IV Congreso de Economía Regional de Castilla y León **Burgos 1994, Comunicaciones 2:** 836-850

UN ANALISIS DE SEGMENTACION EN EL ESTUDIO DEL PARO EN SALAMANCA

Guillermo RAMIREZ Santiago VICENTE TAVERA
Purificación GALINDO VILLARDÓN Maura VASQUEZ
Depto. de Estadística y Matemática Aplicadas.
Universidad de Salamanca

1.- INTRODUCCION

Los métodos **AID** (Automatic Interaction Detection), propuestos inicialmente por Morgan y Sonquist (1963), y sus diferentes versiones posteriores, han sido aplicados con bastante éxito en muy diversas disciplinas científicas. Estos métodos tratan con datos tipo regresión, una variable dependiente y un conjunto de predictores que son de tipo cualitativo. Tienen como característica particular el que su aplicación está prácticamente libre de los supuestos usualmente requeridos por los métodos basados en el modelo lineal.

El objetivo básico de los métodos AID es agrupar los predictorescon perfil similar para la variable dependiente.

El procedimiento general que utilizan los métodos **AID** es de tipo iterativo. Los grupos finales quedan formados después de un proceso por etapas, en cada una de las cuales se selecciona el mejor predictor (en un sentido convenientemente establecido) y con éste se particiona el grupo considerado en varios subgrupos, definidos por las categorías del predictor elegido. Cada subgrupo es a su vez analizado en la misma forma, con los predictores restantes, y se continúa el proceso hasta que no sea posible subdividir ninguno de los subgrupos obtenidos. Debido a esta estrategia de particiones sucesivas de los datos, se conoce a estos métodos con la denominación general de Análisis de Segmentación.

La versión propuesta por Kass (1980) denominada **CHAID** (CHi-square AID) se aplica en el caso de variable dependiente de tipo cualitativo, y debe su nombre a que utiliza el contraste de independencia chi-cuadrado en las diferentes fases del proceso de segmentación.

En este trabajo vamos a desarrollar una aplicación de la técnica **CHAID** a unos datos sobre desempleo recolectados por la Oficina San José del INEM en la ciudad de Salamanca.

Se ha elaborado un programa estadístico, expresamente al efecto, para ordenador PC, en lenguaje Pascal, versión Turbo 6.0 el cual no ha sido aún publicado.

2.- LA TECNICA CHAID

Antes de presentar los resultados obtenidos, creemos conveniente hacer una breve exposición de la técnica CHAID.

2.1.- Fases.

Como ya hemos dicho, se trata de un proceso iterativo cuyas fases son las siguientes:

a.- Agrupación de categorías.

En esta fase se agrupan las categorías de cada predictor cuando éstas tengan un perfil similar en la variable dependiente. Para ello se cruza cada par de categorías de cada predictor con la variable dependiente, y se calcula el estadístico chi-cuadrado del contraste de independencia. El par con mayor valor p, siempre que no sea significativo, formará una sola categoría con los dos valores fusionados. Si se ha fusionado un par de categorías, se repite el proceso con las nuevas categorías agrupadas para ver si se procede o no con nuevas fusiones. Este proceso termina cuando no se pueden realizar más fusiones, bien porque todos los estadísticos chi-cuadrado resultan significativos, o porque ya se han agrupado todas las categorías del predictor, en cuyo caso éste se descarta ya que no produce una discriminación significativa en la variable dependiente, dentro de este grupo. Al final de esta fase, las c categorías iniciales del predictor considerado quedan agrupadas en d categorías ($1 \leq d \leq c$).

Las agrupaciones se llevan a cabo según sea el tipo de predictor empleado, a saber predictores monótono,libres o flotantes. (para mayor información ver ESCOBAR (1992)).

b.- Selección del mejor predictor.

El objetivo de esta fase es encontrar, entre todos los predictores previamente agrupados, el más significativo, es decir aquél que mejor discrimine a los sujetos según la variable

dependiente. Para ello identificamos el predictor con menor valor p, siempre que sea significativo. Para evitar el problema del incremento en el riesgo tipo I, se utiliza la clásica corrección de Bonferroni.

c.- Segmentación.

El grupo considerado se segmenta en d subgrupos, definidos respectivamente por cada una de las categorías del predictor seleccionado en (b).

d.- Iteración del proceso.

Para cada subgrupo formado en (c) se repiten los pasos (a), (b) y (c).

El proceso termina cuando no hay predictores significativos en ninguno de los grupos. Suelen definirse además otras reglas de finalización relacionadas con el tamaño de los grupos y los niveles de segmentación.

