ANÁLISIS MULTIVARIADO I

CLUSTERING Y DISCRIMINANTE

Lucía Coudet
Daniel Czarnievicz
Ramón Talvi

Pseudo Panel de la ECH para el período 1995-2004

	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****		*% de trabajadores precarios*
9	**privado**	Categoría de ocupación	*% de trabajadores del sector privado*
10	**publico**		*% de trabajadores en el sector público*
11	**cpsl**		*% de trabajadores por cuenta propia sin local*
12	**cpcl**		*% de por cuenta propia con local*
13	**profytec**	Condición de ocupación	*% de profesionales y técnicos*
14	**oficina**		*% de empleados de oficina*
15	**manual**		*% de empleados manuales*
16	**indust**	Rama del establecimiento	*% 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*

Investigación en dos etapas

1. Estudio del agrupamiento de las observaciones en función de las variables de Situación de Empleo - Clustering

2. Contribución, o no, del resto de las variables a explicar los grupos formados en la etapa 1 - Análisis discriminante

Matriz de correlaciones

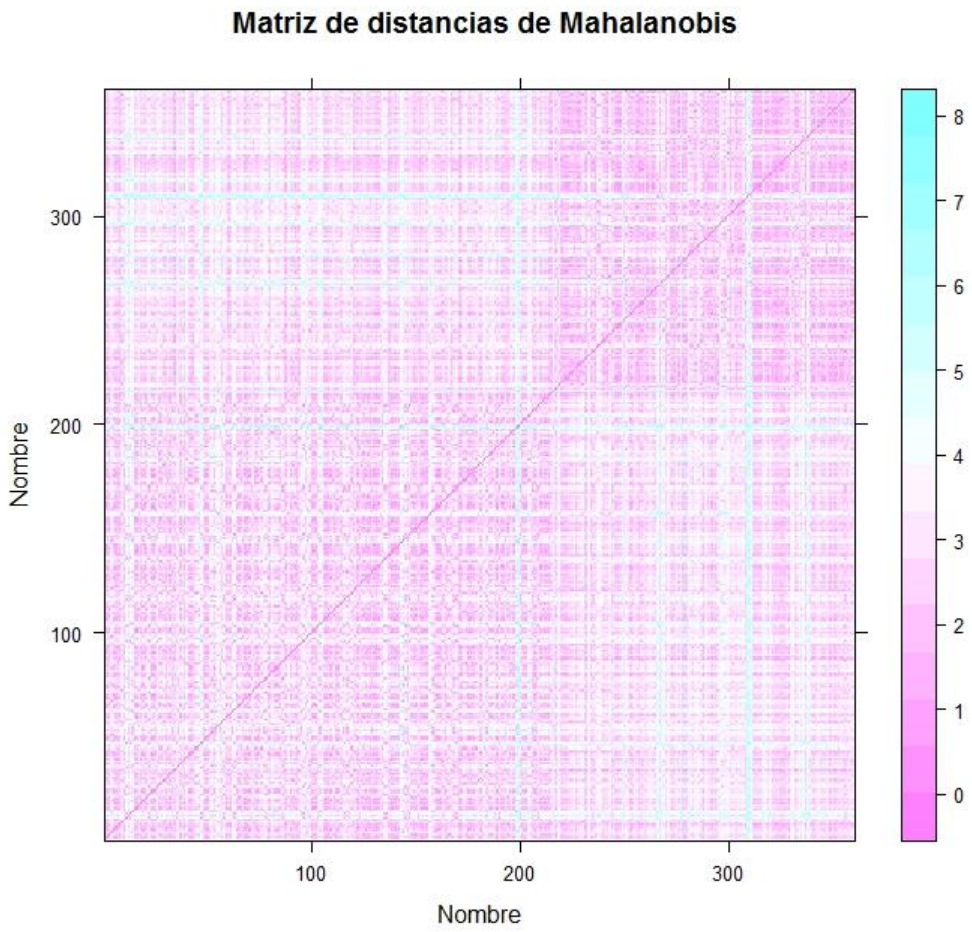
	jefe	desemp	size	tparcial	multiemp	privado	publico	cpsl	cpcl	profytec	oficina	manual	indust	comercio	sfinan	sperson	subemp	precario	ingreso
jefe	1.00	-0.69	-0.03	-0.54	0.22	-0.61	0.17	0.44	0.55	-0.03	-0.32	0.45	0.19	-0.20	0.10	-0.33	-0.43	-0.45	0.44
desemp	-0.69	1.00	0.27	0.65	-0.48	0.77	-0.53	-0.09	-0.61	-0.31	0.10	-0.17	0.02	0.40	-0.25	0.20	0.65	0.76	-0.56
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	-0.54
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	-0.28
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	0.83
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	-0.56
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	0.57
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	-0.33
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	0.23
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	0.69
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	0.15
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	-0.34
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	-0.33
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	-0.52
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	0.71
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	-0.31
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	-0.44
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	-0.70
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	1.00

Correlaciones entre las variables de situación de empleo

	desemp	tparcial	multiemp	subemp	precario
desemp	1.00	0.65	-0.48	0.65	0.76
tparcial	0.65	1.00	-0.11	0.65	0.48
multiemp	-0.48	-0.11	1.00	-0.29	-0.58
subemp	0.65	0.65	-0.29	1.00	0.63
precario	0.76	0.48	-0.58	0.63	1.00

