5. CLASIFICACIÓN DE LOS ACCIDENTES CON SUSTANCIAS PELIGROSAS EN FUNCIÓN DE SU GRAVEDAD

5.1. INTRODUCCIÓN

Los resultados obtenidos en el capítulo anterior ponen de manifiesto la dificultad de analizar una realidad multifactorial (dependiente de múltiples variables) mediante un procedimiento de análisis bivariante. El presente capítulo muestra los resultados obtenidos al aplicar varias técnicas de clasificación multivariante a un conjunto de registros de MHIDAS con el fin de clasificar los accidentes industriales con sustancias peligrosas en función de su gravedad. En primer lugar, este análisis debiera permitir evidenciar qué factores afectan de manera más significativa a la gravedad de los accidentes y en segundo lugar, facilitar una clasificación de los mismos en función de esos factores o incluso posibilitar una predicción de sus consecuencias.

En el apartado 5.2 se presentan algunos de los métodos estadísticos considerados para llevar a cabo esta clasificación y en el apartado 5.3 se analiza la bondad de los datos disponibles (MHIDAS) variable a variable. En los apartados siguientes (5.4 a 5.6) se presentan los resultados obtenidos en la aplicación de los distintos métodos planteados y por último, las conclusiones que ha sido posible alcanzar.

5.2. MÉTODOS DE ANÁLISIS MULTIVARIANTE

El análisis multivariante comprende un conjunto de métodos estadísticos para realizar el tratamiento conjunto de datos relativos a, literalmente, múltiples variables. Algunos de estos métodos son puramente descriptivos, limitándose a proporcionar un estudio de los datos muestrales exclusivamente; otros por el contrario, permiten realizar inferencias sobre diversos parámetros de las poblaciones analizadas [HAIR 1999] [MCCULLAG 1987].

La Tabla 5.1 recoge una clasificación de los diversos métodos de análisis multivariante (adaptada de Uriel [URIEL 1995]), en función de la naturaleza de las variables dependientes e independientes (que pueden ser continuas, categóricas o, simplemente, no existir distinción entre unas y otras, según el método considerado), así como del objetivo del método.

Así, los modelos de regresión tienen por objetivo fundamental analizar las relaciones entre la variable dependiente y la (o las) independientes. Por su parte los modelos logit o el análisis discriminante permiten obtener una clasificación de los individuos analizados, mientras que los análisis de correspondencias o de componentes principales están orientados a reducir el número de factores que describen el conjunto.

De los métodos indicados en la Tabla 5.1, se describen a continuación aquellos que han sido considerados para el desarrollo de la clasificación planteada en este capítulo. El desarrollo expuesto no pretende ser una justificación matemática completa de cada método, sino tan sólo una referencia a su fundamento y la forma en que afronta el problema.

Los modelos de regresión múltiple [CHATERJEE 1991] [HARDY 1993] aplicados a este caso fueron objeto de un completo estudio en un Proyecto Final de Carrera de la Diplomatura de Estadística de la UPC [PFC 2000]. Una de las aportaciones más significativas de este estudio fue la revisión completa de los datos contenidos en MHIDAS y la depuración de los mismos. Además de lo anterior, se consideraron como variables dependientes (descriptores de la gravedad) las variables NPM (número de muertos) y NPH (número de heridos), habiéndose desarrollado modelos de regresión para cada una de ellas. Si bien las posibilidades predictivas de estos modelos no son

buenas, permiten evidenciar las relaciones existentes entre variables y la importancia relativa de cada uno de ellas.

Tabla 5.1. Clasificación de los métodos de análisis multivariante.

Variables dependientes	Variables independientes	Método	Objetivo
Continua. 1 variable	Continuas o categóricas	Modelo de regresión lineal	Analizar Dependencia
Continuas. N variables	Continuas o categóricas	Modelo de regresión multivariante	Analizar dependencia
Categórica. 2 categorías	Continuas o categóricas	Análisis discriminante de dos grupos	Clasificación
		Modelo logit binomial	Clasificación y/o dependencia
Categórica. N categorías	Continuas o categóricas	Análisis discriminante N grupos	Clasificación
		Modelo logit multinomial	Clasificación y/o dependencia
Ninguna	Continuas	Análisis de componentes principales	Reducción de dimensiones
	Categórica	Análisis de correspondencias múltiples	Reducción de dimensiones
	Continuas o categóricas	Análisis de conglomerados	Agrupación de individuos

5.2.1. Análisis de correspondencias múltiples

El Análisis de Correspondencias Múltiples consiste en pasar de las variables categóricas originales observadas a un número más reducido de nuevas variables continuas (o factores) que sinteticen la mayor parte de la información facilitada por las variables iniciales. En este sentido es un procedimiento que tiende a reducir la dimensión del sistema, totalmente equivalente al Análisis de Componentes Principales, pero en este caso aplicado sobre variables categóricas.

Partiendo de la matriz de datos que recoge en cada fila cada uno de los individuos y en cada columna las variables observadas, se construye la matriz disyuntiva completa o tablero lógico (Z) que consta de tantas columnas como categorías consideradas. Si el individuo i posee o pertenece a la categoría j, Z_{ij} toma el valor 1, de lo contrario, el 0. El análisis de correspondencias múltiples no es más que aplicar el análisis factorial de correspondencias a la matriz Z.

De esta forma es posible posicionar cada categoría en diferentes planos factoriales pudiendo observar las relaciones de proximidad-distancia existentes entre ellas. Las agrupaciones de individuos en distintos puntos posibilitan detectar conjuntos de individuos de características similares, más o menos próximos a las categorías analizadas.

Puede consultarse abundante bibliografía al respecto en Jobson [JOBSON 1992], Uriel [URIEL 1995], Etxeberría [ETXEBERRIA 1995], Aluja [ALUJA 1996], Escofier [1990], Lebart [LEBART 1985] [LEBART 1994], Joaristi [JOARISTI 1999], Anderson [ANDERSON 1984] y Becue [BECUE 1991].

Para aplicar el ACM a los datos utilizados en esta tesis se ha empleado el software SPAD.N, especialmente adaptado a este tipo de análisis.

5.2.2. Análisis de conglomerados

Con el Análisis de Conglomerados se pretende dividir la población en diferentes clases o subconjuntos lo más homogéneos posible de una forma automática y mediante algoritmos formalizados matemáticamente. Las clases obtenidas deberán contener individuos parecidos entre sí y lo más diferenciados posible de los individuos de otras clases.

Para llevar a cabo este proceso debe definirse una distancia entre individuos que nos indique en qué medida pueden ser considerados miembros de una misma clase y debe definirse también una estrategia que permita la agregación o separación de individuos para construir las clases.

Se puede establecer la siguiente clasificación de algoritmos:

1. Métodos de partición o no jerárquicos.

    1.1. Método de agregación de centros.
    1.2. Método de centros móviles.
    1.3. Método de las nubes dinámicas.

2. Métodos jerárquicos.

    2.1. Métodos jerárquicos ascendentes.
    2.2. Métodos jerárquicos descendentes.

Los métodos no jerárquicos conducen a la obtención de una partición habiendo definido a priori el número de clases finales a considerar. Se basan en considerar un centro de clase e ir agrupando los individuos a cada uno de ellos en función de su distancia a los mismos.

Los métodos jerárquicos producen una serie de clases anidadas unas dentro de las otras. Los métodos jerárquicos descendentes dividen el conjunto inicial en dos grupos los más homogéneos posible. Sucesivamente se repite esta operación con cada una de las clases obtenidas en la etapa anterior. El proceso se detiene a criterio del analista.

En los procedimientos jerárquicos ascendentes se procede de manera inversa. Se agrupan los individuos más próximos entre sí dos a dos y se reemplazan por los centros de su clase. Se repite el proceso iterativamente hasta que en una clase se agrupan dos colectivos que no pueden ser considerados iguales.