2.2.- Arbol de Segmentación

El resultado final del Análisis de Segmentación suele representarse en un diagrama de árbol en el cual se muestra gráficamente el proceso de las sucesivas segmentaciones. Este diagrama ofrece un resumen parsimonioso de los datos, de gran interés descriptivo y exploratorio. En cada "nudo" del árbol se indica el predictor que produce la segmentación, en cada rama se indica la categoría que define el subgrupo y su tamaño, y dentro de cada rectángulo se indica la distribución porcentual de la variable dependiente en ese subgrupo.

3.- DATOS ANALIZADOS

La base de datos que nos ha suministrado la Delegación Provincial del INEM de Salamanca, contiene información sobre 38 variables y 9439 individuos que demandan trabajo en Salamanca y poblaciones vecinas. Dado que nuestro trabajo tiene un objetivo claramente metodológico, y teniendo en cuenta las limitaciones del programa, hemos efectuado el análisis sobre una muestra de 2000 individuos elegida aleatoriamente de los 6372 esidentes en la ciudad de Salamanca (códigos postales 37006, 37007, 37008 y 37009). Para nuestro estudio hemos hecho también una reducción en el número de variables consideradas al ejegir las 11 variables con mayor interes. Los demandantes de trabajo pueden registrar hasta 4 títulos académicos, hasta 4 ocupaciones (cada una con su respectivo nivel y experiencia profesional) y hasta 4 idiomas. De cada una de estas 5

variables hemos seleccionado sólo una. Hemos descartado otras variables como Código de Oficina, Minusvalías y otras correspondientes a situaciones especiales.

Algunas de las variables tuvieron que ser sometidas a un proceso de recodificación debido a la elevada cantidad de códigos necesaria para mantener información muy detallada pero no para nuestros fines.

Las 11 variables consideradas junto con su denominación y catagorías pueden verse en la siguiente tabla:

1ª Sexo (SEXO)

Categorías: Hombre, Mujer

2ª Edad (EDAD)

Categorías: 16 - 25, 25 - 34, 34 - 43, 43 - 52, 52 - 61 y 61 - 74

3ª Nivel Académico (NIVEACAD)

Categorías:	**Nombre abreviado**
Estudios Primarios	Primarios
Certif. de Escolaridad	Cert. Esc.
Formación Profesional I	FPI
Graduado Escolar	Grad. Esc.
Formación Profesional II	FPII
BUP, COU	BUP-COU
Titulado Medio	Tit. Med.
Titulado Superior	Tit. Sup.

4ª Titulación (TITUL)

Categorías:	**Nombre abreviado**
Sin Título	Sin Tit.
Formación Profesional I	FPI
Formación Profesional II	FPII
Cursos INEM	INEM
Ingenieros Técnicos	Ing. Tec.
Diplomados	Diplomad
Ingenieros	Ingeniero
Economistas	Economist
Abogados	Abogado
Humanistas	Humanist
Científicos	Científic
Profes. de la Salud	Salud

5ª Ocupación (OCUPAC)

Categorías:	**Nombre abreviado**
Profes., Tecnic. y simil.	Pro-Tec
Directores	Director
Administrativos	Administ
Comerciantes	Comerc
Hostelería	Hosteler
Agricultores	Agricult
Minería, Textiles y Alim.	Min-Text
Calzado, Mecán. y Elect.	Cal-Mec
Caucho, Artes Graf., Constr. y Oper. de Maq.	Con-Maq

6ª Nivel Profesional (NIVELPRO)

Categorías:	**Nombre abreviado**
Profesores y otros	Prof-Otros
Directores	Director
Mandos Medios	Medios
Técnicos	Técnico
Maestro	Maestro
Oficiales	Oficial
Aprendices y Peones	Peones

7ªExperiencia Profesional (en años) (EXPERPRO)

Categorías: 0 - 3 ; 3 - 6 y 6 - 9

8ª Causa del Cese (CAUSCESE)

Categorías:	**Nombre abreviado**
No Consta	NC
Cese Voluntario	Cese Vol.
Fin de Contrato	Fin Cont.
Despido	Despido
Regulación de empleo	Reg. Emp.
Inactividad en contrato fijo discontinuo	Inactiv

9ª Idioma (IDIOMA)

Categorías: Ninguno ; Inglés ; Francés y Otros

10ª Actividad Económica (ACTIVECO)

Categorías:	**Nombre abreviado**
Agricultura	Agr.
Energía, Metales y Maq.	Ene-Met
Indust. Manufacturera	Manufac
Construcción	Construc
Comercio y Hostelería	Com-Host
Transporte y Comunicac.	Tran-Com
Instituciones Financieras	Financ
Servicios	Servic