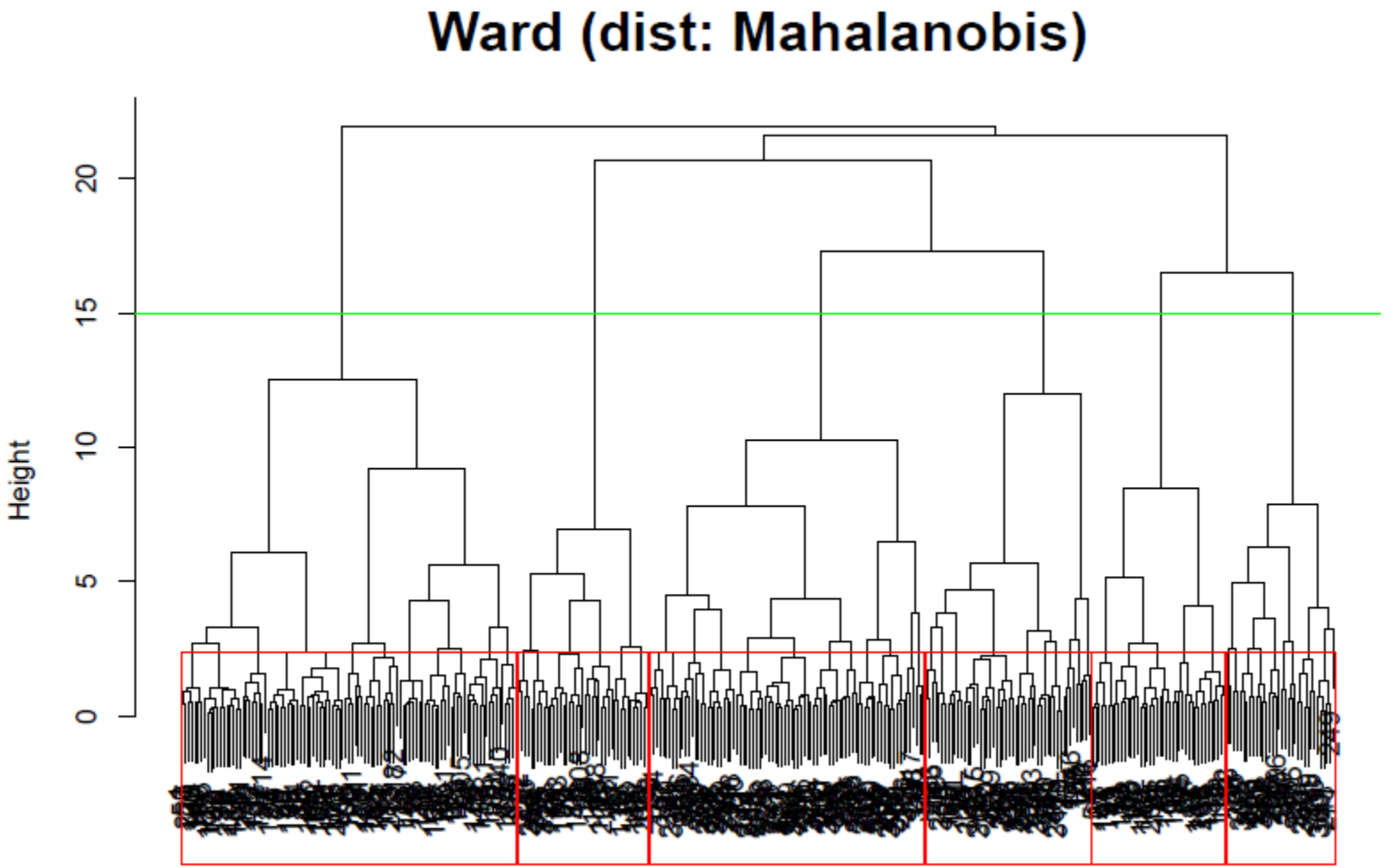
Debido a la alta correlación entre las variables, decidimos utilizar la distancia de Mahalanobis como métrica