En la presente tesis se ha llevado a cabo este análisis después de aplicar sobre el conjunto de datos estudiado un Análisis de Correspondencias Múltiples, permitiendo así obtener varios conjuntos de individuos más o menos homogéneos. Se ha utilizado para ello el programa estadístico SPAD.N

Las referencias bibliográficas sobre esta técnica son muy diversas y extensas. Las más utilizadas en esta tesis han sido Jobson [JOBSON 1992], Aluja [ALUJA 1996], Bisquerra [BISQUERRA 1989] y Etxeberría [ETXEBERRIA 1995].

5.2.3. Modelos de regresión logit

Como generalización de los modelos de regresión, cuando la variable respuesta es dicotómica o politómica es posible emplear una transformación de la misma con el objeto de modelizarla mediante regresión múltiple. Las funciones de transformación más habituales son la lineal, la logit y la probit.

La transformación logit consiste en modelizar la respuesta ln(p/(1-p)) mediante las variables explicativas disponibles (categóricas o no), donde p es la probabilidad de que el individuo pertenezca a uno de los dos grupos considerados y, lógicamente (1-p) la complementaria. Para el caso de múltiples categorías en la variable respuesta (logit multinomial) se toma una categoría de referencia y se calculan tantas funciones como categorías hay, cumpliéndose en todo caso que $p_1 + p_2 + \ldots + p_n = 1$.

Para el cálculo de los coeficientes se emplea la estimación por máxima verosimilitud.

Con la aplicación de estos métodos se obtienen, al igual que en el caso del análisis discriminante, una o más funciones que permiten la clasificación predictiva de los individuos en uno de los grupos o categorías considerados por la variable respuesta, en función de las variables independientes o explicativas. A diferencia del análisis de conglomerados, en el que no se conoce ni el número ni la descripción de las categorías finales, mediante los modelos logit o el análisis discriminante es posible dividir la población en n categorías definidas a priori.

Para evaluar la bondad de los resultados obtenidos, a parte de varios estadísticos de referencia, se emplean las matrices de confusión. En ellas es posible identificar qué proporción de individuos resulta clasificada correctamente.

Para el cálculo de los distintos modelos elaborados en esta tesis se ha empleado el programa MINITAB, por su disponibilidad, facilidad de uso y completa información sobre los modelos obtenidos.

Las referencias bibliográficas sobre estos modelos son muy abundantes, siendo relevantes las de Jobson [JOBSON 1992], Agresti [AGRESTI 1990], Draper [DRAPER 1981], y el propio manual del programa estadístico MINITAB [MINITAB 1996]. Se describen unas aplicaciones interesantes en Bovio [BOVIO 1997], Carrega [CARREGA 1997] y Vega [VEGA 1995].

5.2.4. Árboles de clasificación

Las técnicas de clasificación mediante árboles [SANTESMASES 1997] son particularmente útiles para el análisis de conjuntos de datos multivariantes. La técnica AID (Automatic Interaction Detector) desarrollada por Sonquist [SONQUIST 1973] está en el origen de estos métodos. Sin embargo ha sido criticada por la falta de un test de significación, por la tendencia a identificar relaciones espúreas y sus inadecuadas reglas de finalización [BIGGS 1991]. Estas críticas han ido acompañadas, por lo general, por desarrollos de esta técnica básica que han permitido la aparición de otros métodos más elaborados tales como el CHAID, desarrollado por Kass [KASS 1980].

En la presente tesis se ha hecho uso del método KS implementado en el software estadístico del mismo nombre, acrónimo de Knowledge Seeker desarrollado por Biggs en 1991 y que supone un refinamiento sobre el programa CHAID.

El algoritmo KS, divide recursivamente cada nodo o subconjunto de datos en K nuevos nodos, empezando por el nodo inicial que agrupa la totalidad de los datos. En cada nodo, todas las variables predictoras son consideradas para ser utilizadas como variables discriminantes, utilizándose en cada caso, por defecto, aquella que logra la mejor partición, aunque el analista puede escoger cualquier otra basándose en la propia experiencia o el sentido físico de la clasificación.

Para la selección de la variable que consigue la clasificación más significativa se sigue el siguiente algoritmo:

1. Se selecciona el par de categorías de la variable predictora que sean más similares en función de los test de significación F o χ^2. Se repite este proceso iterativamente hasta que sólo queden dos grupos de categorías.

2. Se calcula el test de significación para cada uno de los subconjuntos anteriores usando los estadísticos F o χ^2 y se selecciona como mejor partición el grupo con la menor probabilidad p.

3. Se determina si la participación seleccionada es estadísticamente significativa. Si no lo es, no puede dividirse más ese nodo en función de esa variable.

Puede hallarse una descripción exhaustiva de este procedimiento en Biggs [BIGGS 1991] y otras referencias a estas metodologías en Kass [KASS 1975] [KASS 1980].

Mediante la aplicación de estas metodologías a los datos manejados en esta tesis se pretende alcanzar una clasificación de accidentes que permita, en la medida de lo posible, estimar la gravedad de sus consecuencias según la tipología de que se trate.

5.2.5. Análisis discriminante

La técnica del análisis discriminante es muy similar al análisis de correlación canónica. Se puede caracterizar este análisis como un análisis de correlación canónica donde uno de los grupos de variables está formado por variables ficticias codificadas que representan la pertenencia de los casos a los distintos grupos.

Matemáticamente el análisis discriminante efectúa una descomposición en valores y vectores propios de la matriz $SCR^{-1}SCM$, donde SCR y SCM son las matrices de sumas cuadráticas del modelo y sumas cuadráticas de error de un MANOVA.

Desde un punto de vista formal, las funciones discriminantes son ecuaciones de regresión que permiten determinar en cada caso la puntuación canónica que les corresponde. Sin embargo el proceso de discriminación es más evidente utilizando las funciones de clasificación de Fisher, que permiten obtener una puntuación directa de cada grupo, atribuyéndose la pertenencia de un caso al grupo que presente mayor puntuación. A los efectos de obtener una clasificación predictiva, la interpretación de este método es verdaderamente simple.

Puede hallarse abundante bibliografía sobre la materia en Jobson [JOBSON 1992], Etxeberría [ETXEBERRIA 1995], Ato [ATO 1994] [ATO 1996] y Uriel [URIEL 1995].

Los resultados obtenidos en la aplicación de este método a los datos disponibles han sido realmente muy poco relevantes, por lo que se obvia su presentación en esta memoria.

5.3. ANÁLISIS DE LAS VARIABLES DISPONIBLES

Partiendo de la selección de registros de MHIDAS (versión enero 2000) y una vez llevadas a cabo las modificaciones descritas en el capítulo precedente, quedan disponibles 5.167 registros de accidente.

Como ya se ha visto anteriormente, de todas las variables disponibles, tan sólo cuatro describen la gravedad de las consecuencias del evento (número de muertos, número de heridos, número de evacuados e importe económico de los daños). De ellas, la variable "número de muertos" es la más significativa permitiendo llevar a cabo un análisis multivariante con ciertas probabilidades de éxito.

El campo "NPM" (número de personas muertas) sólo está informado en 2.216 casos de los 5.167 anteriormente citados. Los principales parámetros estadísticos de posición central o de dispersión de esta variable se presentan en la Tabla 5.2

Tabla 5.2. Parámetros de dispersión y de posición central de la variable NPM.

Parámetro	Valor
Máximo	2.000
Mínimo	0
Mediana	0
Media	6,36
Moda	0
Desviación estándar	59,94
Varianza	3.592

Los parámetros anteriores ponen de manifiesto que la mayoría de los accidentes recopilados, concretamente 1.111 no han causado víctimas mortales. De hecho, la distribución de valores que presenta esta variable se muestra en la Tabla 5.3 siguiente.

Tabla 5.3. Dispersión de los valores de la variable NPM.

Rango NPM	Nº casos	%
0	1.111	50,13
1	360	16,24
2-3	283	12,77
4-6	192	8,66
7-10	97	4,38
11-20	83	3,75
21-50	60	2,71
51-100	17	0,77
101-500	9	0,41
501-2000	4	0,18
TOTAL	**2.216**	**100,00**

Como puede apreciarse, el 93,18% de los casos presenta entre 0 y 10 muertos, siendo muy poco frecuentes los accidentes con un número de muertos superior.