11ª Meses en paro (Hasta el 9/6/94) (MESESPAR)

Categorías: <12 ; 12-24 y >24

4.- RESULTADOS DEL ANALISIS

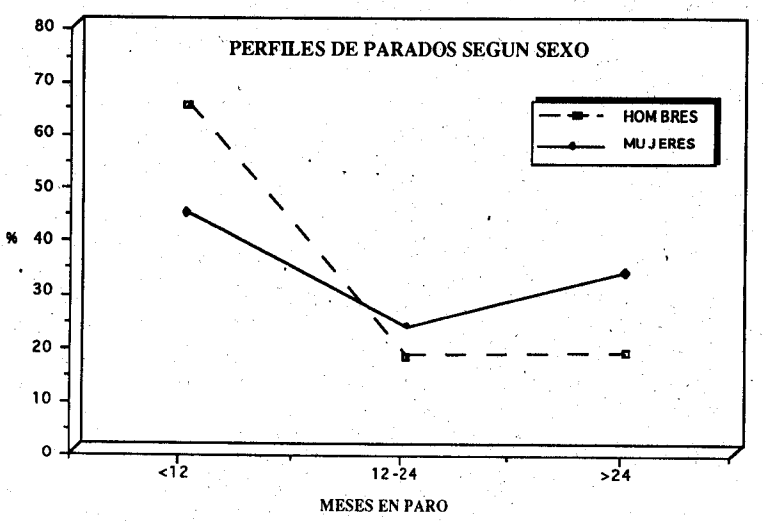
La variable dependiente escogida para el análisis es MESESPAR, es decir, el número de meses en paro que tiene el demandante de ocupación desde su registro en el INEM hasta la fecha tope 9/6/94. Esta varaiable fue recodificada en tres categorías dado que existía una gran asimetría en la distribución por meses lo que dificultaba el análisis que se pretendía llevar a cabo.

Tras agrupar los predictores con perfil similar para la variable dependiente, se procedió a la búsqueda del mejor predictor, es decir aquella variable para la cual el valor de la significación fué menor. Las variables significativas fueron: SEXO; OCUPAC; EDAD; TITUL; ACTIVECO; NIVEACAD; CAUSACESE; NIVELPRO e IDIOMA. De todas ellas se eligió al SEXO por tener nayor significación y mayor valor en el Coeficiente de contingencia.

La muestra en estudio se segmenta en dos grupos de 964 hombres y 1036 mujeres.

El gráfico nº 1 es un diagrama de perfiles en el que se pueden apreciar las diferencias en los porcentajes de parados entre ambos sexos.

GRAFICO 1

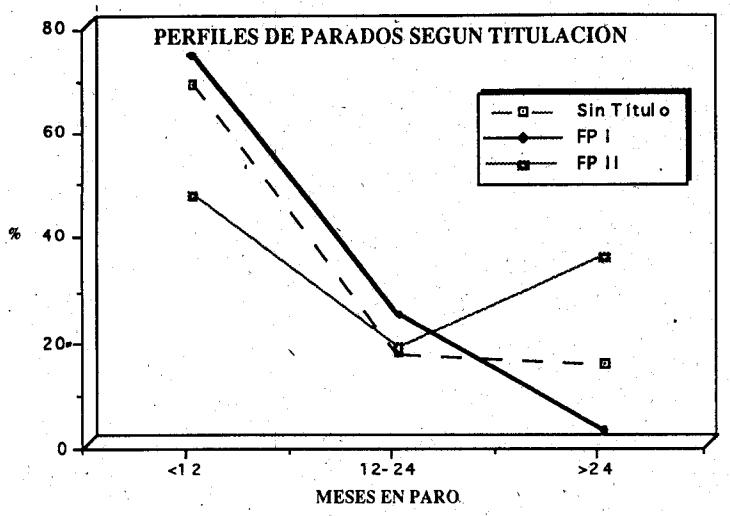


Cabe destacar el alto porcentaje de hombres con menos de 12 meses en paro (64.6%) en contraste con el de las mujeres (44.1%) y con el del grupo total (54.0%). Ocurre además que al incrementarse los meses en paro, se produce un descenso considerable en el porcentaje de hombres desempleados (alrededor del 18% en cada una de las dos categorías restantes). Por su parte, en el grupo de mujeres hay una distribución más equilibrada de los porcentajes en cada una de las categorías de la variable dependiente (44.1%, 23.1% y 32.8% respectivamente).

En cada segmento se busca de nuevo el mejor predictor, encontrando que en el grupo de los hombres todas las variables predictoras tuvieron significación excepto en la variable Experiencia Profesional. De todas las variables significativas la de mayor significación fue Titulación olasificada en tres categorías: los hombres que no poseen título, la segunda integrada básicamente por los que tienen Formación Profesional I, y la tercera, que incluye Formación Profesional II, Cursos del INEM y Diplomaturas.