Matriz de distancias. Métrica: Mahalanobis

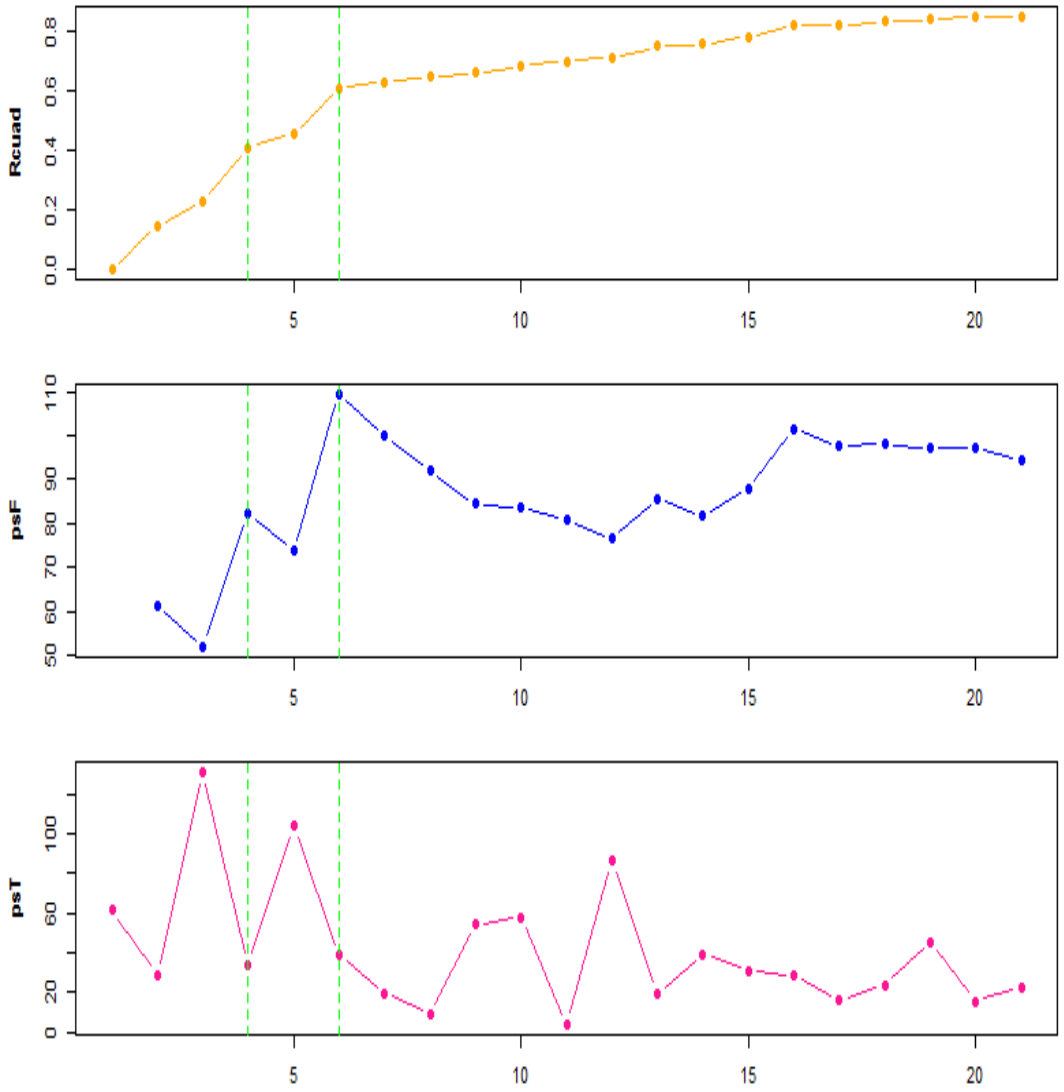


Clustering por Método de Ward

Dendograma

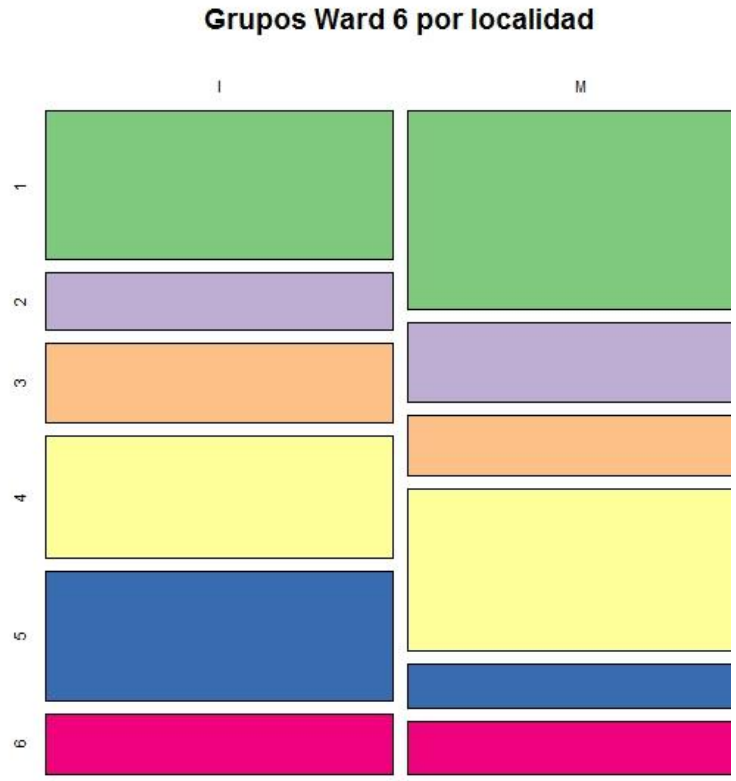


Indicadores (R^2, pseudo-F, pseudo-t)