Los campos de la base de datos MHIDAS adaptada que pueden contener información relevante para llevar a cabo un análisis multivariante sobre este conjunto de registros son los presentados en la Tabla 5.4.

Tabla 5.4. Campos de MHIDAS disponibles para un análisis multivariante.

VARIABLE ORIGINAL	DESCRIPCIÓN	VARIABLES INDICADORAS	DESCRIPCIÓN
AN	Identificación		
AÑO	Año del accidente		
PD	Densidad de población		
MT	Estado Físico	EF-SOLID	Estado físico sólido
		EF-LIQUID	Estado físico líquido
		EF-GAS	Estado físico gas
		EF-PLGAS	Estado físico gas licuado a presión
		EF-DUST	Estado físico polvo
MH	Peligrosidad del material	HZTO	Sustancia tóxica
		HZFI	Sustancia inflamable
		HZEX	Sustancia explosiva
		HZCO	Sustancia corrosiva
		HZRA	Sustancia radioactiva
		HZCD	Sustancia refrigerada
		HZAS	Sustancia asfixiante
		HZOX	Sustancia oxidante
NSUS	Nº de sustancias		
IT	Tipo de incidente	IT-EXPLODE	Explosión
		IT-FIRE	Incendio
		IT-GASCLD	Formación de nube de gas
		IT-RELEASE	Fuga o derrame
OG	Actividad de origen	OG-DOM/COM	Envase o actividad comercial
		OG-PROCESS	Proceso industrial
		OG-STORAGE	Almacenamiento
		OG-TRANSFER	Transferencia de sustancias
		OG-OTHERS	Otras actividades
GC	Causa general	GC-EXTERNAL	Externa
		GC-HUMAN	Error humano
		GC-IMPACT	Impactos
		GC-MECHANICAL	Fallo mecánico
		GC-VREACTION	Reacción violenta
		GC-INSTRUMENT	Fallo de instrumentación
		GC-PROCOND	Variación de las condiciones de proceso
		GC-SERVICE	Fallo en los servicios auxiliares
MUNDO	Nivel de desarrollo	M-1	País desarrollado
		M-2	País en vías de desarrollo
		M-3	País subdesarrollado
NSUS	Nº de sustancias	Continua	
QYN	Cantidad de sustancia	Continua	
NPM	Nº de personas muertas	Continua	

De cada uno de ellos pueden hacerse las siguientes consideraciones:

Variable AÑO: Como se ha visto en el Capítulo 4, concentra la mayor parte de los registros en el último tercio del siglo XX. Considerando sólo los 2.216 registros con la

variable NPM documentada, la distribución es la presentada en la Tabla 5.5. El 84% de los registros (1.867) corresponden al periodo 1970-1998.

Tabla 5.5. Distribución de los Accidentes por año de ocurrencia

Periodo	Nº
<1950	84
1950-1959	67
1960-1969	198
1970-1979	532
1980-1989	858
1990-1999	477
TOTAL	2.216

Variable PD: De las tres categorías que presenta, 595 registros corresponden a "Town" (densidad de población alta) y 1.432 son datos mancantes, por lo que esta variable no se considera en el análisis.

Variable MT: Tiene 248 registros con datos mancantes. Los restantes se distribuyen de la siguiente manera (Tabla 5.6.)

Tabla 5.6. Estado físico de las sustancias.

Estado	Nº
Sólido	227
Líquido	869
Gas	394
Plgas	332
Polvo	65
Varios estados	81
TOTAL	1.968

Variable MH: No presenta datos mancantes. Su distribución se muestra en la Tabla 5.7. Debe tenerse en cuenta que las categorías no son excluyentes.

Tabla 5.7. Categorías de peligro.

Categoría	Nº
Tóxico	645
Inflamable	1.463
Explosivo	313
Corrosivo	158
Radiactivo	4
Refrigerado	15
Asfixiante	9
Oxidante	101

Dado el reducido número de registros que presentan algunas de las categorías de peligro indicadas, sólo se consideran en el análisis las tres primeras (tóxico, inflamable y explosivo).

Variable IT: Describe el tipo de incidente con cuatro categorías no excluyentes, distribuidas como se indica en la Tabla 5.8.

Tabla 5.8. Categorías de peligro.

Categoría	Nº
IT-EXPLODE	1.070
IT-FIRE	1.063
IT-GASCLD	309
IT-RELEASE	753

Variable OG: Presenta 138 registros no informados. Sus seis categorías se distribuyen de la siguiente manera (Tabla 5.9).

Tabla 5.9. Origen del accidente.

Origen	Nº
OG-DOM/COM	177
OG-PROCESS	867
OG-STORAGE	593
OG-TRANSFER	302
OG-WAREHOUSE	104
OG-WASTE	35

Dada la escasa representatividad de las dos últimas categorías, éstas se agrupan en una sola (OG-OTHERS)

Variable GC: Tiene 924 registros no informados. Su distribución es la indicada en la Tabla 5.10.

Tabla 5.10. Causa del accidente.

Causa	Nº
GC-HUMAN	494
GC-VREACT	175
GC-MECHANICAL	598
GC-EXTERNAL	257
GC-PROCOND	63
GC-INSTRUMENTS	40
GC-IMPACT	75
GC-SERVICE	30

Por su interés, y a priori, se incluye en el análisis, si bien en algunos casos termina por eliminarse, dado que la elevada presencia de registros no informados disminuye notablemente la muestra objeto de estudio.

Variable MUNDO: No tiene datos mancantes. 1.733 registros corresponden a países desarrollados, 169 a países en vías de desarrollo y 314 a países del tercer mundo.

Variable NSUS: De los 2.216 registros disponibles, 1.977 se refieren a una única sustancia y 166 a dos sustancias. No se considera esta variable en el análisis.

Variable QYN: Desafortunadamente en 1.614 registros presenta valor nulo, por lo que no se considera su uso. Esta circunstancia limita notablemente las posibilidades del estudio ya que se entiende que esta variable es realmente significativa para poder evaluar la gravedad de los accidentes con sustancias peligrosas.

5.4. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES Y ANÁLISIS DE CONGLOMERADOS

5.4.1. Objetivo

Se pretende llevar a cabo una factorización de los datos recogidos en MHIDAS con el fin de obtener, posteriormente, una partición de los mismos en n clases homogéneas de datos. Para ello se utiliza un procedimiento de clasificación ascendente jerárquico utilizando las primeras coordenadas factoriales y haciendo uso del criterio de agregación de Ward [ALUJA 1996].

Para ejecutar este análisis en el programa estadístico SPAD.N, se hace uso, entre otros de utilidad general para la preparación y descripción de los datos, de los siguientes procedimientos:

SELEC: Selecciona los individuos y las variables que se considerarán activos en el análisis.

CORMU: Ejecuta un análisis de correspondencias múltiples completo, facilitando las coordenadas factoriales de las distintas variables.

DEFAC: Permite una descripción de los ejes factoriales obtenidos mediante el análisis de las modalidades que quedan opuestas por cada eje.

GRAPH: Realiza la representación gráfica de las modalidades en los planos factoriales.

RECIP: Procede a la clasificación ascendente jerárquica de los individuos mediante sus primeras coordenadas factoriales.

PARTI: Realiza la partición en las n clases consideradas y una agregación posterior alrededor de centros móviles.

DECLA: Describe las clases obtenidas a través de las categorías más representativas en cada una de ellas.

5.4.2. Preparación de los datos

Se seleccionan, como ya se ha indicado en el apartado precedente, todos aquellos accidentes con la variable "número de personas muertas" informada y que corresponden a accidentes en instalaciones fijas (no "Transport-" en el campo "Origen"). Se extraen de la selección todos aquellos registros que a su vez presentan datos mancantes en las variables "Estado físico", "Tipo de incidente" u "Origen". Tras esta depuración quedan disponibles para el análisis 1.836 registros.

Se ha valorado la posibilidad de extraer de la muestra aquellos accidentes que por su gravedad excepcional, pudieran distorsionar los resultados. Sin embargo, finalmente se ha optado por incluirlos en el análisis con el fin de tenerlos presentes, aun siendo conscientes del peso que pueden representar en el conjunto.