El gráfico 2 es un diagrama de perfiles correspondiente a este grupo.

GRAFICO 2



HOMBRES

El primero de los tres subgrupos (Hombres sin título), tiene una distribución porcentual muy similar a la del grupo total de hombres, hecho que no resulta sorprendente en vista de que representa más del 75% de este. El subgrupo definido por FPII-INEM-Diplomados, presenta como característica diferenciadora respecto de los otros dos, un alto porcentaje de individuos con más de 2 años en paro. El tercer subgrupo (FPI) sólo cuenta con 49 individuos (menos del 2.5% del total), lo que dificulta una comparación válida con los otros dos.

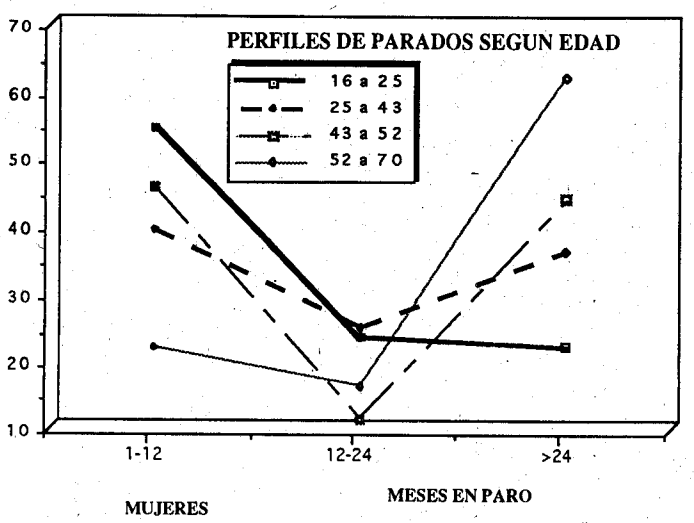
En el grupo de las mujeres, también se encontró significación con todas las variables predictoras excepto Idioma. De todas las variables significativas, la de mayor nivel de significación fue la Edad considerada en cuatro categorías: De 16 a 25 años, de 25 a 43, de 43 a 52 y de 52 a 70.

El gráfico 3 es un diagrama de perfiles correspondiente a este grupo.

El subgrupo de mayor tamaño (25 a 43 años) representa más del 50% de las mujeres estudiadas, y presenta un perfil de paro similar al del grupo total de mujeres. En las mujeres más jóvenes (16 a 25 años) se observa una mayor concentración de desempleadas en la categoría de menos de 1 año de paro, lo que produce un perfil diferenciado con respecto a los subgrupos restantes. Por otra parte, en el grupo de mujeres de mayor edad (52 a 70 años), la mayor concentración de desempleo se observa en la categoría de más de

2 años en paro. Este último subgrupo está sujeto a las mismas reservas en cuanto a tamaño, señaladas en relación con el grupo de hombres con Formación Profesional I.

GRAFICO 3



El subgrupo de mayor tamaño (25 a 43 años) representa más del 50% de las mujeres estudiadas, y presenta un perfil de paro similar al del grupo total de mujeres. En las mujeres más jóvenes (16 a 25 años) se observa una mayor concentración de desempleadas en la categoría de menos de 1 año de paro, lo que produce un perfil diferenciado con respecto a los subgrupos restantes. Por otra parte, en el grupo de mujeres de mayor edad (52 a 70 años), la mayor concentración de desempleo se observa en la categoría de más de 2 años en paro. Este último subgrupo está sujeto a las mismas reservas en cuanto a tamaño, señaladas en relación con el grupo de hombres con Formación Profesional I.

En los gráficos 4, 5 y 6 se identifica una interacción de tipo acumulado entre las variables SEXO, TITUL y MESESPAR. La Titulación discrimina con respecto a los meses en paro, en el grupo de los hombres y no en el de las mujeres.

En forma similar, los gráficos 7, 8 y 9 parecen reflejar una interacción de tipo acumulado entre las variables SEXO, EDAD y MESESPAR. La Edad discrimina con respecto a los meses en paro, en el grupo de las mujeres y no en el de los hombres.