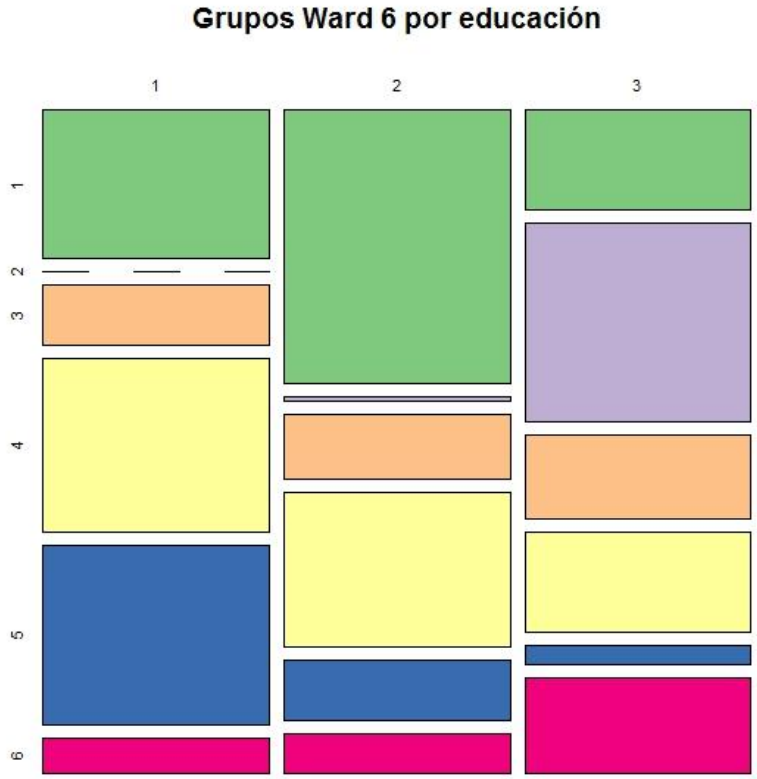
Caracterización de los grupos

En función de los componentes



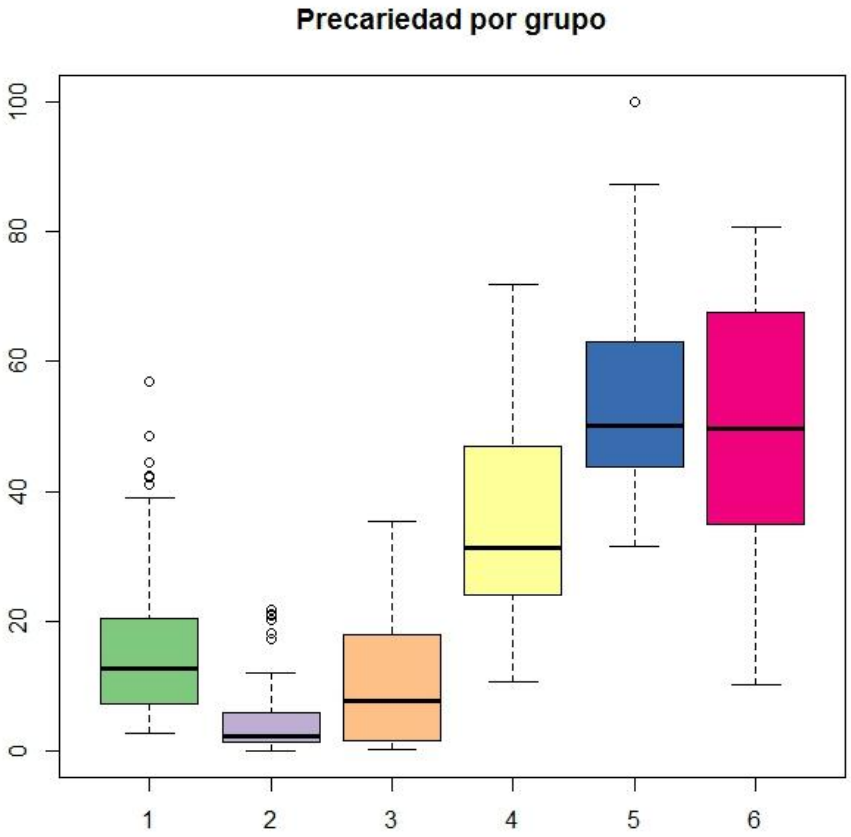
Caracterización de los grupos

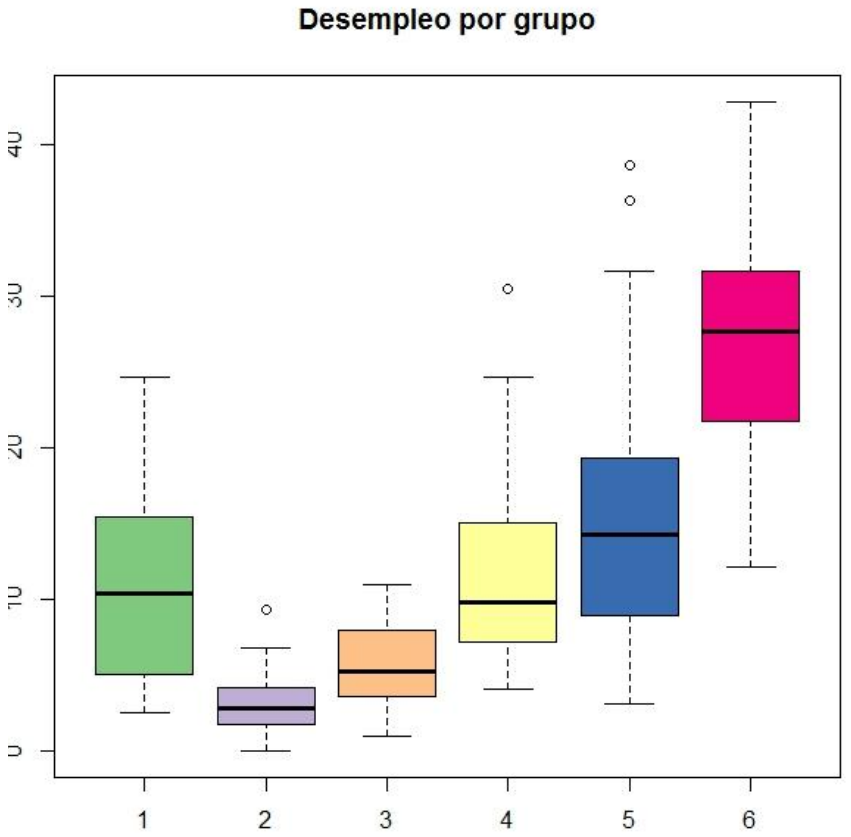
En función de los componentes



Caracterización de los grupos

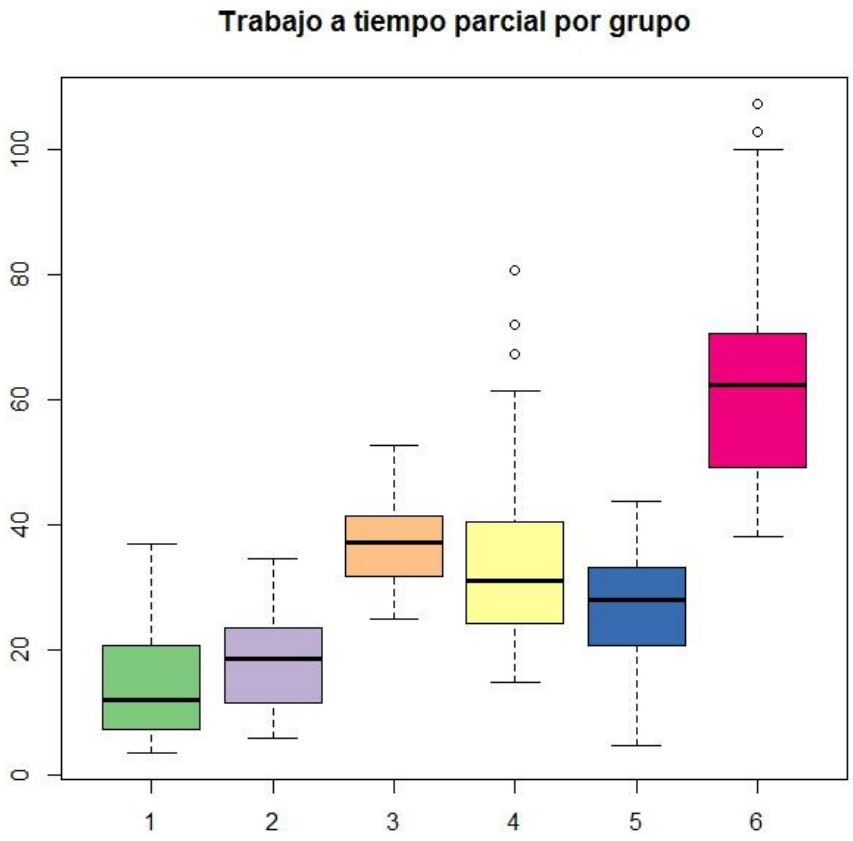
En función de las variables de análisis

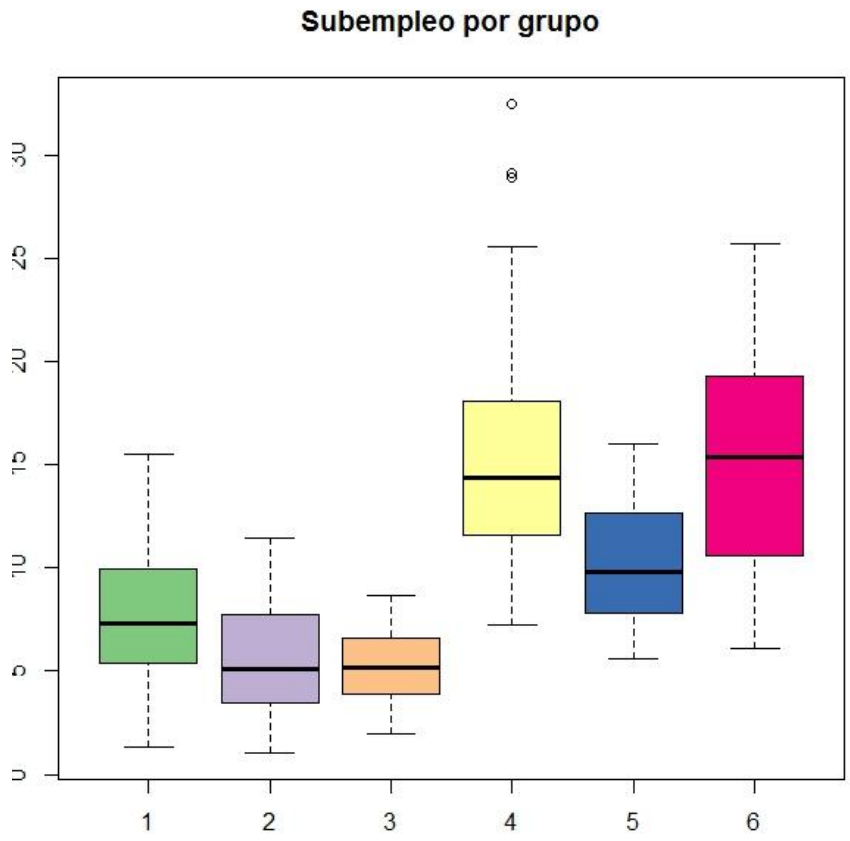




Caracterización de los grupos

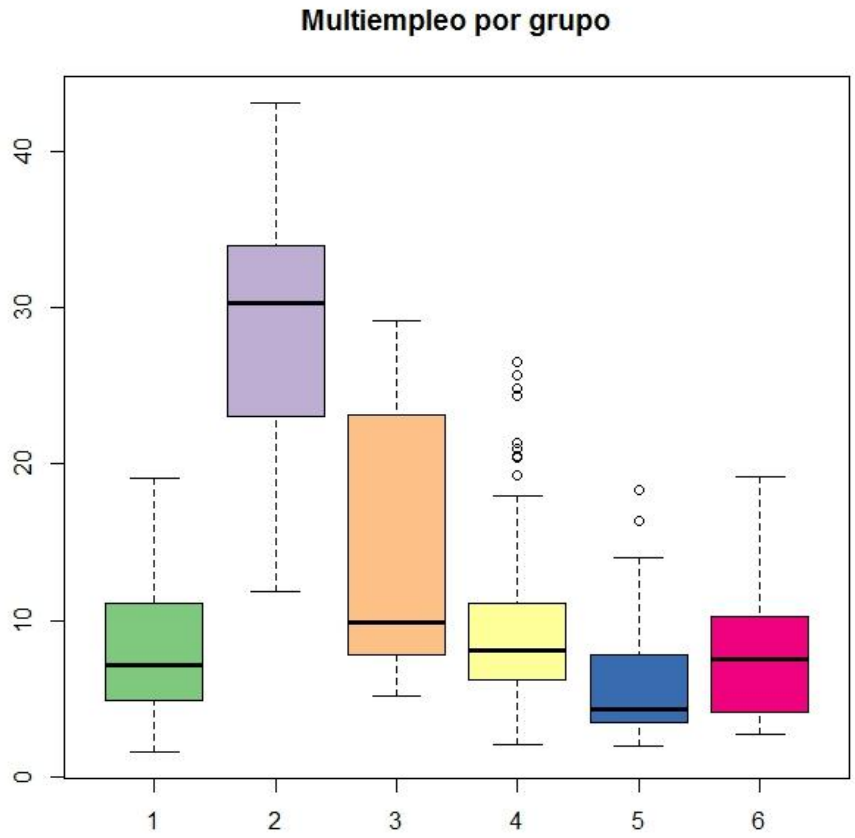
En función de las variables de análisis





Caracterización de los grupos

En función de las variables de análisis



Caracterización de los grupos (resúmen)

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

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%

Caracterización de los grupos

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.

Seguimiento de las unidades en el tiempo

Se buscó hacer un seguimiento de las distintas unidades (los 36 grupos) en el tiempo para evaluar su performance

Se tuvieron en cuenta las variables: desempleo, subempleo, y precariedad, dado que reflejan situaciones involuntarias de los individuos

Cada una de las variables recibió un puntaje del 1 al 6, en función del valor medio de dicha variable en el grupo, en el cual 1 reflejaba la situación más deseable, y 6 la menos deseable

