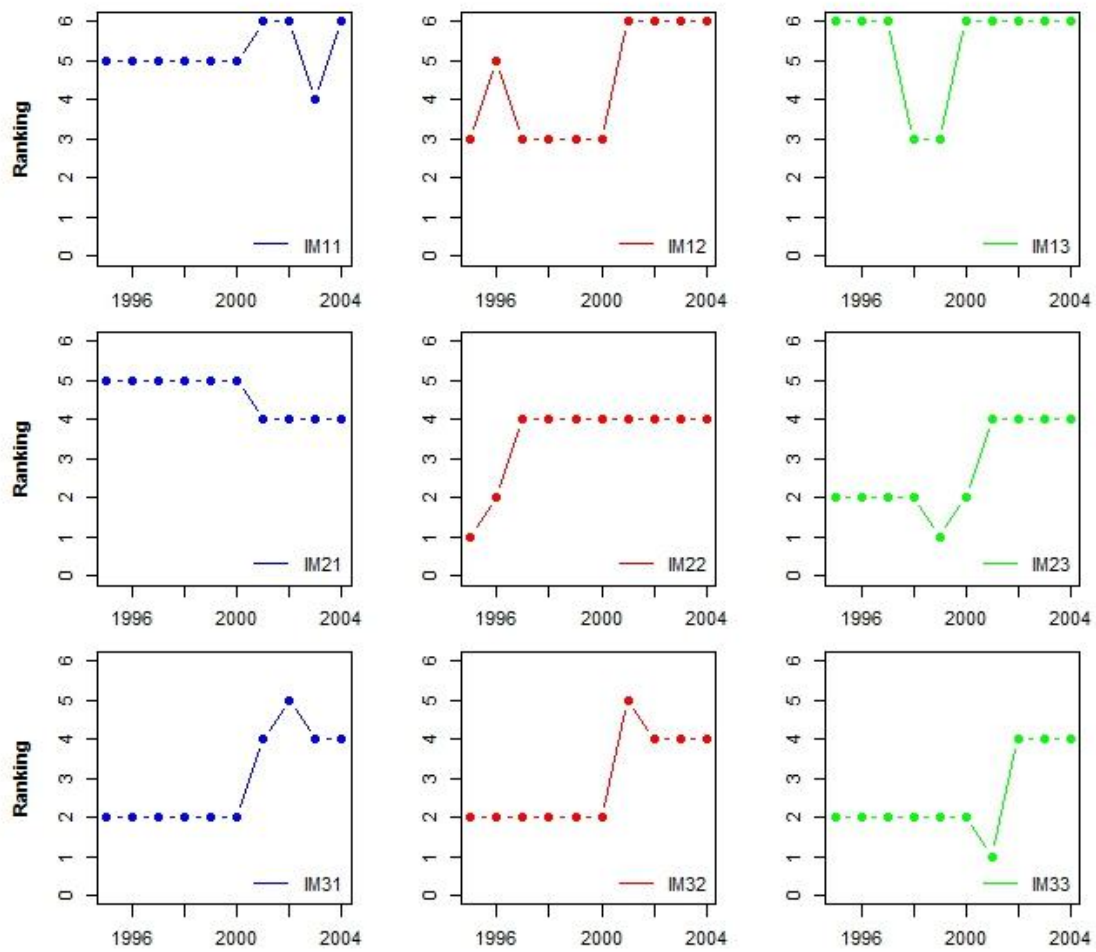
Seguimiento de las unidades en el tiempo Interior-Hombres