Tras múltiples ensayos, las variables que han resultado representativas en este análisis y la forma en la que definitivamente han sido codificadas es la que se presenta en la Tabla 5.11.

Tabla 5.11. Variables y categorías utilizadas en el análisis de correspondencias múltiples.

VARIABLE ORIGINAL	DESCRIPCIÓN	CATEGORÍAS	OBSERVACIONES
AN	Identificación		
DECADA	5 categorías	<1960	No interviene en el análisis final presentado.
		<1970	
		<1980	
		<1990	
		<2000	
EF-SOLID	Estado Físico SÓLIDO	SOL	Variable dicotómica (Si/No)
		NSO	
EF-LIQUID	Estado Físico LÍQUIDO	LIQ	Variable dicotómica (Si/No)
		NLI	
EF-PLGAS	Estado Físico PLGAS	PLG	Variable dicotómica (Si/No)
		NPL	
EF-GAS	Estado Físico GAS	GAS	Variable dicotómica (Si/No)
		NGA	
EF-DUST	Estado Físico DUST	DUS	Variable dicotómica (Si/No)
		NDU	
MH-TO	Toxicidad	STO	Variable dicotómica (Si/No)
		NTO	
MH-FI	Inflamabilidad	SFI	Variable dicotómica (Si/No)
		NFI	
MH-EX	Explosividad	SEX	Variable dicotómica (Si/No)
		NEX	
IT-EXPLODE	Explosión	SXP	Variable dicotómica (Si/No)
		NXP	
IT-FIRE	Incendio	SFR	Variable dicotómica (Si/No)
		NFR	
IT-GASCLD	Nube de gas	SGC	Variable dicotómica (Si/No)
		NGC	
IT-RELEASE	Derrame o fuga	SRE	Variable dicotómica (Si/No)
		NRE	
OG	Actividad de origen	OG-DOM/COM	Envase o actividad comercial
		OG-PROCESS	Proceso industrial
		OG-STORAGE	Almacenamiento
		OG-TRANSFER	Transferencia de sustancias
		OG-OTHERS	Otras actividades
MUNDO	Nivel de desarrollo	MU1	País desarrollado
		MU2	País en vías de desarrollo
		MU3	País subdesarrollado
NPM	Nº de personas muertas	NM1(=0)	
		NM2(<3)	
		NM3(<15)	
		NM4(≥15)	

Las variables que no se han incluido en la tabla anterior han sido eliminadas del análisis por no aportar ningún resultado significativo. Es el caso, por ejemplo, de la variable GC (Causa general). Esta variable presenta aproximadamente un 50% de registros vacíos, por los que su inclusión en la muestra reduce notablemente el número de registros disponibles.

5.4.3. Resultados

Tras el análisis de correspondencias múltiples los resultados proporcionados por SPAD son los que, de forma resumida, se presentan a continuación.

El histograma de los primeros valores propios permite determinar el nivel de información retenido al conservar un número reducido de factores. Como puede apreciarse, conservando los primeros 5 factores sólo se retiene el 47,44% de la información proporcionada por todas las variables originales.

Tabla 5.12. Primeros valores propios

Factor	Valor Propio	%	% acumulado	
1	0,2191	15,65	15,65	**********
2	0,1696	12,11	27,77	**********
3	0,1006	7,19	34,95	********
4	0,0947	6,76	41,72	*******
5	0,0801	5,72	47,44	******
...	...	...	...	
TOTAL (Traza)	1,4000		100,00	

Por lo que al cuadro de coordenadas y valores test se refiere, puede apreciarse que todas las modalidades consideradas son representativas (valores test superiores a 2 en valor absoluto) en uno u otro eje (Tabla 5.13)

El valor test es el criterio que evalúa estadísticamente la desviación entre la media sobre el grupo y la media sobre la población. Se expresa en número de desviaciones tipo de una ley normal. Así, cuanto mayor sea el valor test observado (y superior al umbral de 2 desviaciones tipo) mejor caracterizará el elemento a la categoría de individuos. [LEBART 1994]

Básicamente, el eje 1 opone los accidentes con sustancias sólidas y explosivas, con un elevado número de muertos a los accidentes producidos por fugas con posterior formación de nubes de gas, con sustancias tóxicas.

El eje 2 contrapone los accidentes con sustancias sólidas y explosivas con elevado número de muertos a los accidentes por incendio producidos por sustancias líquidas inflamables.

El tercer eje a su vez opone varias categorías de la variable OG. Así los accidentes en instalaciones de proceso o por actividades comerciales se oponen a los accidentes en zonas de almacenamiento o los producidos en operaciones de mantenimiento. El mismo eje opone las sustancias en estado gaseoso a las que se encuentran en estado sólido.

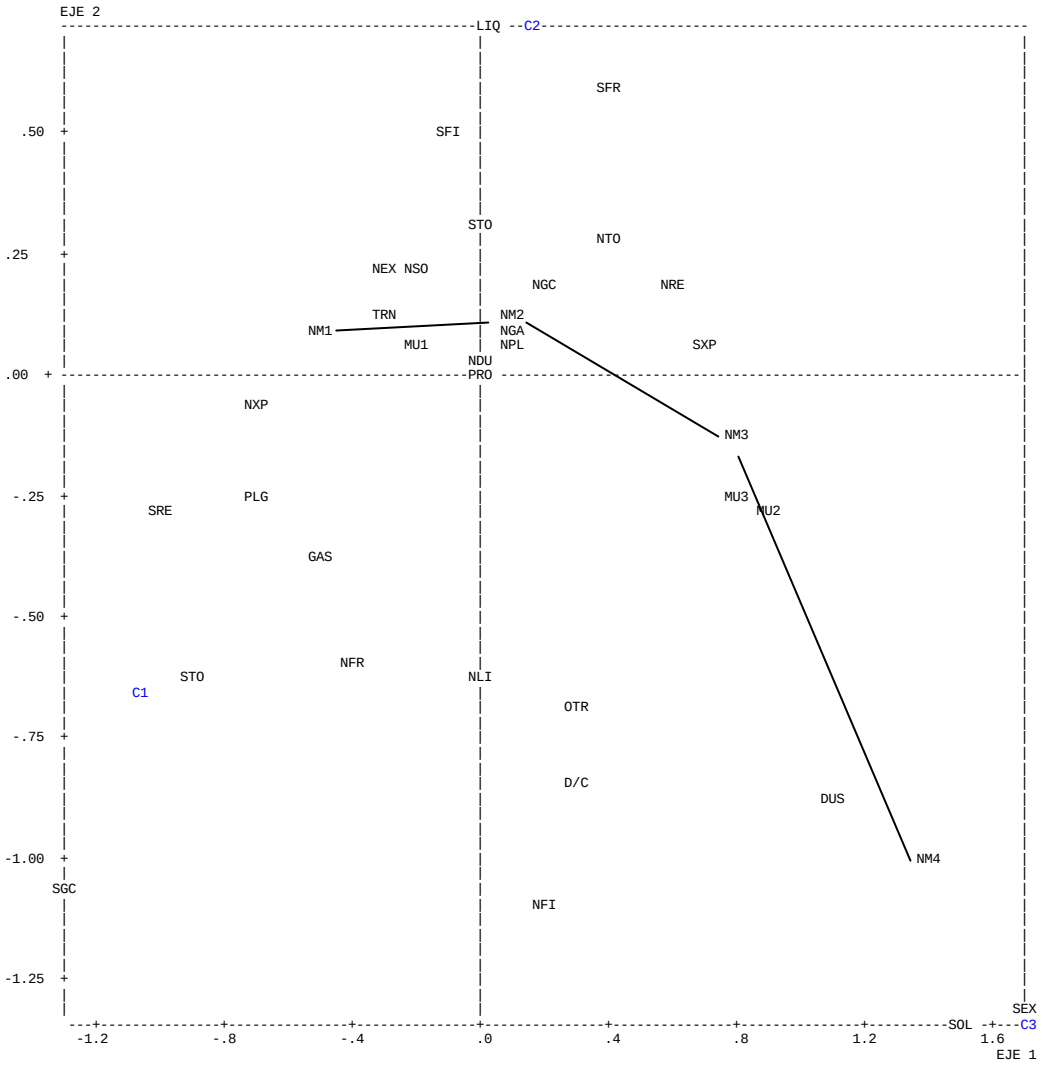
La Figura 5.1 presenta el primer plano factorial (ejes 1 y 2) y la ubicación de cada categoría sobre el mismo. En él es posible observar a priori que en el extremo del cuarto cuadrante se ubican los accidentes más graves, producidos por sustancias sólidas (y en polvo) y explosivas y que los accidentes ocurridos en países desarrollados (MU1) se ubican próximos a los accidentes menos graves (NM1) frente a los de las categorías MU2 y MU3 (muy próximos) y que se concentran próximos a las categorías NM3 y NM4 (elevado número de muertos).