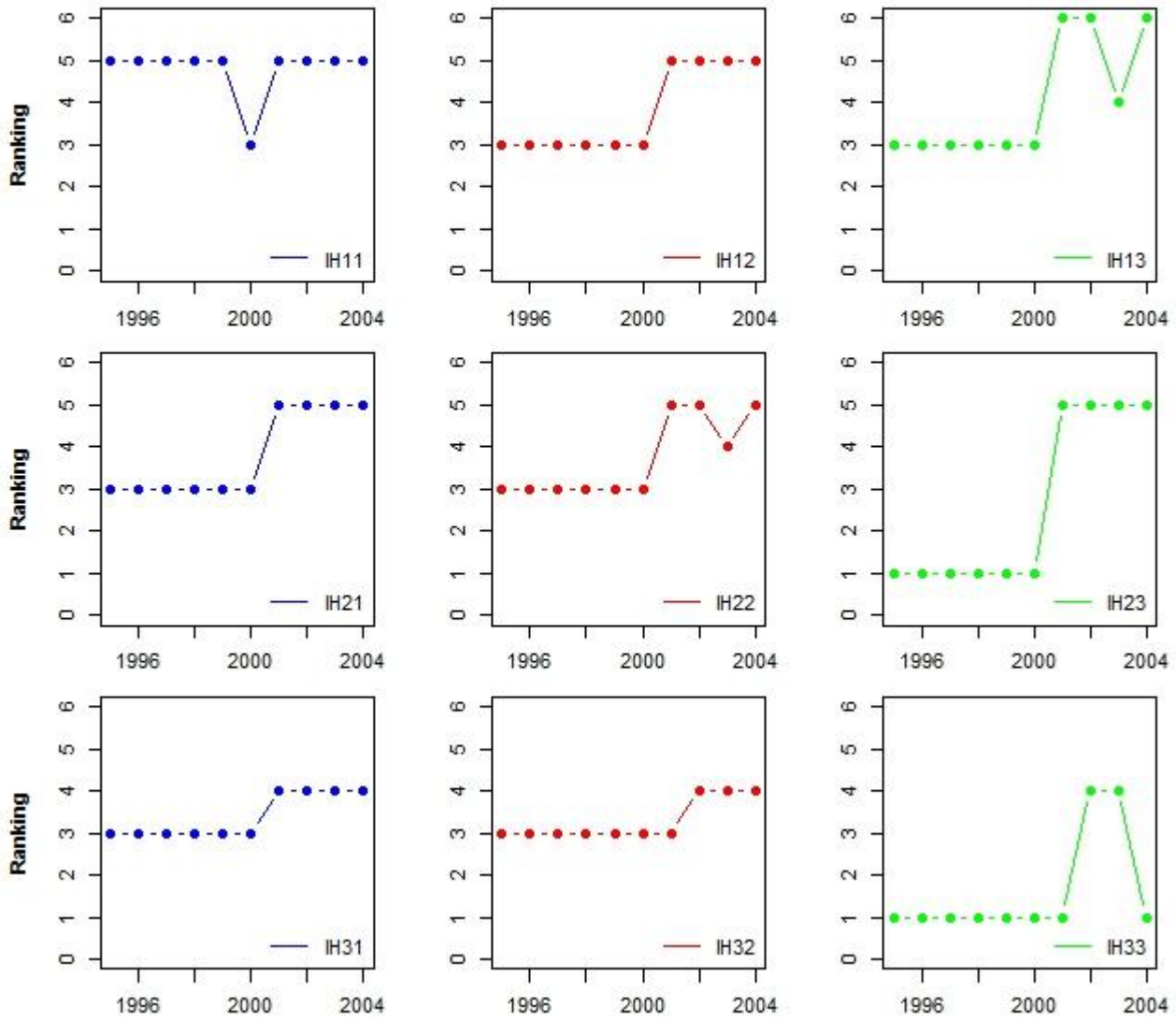
La posición global del grupo en el ranking se obtuvo mediante una Suma de Bentham (sumas horizontales) de la posición que cada grupo obtuvo en el ranking de cada variable

Seguimiento de las unidades en el tiempo

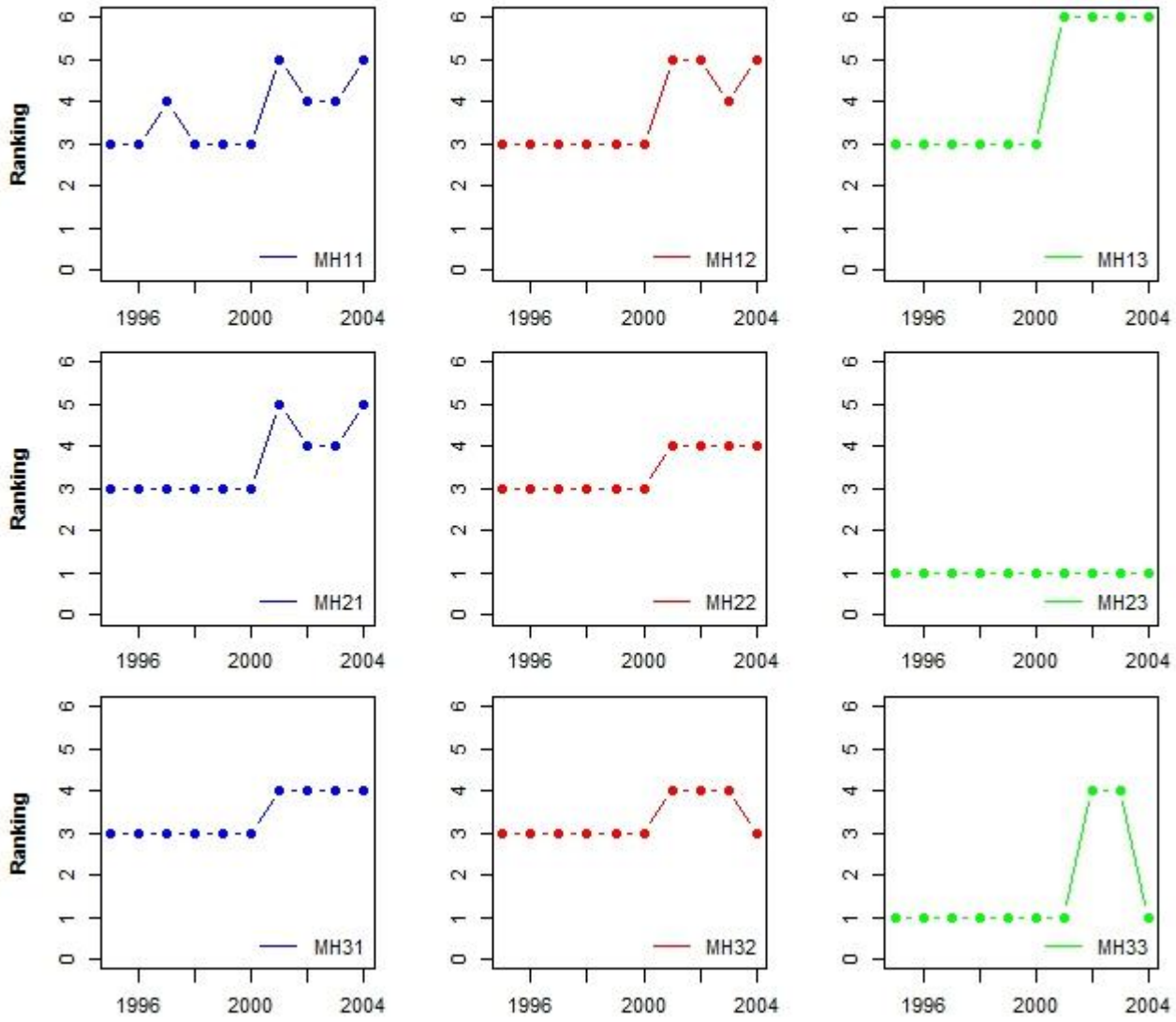
Ranking					
Grupos	**Desempleo**	**Subempleo**	**Precariedad**	**Suma**	**Posición**
1	3	3	3	9	3
2	1	2	1	4	1
3	2	1	2	5	2
4	4	5	4	13	4
5	5	4	6	15	5
6	6	6	5	17	6