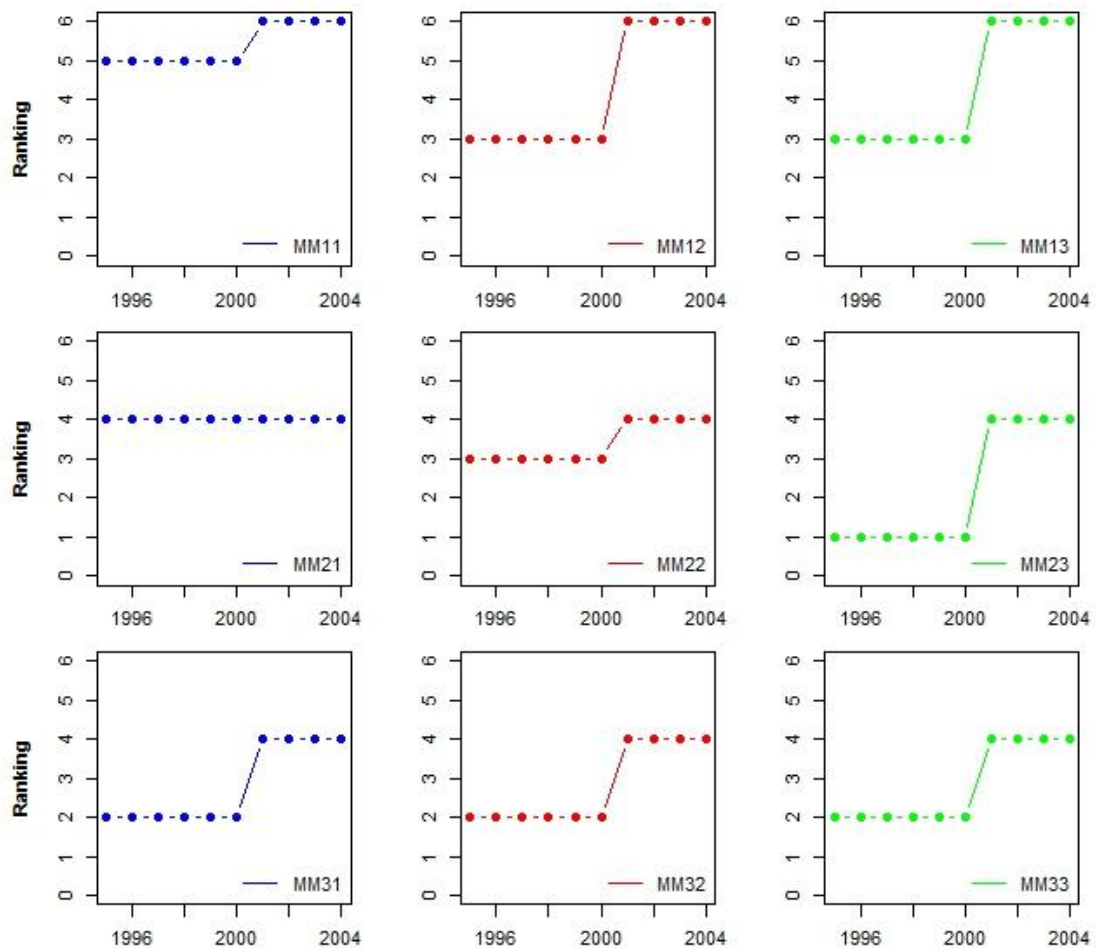
Seguimiento de las unidades en el tiempo Montevideo-Hombres



Seguimiento de las unidades en el tiempo Interior-Mujeres



Seguimiento de las unidades en el tiempo Montevideo-Mujeres



ANEXO 7: SALIDAS DEL TEST DE MARDIA

7.A - TESTEO DE NORMALIDAD MULTIVARIADA PARA TODA LA POBLACIÓN

Mardia's Multivariate Normality Test

data : .

g1p : 75.56817
chi.skew : 4534.09
p.value.skew : 0

g2p : 251.8035
z.kurtosis : 12.46184
p.value.kurt : 0

chi.small.skew : 4576.946
p.value.small : 0

Result : Data are not multivariate normal.

7.B - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 1

Mardia's Multivariate Normality Test

data : .

g1p : 99.12792
chi.skew : 1734.739
p.value.skew : 2.857575e-120

g2p : 268.15
z.kurtosis : 10.687
p.value.kurt : 0

chi.small.skew : 1791.061
p.value.small : 1.232542e-128

Result : Data are not multivariate normal.