GRAFICO 4

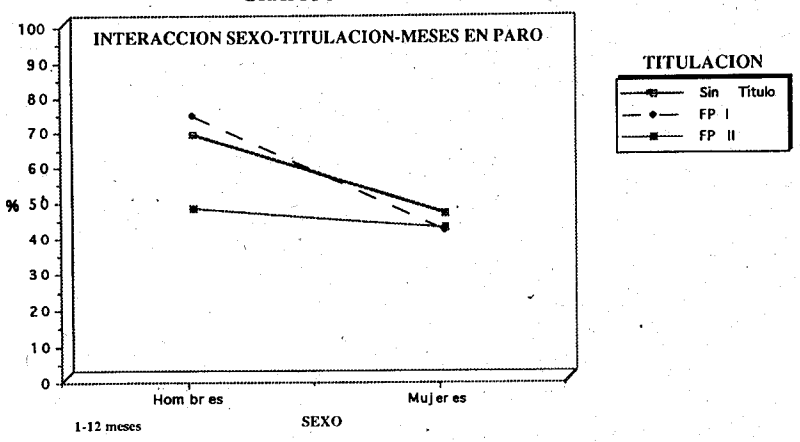


GRAFICO 5

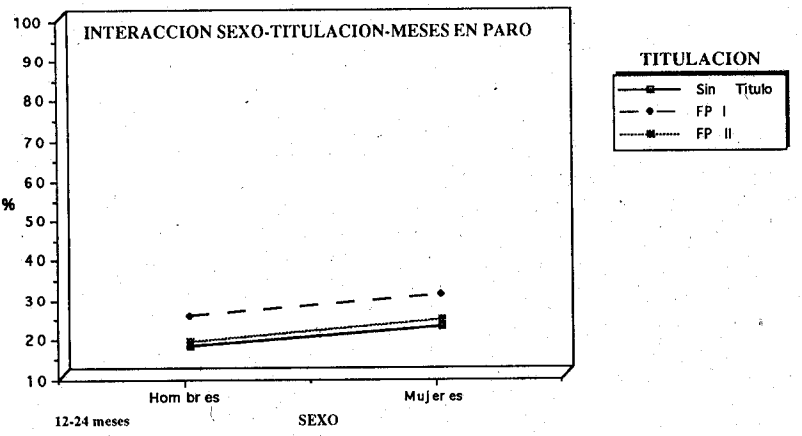


GRAFICO 6

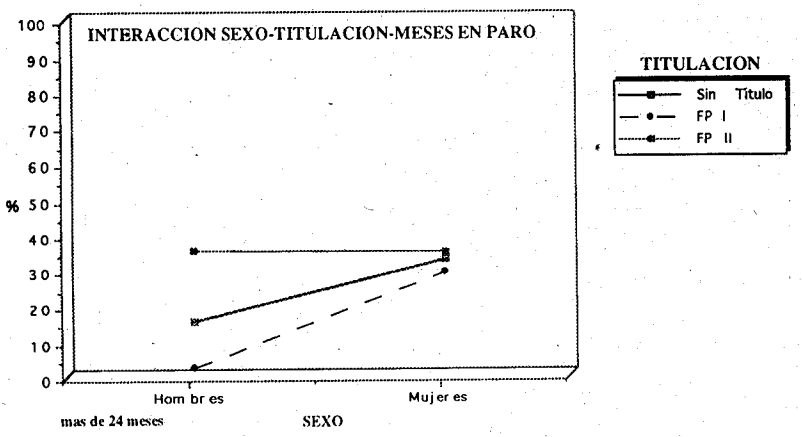


GRAFICO 7