Tabla 5.13. Valores test y coordenadas de cada categoría en los cinco primeros ejes factoriales.

	VALORES TEST					COORDENADAS				
	1	2	3	4	5	1	2	3	4	5
SOL - SI SOLIDO	24.4	-22.4	-9.6	5.2	-13.6	1.45	-1.33	-.57	.31	-.81
NSO - NO SOLIDO	-24.4	22.4	9.6	-5.2	13.6	-.22	.21	.09	-.05	.13
LIQ - SI LIQUIDO	-.5	28.4	-13.1	16.2	8.4	-.01	.71	-.32	.40	.21
NLI - NO LIQUIDO	.5	-28.4	13.1	-16.2	-8.4	.01	-.62	.29	-.36	-.18
PLG - SI PLGAS	-13.5	-4.9	-9.7	-33.1	-3.8	-.66	-.24	-.48	-1.63	-.19
NPL - NO PLGAS	13.5	4.9	9.7	33.1	3.8	.15	.05	.11	.37	.04
GAS - SI GAS	-11.0	-8.5	33.2	7.9	-9.0	-.49	-.38	1.48	.35	-.40
NGA - NO GAS	11.0	8.5	-33.2	-7.9	9.0	.13	.10	-.41	-.10	.11
DUS - SI DUST	9.2	-7.5	-.6	2.3	27.2	1.08	-.87	-.08	.27	3.19
NDU - NO DUST	-9.2	7.5	.7	-2.3	-27.2	-.04	.03	.00	-.01	-.13
STO - TOXICO	-23.6	-17.9	1.3	6.0	-3.3	-.84	-.64	.05	.22	-.12
NTO - NO TOXICO	23.6	17.9	-1.3	-6.0	3.3	.36	.27	-.02	-.09	.05
SFI - INFLAMABLE	-4.6	31.1	8.7	-10.3	-10.2	-.07	.49	.14	-.16	-.16
NFI - NO INFLAMABLE	4.6	-31.1	-8.7	10.3	10.2	.16	-1.07	-.30	.35	.35
SEX - EXPLOSIVO	29.0	-21.9	-6.7	2.0	6.9	1.71	-1.29	-.40	.12	.41
NEX - NO EXPLOSIVO	-29.0	21.9	6.7	-2.0	-6.9	-.27	.20	.06	-.02	-.06
SXP - SI-IT1EXPL	30.2	2.4	9.5	-7.7	4.7	.73	.06	.23	-.19	.11
NXP - NO-IT1EXPL	-30.2	-2.4	-9.5	7.7	-4.7	-.68	-.05	-.22	.18	-.11
SFR - SI-IT1FIRE	15.2	24.6	-2.6	2.0	-13.7	.35	.57	-.06	.05	-.32
NFR - NO-IT1FIRE	-15.2	-24.6	2.6	-2.0	13.7	-.36	-.58	.06	-.05	.32
SGC - SI-ITGCD	-23.0	-19.5	-3.3	.0	-2.6	1.24	-1.05	-.18	.00	-.14
NGC - NO-IT1GCD	23.0	19.5	3.3	.0	2.6	.23	.20	.03	.00	.03
SRE - SI-IT1RELE	-31.0	-9.5	-1.8	-.7	6.5	-.94	-.29	-.06	-.02	.20
NRE - NO-IT1RELE	31.0	9.5	1.8	.7	-6.5	.55	.17	.03	.01	-.12
D/C - DOM/COM	3.2	-10.2	6.9	-9.2	-4.7	.26	-.83	.56	-.75	-.39
PRO - PROCESS	1.5	.3	23.3	13.2	5.8	.04	.01	.66	.38	.17
STO - STORAGE	-.8	8.3	-13.9	-2.1	9.9	-.03	.30	-.50	-.08	.36
TRN - TRANSFER	-5.4	2.2	-8.7	-19.1	-2.4	-.29	.12	-.47	-1.03	-.13
OTR - OTROSOG1	3.1	-8.0	-15.1	15.6	-20.8	.27	-.70	-1.31	1.35	-1.81
MU1 - MUNDO1	-17.7	5.5	-3.3	5.9	9.7	-.21	.06	-.04	.07	.11
MU2 - MUNDO2	10.6	-3.1	4.3	-8.3	-2.4	.93	-.27	.38	-.73	-.21
MU3 - MUNDO3	13.1	-4.2	.8	-.9	-9.7	.77	-.25	.05	-.05	-.57
NM1 - NM=0	-21.9	3.8	-14.6	12.4	-5.2	-.50	.09	-.33	.28	-.12
NM2 - NM<3	2.9	3.1	11.1	-6.2	7.8	.12	.13	.46	-.26	.32
NM3 - NM<15	16.9	-2.5	8.6	-7.1	-2.5	.80	-.12	.41	-.33	-.12
NM4 - NM>=15	13.5	-10.0	-3.9	-3.3	1.2	1.34	-1.00	-.39	-.33	.12

También es posible observar la oposición de categorías descrita para cada uno de los dos primeros ejes. Debe observarse cierta prudencia al analizar estos gráficos pues aunque en los mismos dos categorías aparezcan muy próximas por razón de la perspectiva, en realidad pueden estar absolutamente alejadas en función de alguno de los ejes restantes.

Fig. 5.1. Representación gráfica del primer plano factorial.



Tras la ejecución del procedimiento de SPAD.N denominado PARTI y aplicando una técnica de consolidación alrededor de tres centros, ha sido posible obtener tres clases que agrupan 575, 999 y 262 accidentes respectivamente. La inercia total observada es de 0,664, la inercia interclases es de 0,293 y el cociente resulta 0,442, valor relativamente reducido como para obtener unas clases claramente definidas.

Se han ensayado otras clasificaciones con más categorías pero dadas las limitaciones de los datos y las pocas variables útiles, los resultados obtenidos correspondían más a distribuciones aleatorias que a categorías con algún sentido físico observable.

La Tabla 5.14 recoge los valores test de cada una de las tres clases antes citadas y sus coordenadas en los cinco primeros ejes. Los valores test permiten afirmar que las clases obtenidas son claramente significativas en los primeros ejes.

Tabla 5.14. Valores test y coordenadas factoriales de las clases obtenidas.

		VALORES TEST					COORDENADAS				
	Efectivos	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
CLASE 1	575	-30,3	-20,7	1,4	3,3	0,3	-1,05	-0,71	0,05	0,12	0,01
CLASE 2	999	7,5	35,2	3,0	-4,4	-4,9	0,16	0,75	0,06	-0,1	-0,11
CLASE 3	262	29,5	-22,8	-6,1	1,9	6,6	1,69	-1,3	-0,35	0,11	0,38

Estas clases han sido representadas en la Figura 5.1 indicadas como C1, C2 y C3 respectivamente.

Las clases consideradas pueden definirse en función de las categorías más próximas a ellas, tal como se representan en el plano factorial definido por los dos primeros ejes.

Así la clase C1 agrupa accidentes por formación de nubes de gas tras una fuga de sustancias gaseosas (categoría PLG incluida) y tóxicas, sin incendio ni explosión posterior y con un número de muertos significativamente reducido.

La clase C2 agrupa accidentes de sustancias líquidas, inflamables, no explosivas y no tóxicas producidos por incendio y que han causado un número reducido de muertos (NM2). Esta clase concentra más de la mitad de los registros de la muestra. Se han ensayado procedimientos para dividirla en dos o más clases con un número de elementos menor, pero las clases obtenidas no tenían sentido físico alguno.