7.C - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 2

Mardia's Multivariate Normality Test

```
-----  
data : .  
  
g1p          : 124.6283  
chi.skew     : 851.6266  
p.value.skew : 2.011102e-14  
  
g2p          : 243.3391  
z.kurtosis   : 2.925232  
p.value.kurt : 0.003441997  
  
chi.small.skew : 922.7286  
p.value.small  : 3.329945e-20  
  
Result       : Data are not multivariate normal.  
-----
```

7.D - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 3

Mardia's Multivariate Normality Test

```
-----  
data : .  
  
g1p          : 95.39472  
chi.skew     : 667.763  
p.value.skew : 0.001130622  
  
g2p          : 219.0086  
z.kurtosis   : -0.7641452  
p.value.kurt : 0.4447807  
  
chi.small.skew : 722.1783  
p.value.small  : 4.028966e-06  
  
Result       : Data are not multivariate normal.  
-----
```

7.E - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 4

Mardia's Multivariate Normality Test

```
-----  
data : .  
  
g1p          : 117.0083  
chi.skew     : 1677.12  
p.value.skew : 7.620153e-112  
  
g2p          : 267.4753  
z.kurtosis   : 9.52408  
p.value.kurt : 0  
  
chi.small.skew : 1743.641  
p.value.small  : 1.387116e-121  
  
Result       : Data are not multivariate normal.  
-----
```

7.F - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 5

Mardia's Multivariate Normality Test

```
-----  
data : .  
  
g1p          : 113.8093  
chi.skew     : 986.3473  
p.value.skew : 5.579368e-26  
  
g2p          : 245.3725  
z.kurtosis   : 3.640725  
p.value.kurt : 0.0002718713  
  
chi.small.skew : 1051.185  
p.value.small  : 2.230625e-32  
  
Result       : Data are not multivariate normal.  
-----
```

7.G - TESTEO DE NORMALIDAD MULTIVARIADA PARA EL GRUPO 6

Mardia's Multivariate Normality Test

data : .

g1p : 107.8603
chi.skew : 611.2082
p.value.skew : 0.06608338

g2p : 218.9001
z.kurtosis : -0.7024729
p.value.kurt : 0.4823843

chi.small.skew : 672.8278
p.value.small : 0.0007164566

Result : Data are multivariate normal.

ANEXO 8: COEFICIENTES ESTIMADOS DEL MODELO MULTILOGIT (GRUPO DE REFERENCIA: GRUPO 1)

Coeficientes estimados					
	2	3	4	5	6
Intercepto	-38.05	20.09	13.05	22.08	34.41
jefe	0.12	-0.04	-0.02	0.06	-0.60
size	0.00	0.14	0.13	0.10	-0.11
privado	0.24	-0.44	-0.16	-0.18	0.04
publico	0.16	-0.28	-0.34	-0.24	-0.41
cpsl	-0.57	-0.22	-0.04	-0.44	-0.78
cpcl	0.72	0.04	-0.14	-0.51	0.42
profytec	0.24	0.15	0.25	0.28	0.41
oficina	-0.10	-0.07	0.22	0.15	0.36
manual	-0.20	-0.65	-0.11	-0.03	-0.01
indust	0.28	0.54	0.01	0.15	0.33
comercio	0.23	0.08	0.05	0.15	-0.36
sfinan	-0.52	-1.01	-2.19	-2.68	-2.55
sperson	-0.02	0.09	-0.07	0.01	-0.76
ingreso	0.85	1.83	1.81	-2.90	-2.96

ANEXO 9: FUNCIÓN DE CROSS-VALIDATION¹⁴

```

totalAccuracy <- c()
cv <- 359
cvDivider <- floor(nrow(ech_wa) / (cv+1))

for (cv in seq(1:cv)) {
  # assign chunk to data test
  dataTestIndex <- c((cv * cvDivider):(cv * cvDivider +
cvDivider))
  dataTest <- ech_wa[dataTestIndex,]

  # everything else to train
  dataTrain <- ech_wa[-dataTestIndex, ]
  crossval <- multinom(grupos_agnes_mah_redu_wa_6~ .,
data=dataTrain, maxit=500, trace=F)
  pred <- predict(crossval, newdata=dataTest, type="class"

  # classification error
  cv_ac <- postResample(dataTest$grupos_agnes_mah_redu_wa_6,
pred)[[1]]
  print(paste('Current Accuracy:',cv_ac,'for CV:',cv))
  totalAccuracy <- c(totalAccuracy, cv_ac)
}

mean(totalAccuracy)

```

¹⁴ La misma fue programada por Manuel Amunategui y puede encontrarse en su página de github (<http://amunategui.github.io/multinomial-neuralnetworks-walkthrough/>)