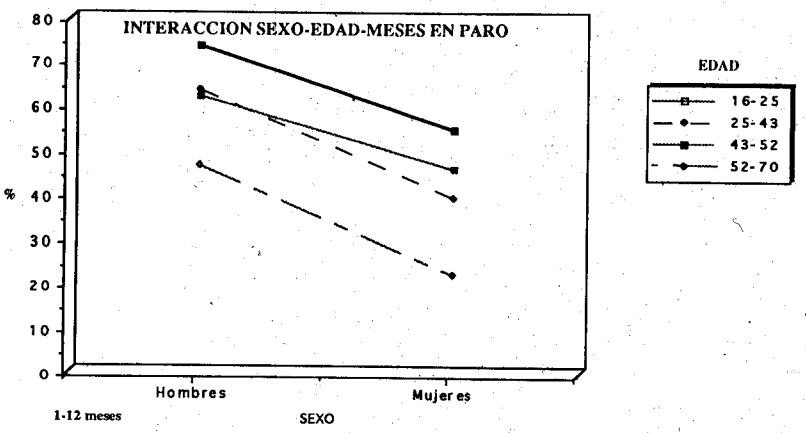


GRAFICO 8

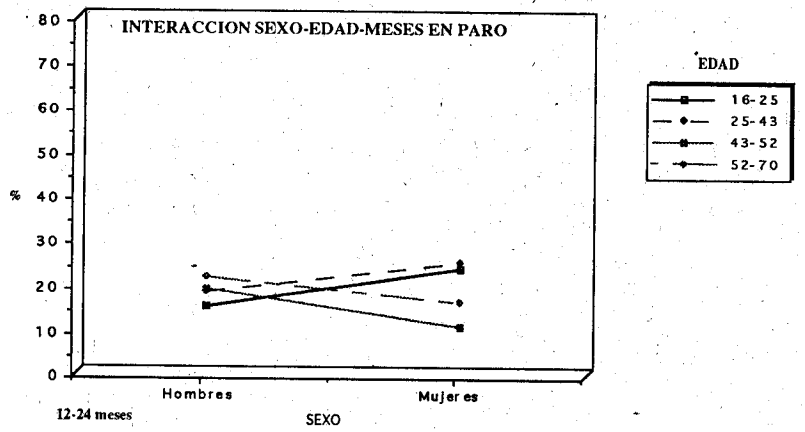
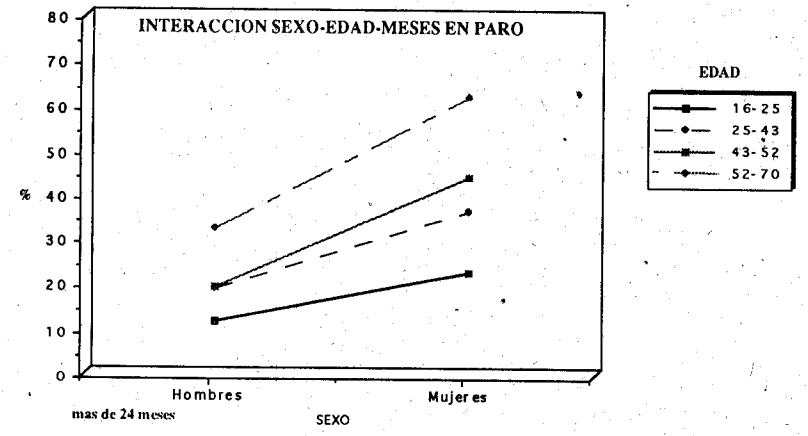
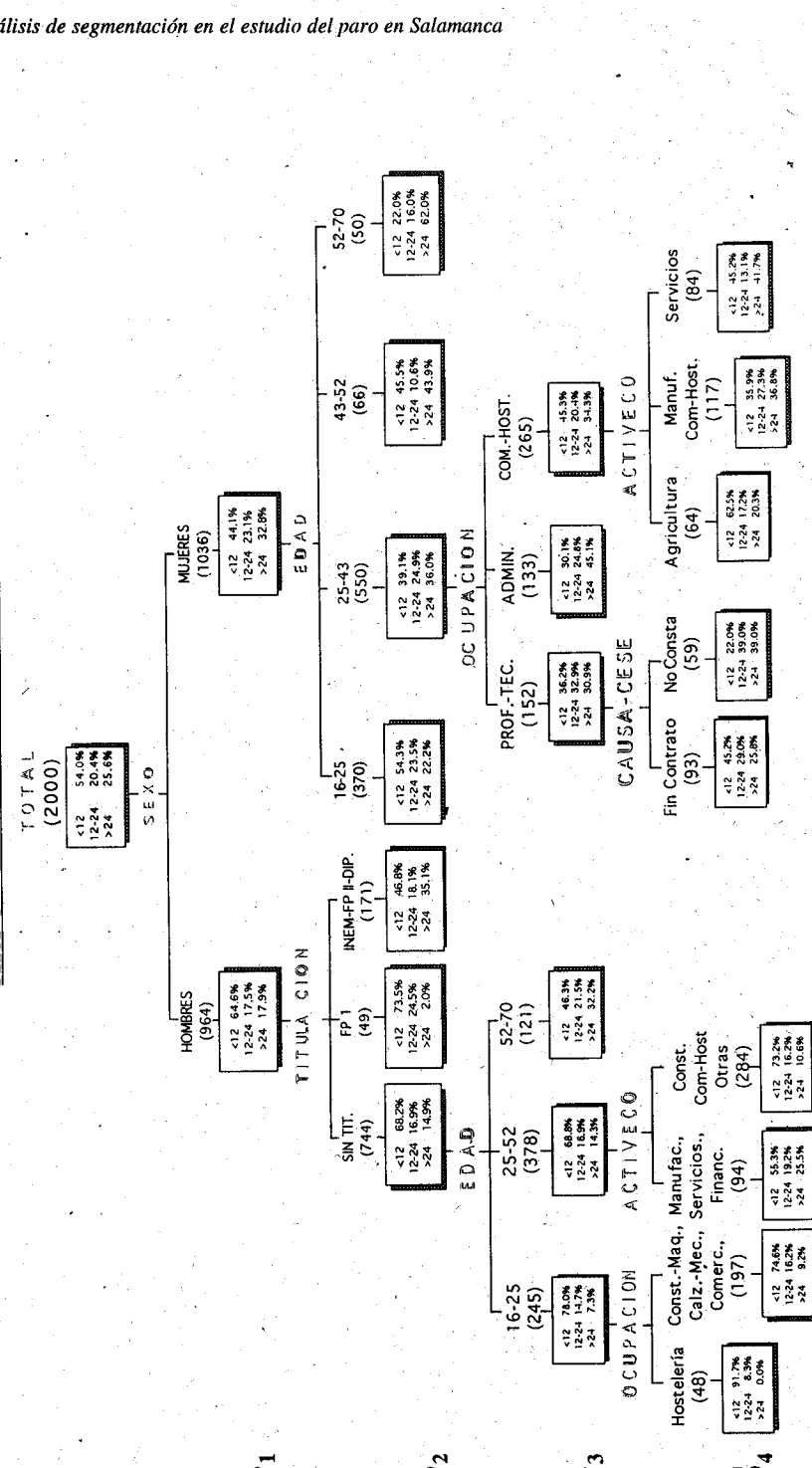