Se concluyó que los integrantes del grupo 2 contaban con la Situación de Empleo más deseable, mientras que los del grupo 6 presentaron la situación menos deseable

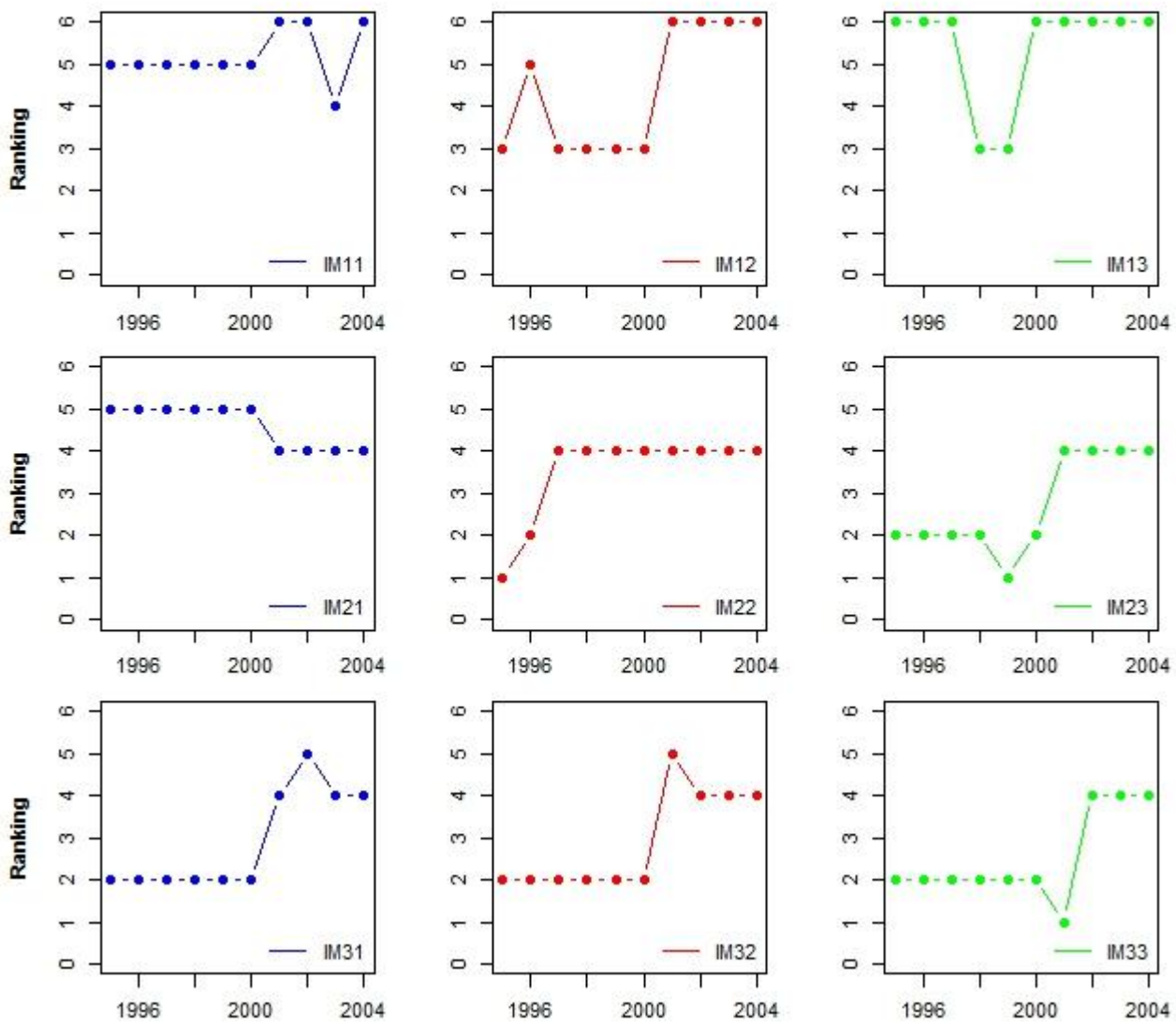
Seguimiento de las unidades en el tiempo
Interior-Hombres