Y la clase C3 representa accidentes con sustancias explosivas, no inflamables, en estado sólido o en polvo, producidos en actividades comerciales o instalaciones de proceso, con elevado número de muertos y mayoritariamente en países no desarrollados. Es normal que esta categoría agrupe muchos menos efectivos que las restantes ya que como se ha visto reiteradamente los accidentes graves son excepcionales.

Con el fin de verificar el grado de fiabilidad de los resultados obtenidos, se ha procedido a valorar la gravedad media (en número de muertos) para cada una de las clases obtenidas. Los resultados han sido los recogidos en la Tabla 5.15.

Tabla 5.15. Gravedad media en cada clase.

		Nº DE MUERTOS		
	EFECTIVOS	**TOTAL NPM**	**MEDIA**	**MÁXIMO**
CLASE 1	575	2.717	4,73	2.000
CLASE 2	999	3.325	3,33	500
CLASE 3	262	6.155	23,49	1.377
TOTAL	**1.836**	**12.197**	**6,64**	

Sin el accidente AN 1098 (Bhopal, 1984), la Clase 1 tendría una media de 1,24 muertos por accidente.

Como era previsible en razón de la descripción de las clases obtenidas, la clase 1 tiene una media de 23 muertos por accidente, frente a valores mucho más bajos en las demás categorías (la media general por accidente es de 6,6 muertos). De todas maneras hay que destacar que la variabilidad de los datos dentro de cada categoría es tan grande, que los márgenes de confianza de las medias obtenidas para cada clase se solapan claramente.

Asimismo debe tenerse presente que en la Clase 1 se encuentra incluido el accidente de Bhopal (Bhopal, 1984) con 2.000 muertos registrados. De no tener en cuenta este registro claramente extraordinario, la media de esa categoría sería de 1,24 muertos por accidente, resultado mucho más coherente con la descripción de esta clase que se ha hecho en este mismo apartado.

El estudio realizado de los accidentes seleccionados de la base de datos MHIDAS ha permitido obtener una clasificación de los mismos en tres clases claramente diferenciadas en función de su gravedad.

Se ha puesto de manifiesto el potencial que presenta la utilización combinada del análisis de correspondencias múltiples y del análisis de conglomerados con el fin de clasificar un conjunto de individuos dependiente de múltiples variables.

Sin embargo, la limitada información contenida en MHIDAS relativa a cada accidente no ha permitido ir más allá de una confirmación estadística de lo que el juicio experto y la experiencia ya han puesto de manifiesto sobradamente.

Una aportación significativa de este estudio es que la clasificación realizada ha permitido identificar qué factores permiten una máxima discriminación de los individuos y agruparlos en colectivos más homogéneos sobre los que se pueden desarrollar estudios posteriores más detallados. A la vista de la existencia de estas tres clases claramente diferenciadas, es obvio que un tratamiento masivo de estos accidentes no puede ser viable. De hecho, en los estudios posteriores presentados en el próximo capítulo, se ha trabajado con conjuntos limitados y homogéneos de registros para obtener unos modelos estadísticamente significativos.

5.5. MODELOS DE REGRESIÓN LOGIT

5.5.1. Objetivo

Se pretende estimar un modelo que permita la clasificación predictiva de los accidentes en función de su gravedad o, cuanto menos, conocer qué factores son más relevantes para determinar la gravedad de los accidentes.

Para llevar a cabo este análisis se ha utilizado el programa estadístico MINITAB, que dispone de un procedimiento estandarizado de regresión logit, con tres opciones disponibles: respuesta binomial, respuesta ordinal y respuesta nominal.

La respuesta binomial se utiliza cuando la variable respuesta o dependiente es binaria (si/no, verdadero/falso). El método ordinal es adecuado para aquellos casos en los que existe una relación de orden entre las n categorías de la variable respuesta, como puede ser el caso "muy grave", "grave", "normal", "leve" y "muy leve". Por último, el modelo nominal se utiliza para aquellos casos en los que no hay relación de orden entre las categorías de la variable dependiente (por ejemplo, "azul", "rojo" o "verde").

La diferencia analítica entre los dos últimos métodos es que en el caso de la respuesta ordinal se calcula una única pendiente para la función de regresión con diferentes ordenadas en el origen, mientras que para el último caso (respuesta nominal), se calcula una pendiente para cada categoría.

5.5.2. Preparación de los datos

Con el fin de adecuar los datos disponibles a los requisitos de esta técnica, se han preparado dos variables auxiliares que tienen la función de variables respuesta. La primera de ellas, denominada NPM2, clasifica los accidentes en dos categorías en función del número de muertos producidos, siendo 0 para accidentes sin muertos y 1 para aquellos accidentes en los que se ha producido, por lo menos, un fallecido. NPM2 es por lo tanto una variable binaria y por ello admite un tratamiento mediante el método de regresión logística binaria.

La segunda de las variables auxiliares se ha denominado NPM4 y categoriza el número de muertos en cinco grupos diferentes: "0", si el accidente no ha provocado víctimas mortales, "1" si se han producido menos de tres muertos, "2" si el número de muertos está entre 3 y 10, ambos incluidos; "3" si el número de muertos está entre 11 y 25 y "4" si es mayor de 25. Esta variable responde claramente a la tipología de ordinal descrita anteriormente.

Por lo que respecta a las restantes variables, se han utilizado las mismas que se han descrito en la Tabla 5.11 para el análisis de correspondencias múltiples, con la salvedad de que en este caso se ha conservado la variable original continua "año" y no la agrupada "década".

El conjunto de registros seleccionado para el análisis es el mismo que se ha utilizado en el apartado 5.4, con 2.216 registros.

5.5.3. Resultados

Se presentan en primer lugar los resultados obtenidos al aplicar este método estadístico a los datos de partida utilizando la variable NPM2 como variable dependiente o respuesta.

La distribución de los registros en función de la variable NPM2 es la siguiente: pertenecen a la categoría "0" (sin muertos) 999 registros y a la categoría "1" (con uno o más muertos) 918 registros. Los restantes 299 registros (hasta los 2.216 originales) corresponden a accidentes que tienen alguna de las variables significativas no definidas y por lo tanto no son utilizados en el cálculo de los coeficientes.

Tras múltiples ensayos en los que se han determinado las variables estadísticamente relevantes, se ha obtenido el resultado presentado en la Tabla 5.16.

Tabla 5.16. Resultados de la regresión logit binaria.

Predictor	Coeficiente	Desviación estándar	Z	P
Constante	-0,2252	0,3134	-0,72	0,472
Año	-0,011870	0,003905	-3,04	0,002
MHEX	1,2796	0,1994	6,42	0,000
IT1EXPLO	1,3702	0,1133	12,09	0,000
IT1GASCLD	-0,5157	0,1638	-3,15	0,002
MUNDO				
2	1,9197	0,2774	6,92	0,000
3	1,1241	0,1660	6,77	0,000
EFGAS	0,4620	0,1285	3,6	0,000

La interpretación de este resultado es la siguiente:

Si s es la probabilidad de pertenencia al grupo "0", lógicamente (1-s) es la probabilidad de pertenencia al grupo "1". El modelo logit establece que:

$$\ln\left[\frac{s}{1-s}\right] = d = \beta_0 + \sum_{j=1}^{j=c} \beta_j x_j \qquad (5.1)$$

Donde los coeficientes β son los presentados en la Tabla 5.16. Una observación de estos coeficientes permite detectar las mismas tendencias observadas en el apartado anterior gracias al análisis de correspondencias múltiples:

1. Las variables más significativas para determinar la gravedad de los accidentes son IT1EXPLO, MHEX y MUNDO.

2. Cuanto más reciente es el accidente (mayor AÑO), menos graves tienden a ser sus consecuencias.

3. SI la sustancia es explosiva (MHEX=1), la probabilidad de pertenecer al grupo "1" aumenta.

4. Sucede lo mismo si el accidente es por explosión (IT1EXPLO=1).

5. Sin embargo la formación de nubes de gas (IT1GASCLD=1) tiende a reducir la probabilidad de pertenencia al grupo "1".

6. Los accidentes ocurridos en países industrializados (MUNDO=1) son menos graves que los ocurridos en países en vías de desarrollo o no desarrollados.