GRAFICO 9




ARBOL DE SEGMENTACION
TOTAL
(2000)
<12 54.0%
12-24 20.4%
>24 25.6%
SEXO
HOMBRES
(964)
<12 64.6%
12-24 17.5%
>24 17.9%
MUJERES
(1036)
<12 44.1%
12-24 23.1%
>24 32.8%
TITULACION
SIN TIT.
(744)
<12 68.2%
12-24 16.9%
>24 14.9%
FP I
(49)
<12 73.5%
12-24 24.5%
>24 2.0%
INEM-FP II-BUP.
(171)
<12 46.8%
12-24 18.1%
>24 35.1%
EDAD
16-25
(370)
<12 54.3%
12-24 23.5%
>24 22.2%
25-43
(550)
<12 39.1%
12-24 24.9%
>24 36.0%
43-52
(66)
<12 45.5%
12-24 10.6%
>24 43.9%
52-70
(50)
<12 22.0%
12-24 16.0%
>24 62.0%
OCUPACION
PROF.-TEC.
(152)
<12 36.2%
12-24 32.9%
>24 30.9%
ADMIN.
(133)
<12 30.1%
12-24 24.8%
>24 45.1%
COM.-HOST.
(265)
<12 45.3%
12-24 20.4%
>24 34.3%
EDAD
16-25
(245)
<12 78.0%
12-24 14.7%
>24 7.3%
25-52
(378)
<12 68.8%
12-24 16.9%
>24 14.3%
52-70
(121)
<12 46.3%
12-24 21.5%
>24 32.2%
CAUSA-CESE
Fin Contrato
(93)
<12 45.2%
12-24 29.0%
>24 25.8%
No Consta
(59)
<12 22.0%
12-24 39.0%
>24 39.0%
ACTIVECO
Agricultura
(64)
<12 62.5%
12-24 17.2%
>24 20.3%
Manuf.
Com-Host.
(117)
<12 35.9%
12-24 27.3%
>24 36.8%
Servicios
(84)
<12 45.2%
12-24 13.1%
>24 41.7%
OCUPACION
Hosteleria
(48)
<12 91.7%
12-24 8.3%
>24 0.0%
Const.-Maq.,
Calz.-Mec.,
Comerc.,
(197)
<12 74.6%
12-24 16.2%
>24 9.2%
ACTIVECO
Manufac.,
Servicios.,
Financ.
(94)
<12 56.3%
12-24 19.2%
>24 25.5%
Const.
Com-Host
Otras
(284)
<12 73.2%
12-24 16.2%
>24 10.6%
S1
S2
S3
S4

Por tanto, en el segundo nivel de segmentación S_2 (ver árbol de segmentación)) obtenemos los siete segmentos ya descritos. Tras analizar cada estrato, encontramos que el segmento Hombres- FP I y el segmento Hombres-INEM, FP II, DIP. junto a los segmentos Mujeres-(16-25), Mujeres-(43-52) y Mujeres-(52-70) son grupos terminales por razones de tamaño y por no encontrarse asociación significativa con ninguno de los predictores; sin embargo los segmentos de hombres sin Titulación y Mujeres-(25-43) siguen presentando una interacción significativa, por lo que se sometieron a un nuevo proceso de segmentación, encontrando que la variable Edad definida por tres categorías: 16-25, 25-52 y 52-70 es la variable mas significativa en los hombres sin titulación y la de Ocupación clasificada esencialmente por tres categorías: Prof.-Tec., Admin. y Com-Host. en el segmento Mujeres-(25-43).