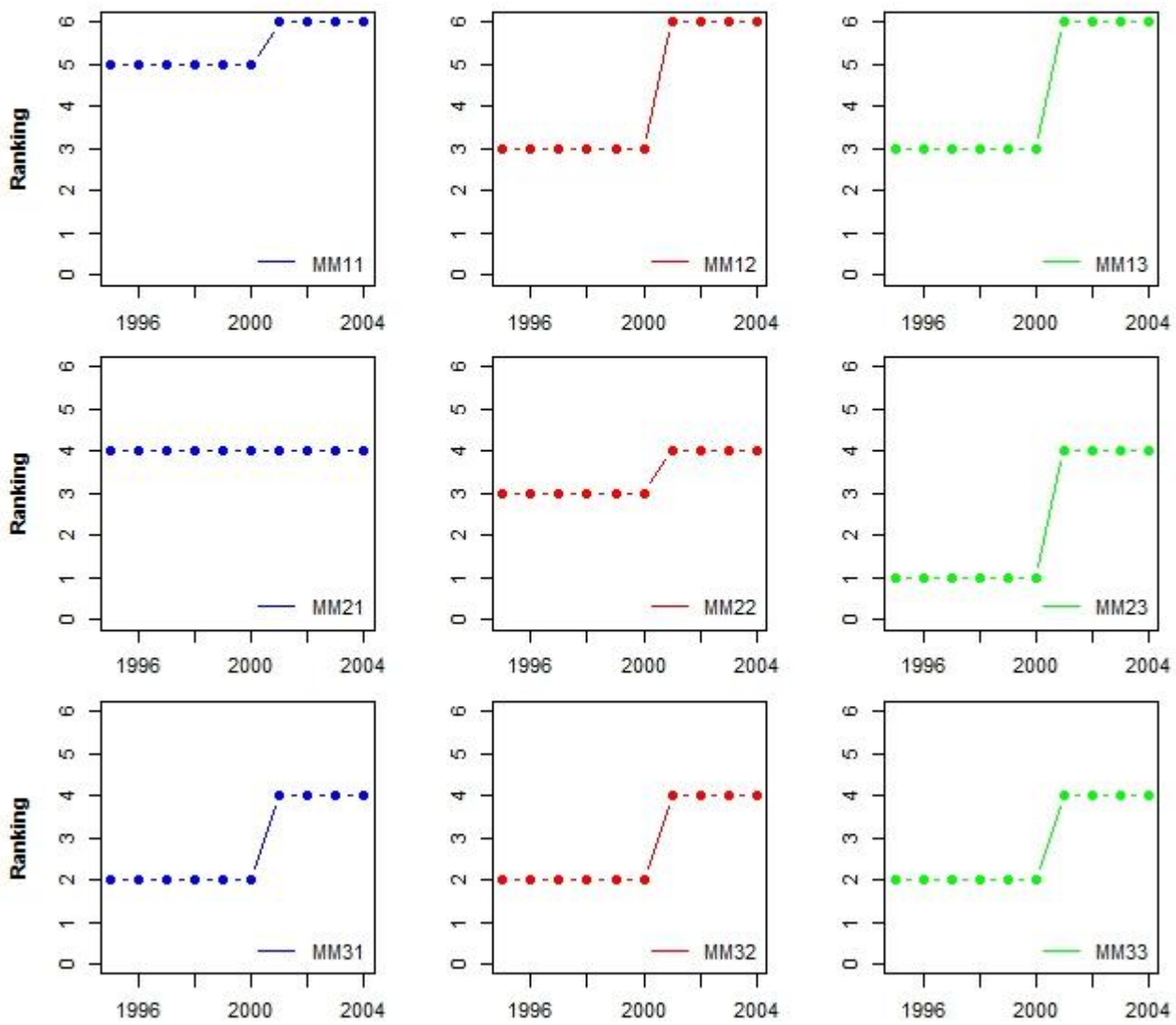
Seguimiento de las unidades en el tiempo
Montevideo-Hombres



Seguimiento de las unidades en el tiempo
Interior-Mujeres



Seguimiento de las unidades en el tiempo
Montevideo-Mujeres



Seguimiento de las unidades en el tiempo

Dos resultados se destacan:

1. Si bien la gran mayoría de las unidades sufren una desmejora durante la recesión y crisis de principio de siglo XXI, las que no cayeron, no fueron las de nivel educativo 3 (IH11, MH23, IM21, IM22, y MM21).

2. Las unidades que más rápidamente se recuperaron fueron los hombres de mayor edad y nivel educativo, independientemente de la localidad de residencia (IH33, MH33, MH32).

Análisis Discriminante

Se buscó analizar si las variables explicativas de las restantes clases (hogar, categoría de ocupación, condición de ocupación, y rama del establecimiento), explicaban la pertenencia a los grupos previamente definidos (grupos Ward 6)

El test de Mardia rechazó la hipótesis nula de normalidad en los grupos, por lo que se estimó un discriminante multilogístico (dado que la variable de respuesta, “grupos”, tiene 6 niveles)

Se tomó como grupo de referencia al grupo 1

El objetivo es crear un algoritmo que nos permita clasificar nuevas observaciones

Predicciones correctas

Grupos	Valores Predichos 1	2	3	4	5	6	Total
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

Porcentaje de observaciones correctamente clasificadas: 87.22%

Grupos	Valores Predichos 1	2	3	4	5	6	Total
1	93.3%	0.0%	0.0%	3.8%	1.9%	1.0%	29.2%
2	2.4%	90.2%	2.4%	4.9%	0.0%	0.0%	11.4%
3	2.4%	2.4%	92.9%	2.4%	0.0%	0.0%	11.7%
4	4.7%	2.3%	1.2%	79.1%	11.6%	1.2%	23.9%
5	7.7%	0.0%	0.0%	15.4%	76.9%	0.0%	14.4%
6	0.0%	0.0%	0.0%	2.9%	2.9%	94.1%	9.4%
Total	30.0%	11.1%	11.4%	23.3%	14.7%	9.4%	87.8%

Cross-validation

```
totalAccuracy <- c()
cv <- 360

for (cv in seq(1:cv)) {
      # assign chunk to data test
      dataTestIndex <- cv
      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)
```

La función realiza Cross Validation por leave-one-out y obtiene una media de predicciones correctas de 79.22%

Conclusiones generales

- Clustering
  - ✓ Solo en uno de los grupos (grupo 5) la característica localidad contribuyó a la clasificación.
  - ✓ Las restantes características (sexo, edad, y nivel educativo) fueron relevantes para la clasificación en todos los grupos.
  - ✓ Se encuentra una desmejora generalizada en la calidad de la situación de empleo como efecto de la recesión y crisis de principios de siglo.
  - ✓ El efecto fue generalizado, pero lograron recuperarse más rápidamente los grupos de mayor nivel educativo.
- Análisis Discriminante
  - ✓ Las variables no incluidas en Clustering contribuyen a explicar la formación de grupos.

Bibliografía

- 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