Los valores de p inferiores a 0,05 en la Tabla 5.16 indican que los coeficientes de las variables seleccionadas como representativas son significativamente no nulos. Además, como medida adicional, se ha descartado la hipótesis de que todos los coeficientes sean nulos.

Se han obtenido dos parámetros que evalúan la bondad de los resultados alcanzados. El test de la D de Somers da un resultado de 0,57 en un rango de entre 0 y 1 (cuanto más próximo a uno mejor se considera el ajuste) y el test de la Gamma de Goodman-Kruskal da también como resultado 0,57 y debe interpretarse de la misma forma.

Para estimar la utilidad predictiva del modelo se han llevado a cabo dos análisis concretos. Uno de ellos proporcionado por el propio programa MINITAB y otro aplicando las matrices de confusión a los datos utilizados y a otros de nuevos.

Para el primero de los análisis se cruzan los datos de las dos clases dos a dos, con lo que se obtienen 917.082 pares (999 x 918). Aplicando la función logit obtenida a cada una de estas parejas, se considera "acierto" si se obtiene mayor probabilidad de pertenencia a la clase de referencia para el individuo que en realidad pertenece a esa clase y viceversa, se considera "fallo" si las probabilidades obtenidas tienden a indicar lo contrario.

Aplicando este procedimiento con el modelo obtenido se obtiene un 78,3% de aciertos frente a un 21,3% de fallos (y un 0,4% de igualdad de probabilidades).

Aplicando la matriz de confusión a los datos de partida, los resultados obtenidos son similares, tal como se indica en la Tabla 5.17.

Tabla 5.17. Matriz de confusión para el modelo logit de dos categorías.

	REAL		
ESTIMADA	CLASE 0	CLASE 1	TOTAL
CLASE 0	707	253	960
CLASE 1	292	665	957
TOTAL	**999**	**918**	**1.917**

El porcentaje de aciertos en la clasificación apriorística en este caso es del 71,6% (1.372 aciertos/1.917 individuos totales). Considérese que hay 299 individuos no clasificables por carecer de información en alguna de las variables relevantes para la aplicación del modelo.

Aplicando el mismo procedimiento a un conjunto de 36 individuos de referencia no utilizados en la obtención de los parámetros del modelo, el porcentaje de aciertos ha sido del 70,9% (20 puntos porcentuales más que si se hubiera hecho la clasificación de forma totalmente aleatoria)

Estos resultados confirman con rigor estadístico lo que la práctica o el juicio experto ya han puesto de manifiesto con anterioridad. El modelo obtenido tiene una eficacia predictiva muy limitada, principalmente como consecuencia de que las variables disponibles en MHIDAS no son las realmente influyentes en la gravedad del evento.

Es importante resaltar que variables como el origen "OG" o la causa general "GC" no han aparecido en el modelo final como consecuencia de un significativo número de registros no informados.

Se presenta a continuación el resultado obtenido utilizando el modelo logit ordinal. Mediante la variable NPM4 tal como se ha definido, la clasificación obtenida es la presentada en la Tabla 5.18.

Tabla 5.18. Clasificación de los accidentes mediante la variable NPM4.

NPM4	**CRITERIO**	**Nº ACCTES.**
0	NPM=0	999
1	0<NPM<3	448
2	2<NPM<11	330
3	10<NPM<26	85
4	NPM<25	55
TOTAL		**1.917**

Es evidente que la distribución no presenta en absoluto normalidad.

Los coeficientes obtenidos en este caso son los que se presentan en la Tabla 5.19.

Tabla 5.19. Resultados de la regresión logit ordinal.

Predictor	Coeficiente	Desviación estándar	Z	P
Constante 1	-0,6988	0,2679	-2,61	0,009
Constante 2	0,6868	0,2685	2,56	0,011
Constante 3	2,4723	0,2775	8,91	0,000
Constante 4	3,6666	0,2984	12,29	0,000
Año	0,020578	0,003220	6,39	0,000
MHEX	-1,0390	0,1469	-7,07	0,000
IT1EXPLO	-1,4208	0,1011	-14,05	0,000
EFPLGAS	0,2804	0,1397	2,01	0,045
MUNDO				
2	-1,8877	0,1778	-10,62	0,000
3	-1,3062	0,1330	-9,82	0,000
EFLIQUID	0,3592	0,1073	3,35	0,001

Los valores de p inferiores a 0,05 garantizan que las variables seleccionadas son ciertamente relevantes.

Igual que en el caso anterior se ha verificado que al menos no todos los coeficientes obtenidos son nulos y se ha calculado la D de Somer (0,52) y la Gamma de Goodman-Kruskal (0,52).

Aplicando la función logit obtenida a cada una de las parejas que es posible obtener al combinar individuos de grupos diferentes se obtienen 1.178.418 pares. En este caso se considera "acierto" si el individuo de un grupo superior tiene una probabilidad más elevada de pertenecer a un grupo superior al otro y "fallo" si sucede lo contrario.

Con el modelo calculado se obtiene un 75,7% de aciertos y un 23,9% de fallos.

Sin embargo la matriz de confusión obtenida al aplicar este modelo a los datos de partida presenta los resultados recogidos en la Tabla 5.20.

Tabla 5.20. Matriz de confusión para el modelo logit ordinal.

	ESTIMADA					
REAL	CLASE 0	CLASE 1	CLASE 2	CLASE 3	CLASE 4	TOTAL
CLASE 0	921	17	61	0	0	999
CLASE 1	335	26	87	0	0	448
CLASE 2	168	20	142	0	0	330
CLASE 3	31	3	51	0	0	85
CLASE 4	10	2	40	0	3	55
TOTAL	**1.465**	**68**	**381**	**0**	**3**	**1.917**

El porcentaje de aciertos en esta caso es del 56'9% (si se consideran como aciertos las clases adyacentes a la correcta, este porcentaje se eleva al 83,6%). Teniendo en cuenta que una clasificación aleatoria (considerando las frecuencias de cada clase) hubiera significado un 35,9% de aciertos, el modelo permite mejorar en 20 puntos porcentuales esta clasificación. Sin embargo, es evidente que la capacidad discriminante del mismo es realmente limitada: no atribuye individuos a la clase "3" y sólo asigna 3 a la categoría de accidentes "muy graves" (clase "4"). Estas limitaciones

son principalmente debidas a que los registros disponibles se concentran mayoritariamente en las primeras categorías. De otro lado son evidentes las limitaciones de los datos de partida con respecto a la densidad de población en el lugar del accidente, tipología del evento, energía de combustión o toxicidad de la sustancia interviniente, etc.

5.6. ÁRBOLES DE CLASIFICACIÓN

5.6.1. Objetivo

Se pretende llevar a cabo una clasificación de los accidentes seleccionados en función de su gravedad (valorada en función del número de víctimas mortales ocasionadas) mediante árboles de decisión. Para ejecutar este análisis se ha hecho uso del programa "Knowledge Seeker" especializado en la aplicación de esta técnica según se ha descrito en el apartado 5.1.

La principal diferencia respecto a la clasificación obtenida mediante el análisis de correspondencias múltiples combinado con el análisis de conglomerados responde a que mientras en este último caso los grupos obtenidos finalmente no están definidos a priori, en el método que ahora se presenta (igual que en la aplicación de modelos logit), los grupos en los que se clasifican los individuos están definidos a priori y responden a unos criterios lógicos determinados.

5.6.2. Preparación de los datos

Para aplicar este método a los registros de MHIDAS se ha hecho uso de la misma estructura de la información dispuesta para los modelos logit, con la salvedad de que se ha recategorizado la variable NPM4 en cuatro categorías. Así en este caso, la variable NPM4 distribuye los accidentes de la siguiente manera (Tabla 5.21):

Tabla 5.21. Clasificación de los accidentes mediante la variable NPM4.