El proceso continuó hasta una cuarta segmentación donde se consigue que todos los grupos sean terminales o bien por razón de tamaño o por no encontrar un predictor significativo.

Los resultados se sintetizan en el árbol de las páginas siguientes.

5.- DICUSION

El análisis de la muestra considerada sin llevar a cabo una segmentación de la información, nos llevaría a una estimación de la distribución de parados cuantificable en los siguientes términos: el 54 % lleva en paro menos de un año, aproximadamente el 75% menos de dos años y solo un 25% mas de dos años.

Analizando la interacción, las conclusiones son bien diferentes ya que el 82% de los hombres lleva en paro menos de dos años y sin embargo para las mujeres esta cifra decrece hasta un 67%.

Obviamente, el análisis pone de manifiesto que esta estratificación es demasiado simple y dentro de cada sexo los porcentajes varían segun otras características. Por ejemplo, en el caso de las mujeres menores de 25 años, el 45,7% lleva en paro mas de un año, para aquellas cuya edad oscila entre 25 y 43 años el porcentaje es del 60,9% llegando a ser del 78% en el grupo de las mayores. Solo en el grupo cuya edad esta comprendida entre 25 y 43 años, se ha encontrado una distribución porcentual diferente dependiendo de la Ocupación a la que se dedican, de hecho para las Administrativas el 69,9% llevan mas de un año en paro y solo el 57,7% de las que se dedican al Comercio y Hostelería. Dentro de este último grupo aún la distribución es significativamente diferente, dependiendo de su Actividad Económica, así para el sector Servicios el 58,3% lleva menos de dos años en paro y sin embargo en el sector Agricultura este porcentaje se incrementa hasta un 79,7%.

Algo similar hemos encontrado en el grupo de los varones, en el cual parece que la distribución porcentual esta mas ligada a su titulación que a su edad. La edad parece ser importante en el grupo Sin Titulación en el cual los mas jóvenes (entre 16 y 25 años) el 78% lleva en paro menos de 12 meses , sin embargo el porcentaje decrece hasta un 46,3%en el grupo que sobrepasa los 52 años.

En los mas jóvenes, la distribución esta ligada a la ocupación, así en el ramo de Hostelería solo el 8,3% lleva en el paro mas de un año y sin embargo en las otras ramas este porcentaje se incrementa hasta el 25,4%.

En el grupo de edad entre 25 y 52 años la distribución presenta una mayor relación con la Actividad Económica, así el 55,3% de parados cuya actividad esta en las Manufacturas, Servicios y Entidades Financieras contabiliza menos de un año de paro y sin embargo en el caso de la Construcción, Comercio y Hostelería y otras este porcentaje sobrepasa el 73%.

Un estudio exhaustivo de cada rama del árbol nos permitiría incrementar los detalles. Lo que resulta evidente es que un análisis global de la situación nos llevaría a resultados sesgados ya que claramente hay un interacción en la información que requiere una subdivisión en estratos homogéneos en relación a la variable dependiente para obtener estimadores fiables.

8.- BIBLIOGRAFIA

CELLARD J. y LABBE B. (1967): Le Programme ELISEE, présentation et aplication, Metra, Vol. VI, **3**, 511-520.

DIDAY E. (1992): Analyse des données et classification automatique et symbolique, Seminario Internacional de Estadística en Euskadi, Instituto Vasco de Estadística, Cuaderno 27, Eustat.

ESCOBAR M. (1992): El Análisis de Segmentación: Concepto y Aplicaciones, Estudio/Working Papers, **31**, Instituto Juan March de Estudios e Investigaciones, Madrid.

HAWKINS D. y KASS G. (1982): Automatic Interaction Detection, Topics in Applied Multivariate Analysis, Cambridge Press, 269-302.

KASS G.V. (1975): Significance testing in automatic interaction detection, Applied Statistics, **24**, 178-189.

KASS G.V. (1980): An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics, **29**, 119-127.

MORGAN J. Y SONQUIST J. (1963): Problems in theAnalysis of Survey Data and a Proposal, Journal of the American Statistical Association, **58**, 415-434.

ROM D. (1990): A Sequentially Rejective Test Procedure based on a Modified Bonferroni Inequality, Biometrika, **77**, 663-665.

VOLLE M. (1981): Analyse des Données, Collection "Ecpnomie et Statistiques Avancées", 2eme édition, Ed. Economica, Paris.