NPM4	CRITERIO	Nº ACCTES.
A	NPM<3	1.033
B	2<NPM<11	246
C	10<NPM<26	80
D	NPM>25	60
TOTAL		**1.419**

Por lo demás las variables utilizadas en esta aplicación han sido las mismas que se indican en la Tabla 5.11.

5.6.3. Resultados

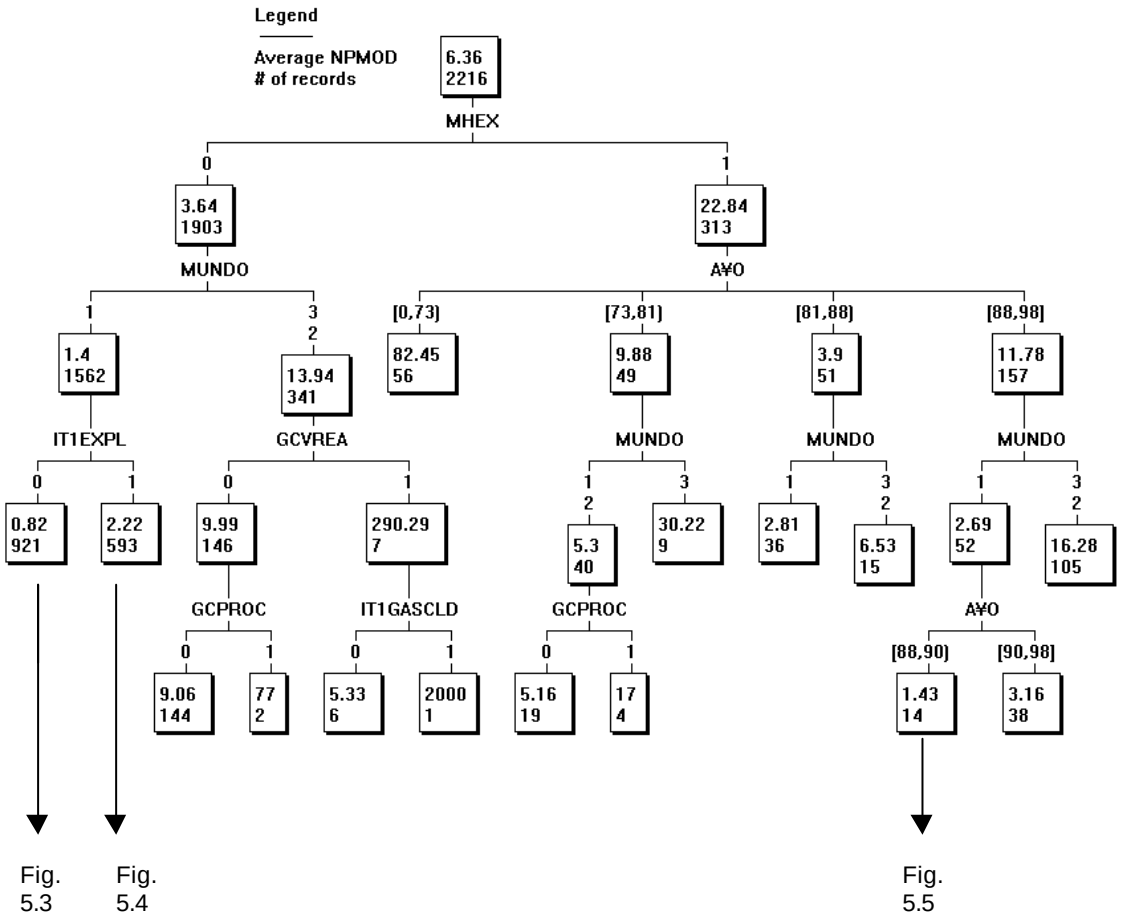
La aplicación de esta metodología permite la obtención de una representación gráfica de la clasificación final obtenida en forma de árbol.

Al ser una clasificación jerárquica, se puede identificar en cada nodo cual es la variable que permite una mejor discriminación entre los individuos garantizando la

representatividad estadística de esta partición para un determinado margen de confianza (en este caso α=0,05). Asimismo, para cada nodo, se puede conocer el número de individuos atribuidos a cada grupo y la media y la desviación estándar de la variable clasificadora (en este caso la variable continua NPM).

La clasificación en los primeros niveles es la que se muestra en la Figura 5.2.

Fig. 5.2. Clasificación en los primeros niveles mediante el programa KS (1).



La Figura 5.2 permite apreciar que la variable que en primer término resulta más significativa para clasificar los accidentes analizados es MHEX. Es decir, que el factor más determinante para establecer la gravedad de un accidente es saber si en el mismo participan sustancias explosivas. En general, la media del número de muertos en todo el conjunto de registros es de 6,36 (primer nodo). Sin embargo, para el conjunto de accidentes sin sustancias explosivas (MHEX=0), esta media se reduce a 3,64 frente a los 22,84 de los accidentes en los que intervienen sustancias explosivas.

Esta primera separación permite dividir el conjunto inicial en dos colectivos de 1.903 y 313 registros respectivamente.

A su vez, los accidentes que se producen sin intervención de sustancias explosivas se dividen en un segundo nivel en función del grado de desarrollo del país de ocurrencia. Si se trata de un país avanzado industrialmente, la media de número de muertos es

1,4 (1.562 registros) frente a 13,94 (341 registros) de los países en vías de desarrollo o no desarrollados.

La clasificación sigue sucesivamente hasta conformar la totalidad del árbol de decisión. El proceso se detiene cuando en un nodo determinado no hay evidencia estadística de que dos o más grupos sean distintos entre sí para el margen de confianza dado.

Es importante destacar que esta metodología permite adoptar la variable clasificadora que mayor potencial discriminante aporta en cada nodo adaptando en cada caso la necesaria, mientras que en los modelos logit descritos, las variables independientes utilizadas en la clasificación son las mismas para todo el conjunto. Ello hace posible que en ciertos niveles del árbol aparezcan variables significativas como GCHUMAN o OG, que en las otras técnicas han resultado no significativas.

Resulta también relevante que determinadas terminaciones del árbol están compuestas por pocos individuos (en ocasiones uno). A pesar de la poca relevancia estadística de estos grupos, se mantienen por separado para poner de evidencia que el nodo inmediatamente superior está claramente afectado por la presencia de individuos heterogéneos. Tal es el caso del nodo terminal que presenta un único individuo con un total de 2.000 muertos (accidente de Bhopal en 1984).

Las Figuras 5.3 a 5.5 presentan la totalidad de los árboles de clasificación obtenidos.

A cada terminación del árbol de clasificación obtenido se le asigna el grupo A, B, C o D en función del número de muertos medio y según la Tabla 5.21 presentada. Esta categorización de la variable respuesta permite analizar la bondad de los resultados en forma de matriz de confusión, como en los casos anteriores.

Tabla 5.22. Matriz de confusión para el modelo obtenido con el método KS.

	ESTIMADA				
REAL	CLASE A	CLASE B	CLASE C	CLASE D	TOTAL
CLASE A	666	245	19	103	1.033
CLASE B	43	108	43	52	246
CLASE C	6	27	27	20	80
CLASE D	0	11	20	29	60
TOTAL	**715**	**391**	**109**	**204**	**1.419**

Con esta clasificación se obtienen 830 aciertos (un 58,5%). Si se admite como acierto la asignación a las clases contiguas a las reales el porcentaje de acierto se eleva al 83,5%.

Debe observarse que la detección de individuos de las clases que representan mayor gravedad es sustancialmente diferente a la obtenida en el apartado anterior utilizando los modelos logit. De los 60 accidentes clasificados como graves (más de 25 muertos), la clasificación obtenida permite "detectar" 49 (entre 11 y 2.000 muertos). Pero a su vez también clasifica 155 accidentes como graves cuando en realidad no han llegado a superar los diez muertos. Este tipo de error es el que se identifica habitualmente como "falsos positivos" y por lo tanto la clasificación obtenida utilizada con finalidad predictiva tiende a pronosticar resultados más graves de los que se producen en realidad.

Fig. 5.3. Clasificación mediante el programa KS (2).

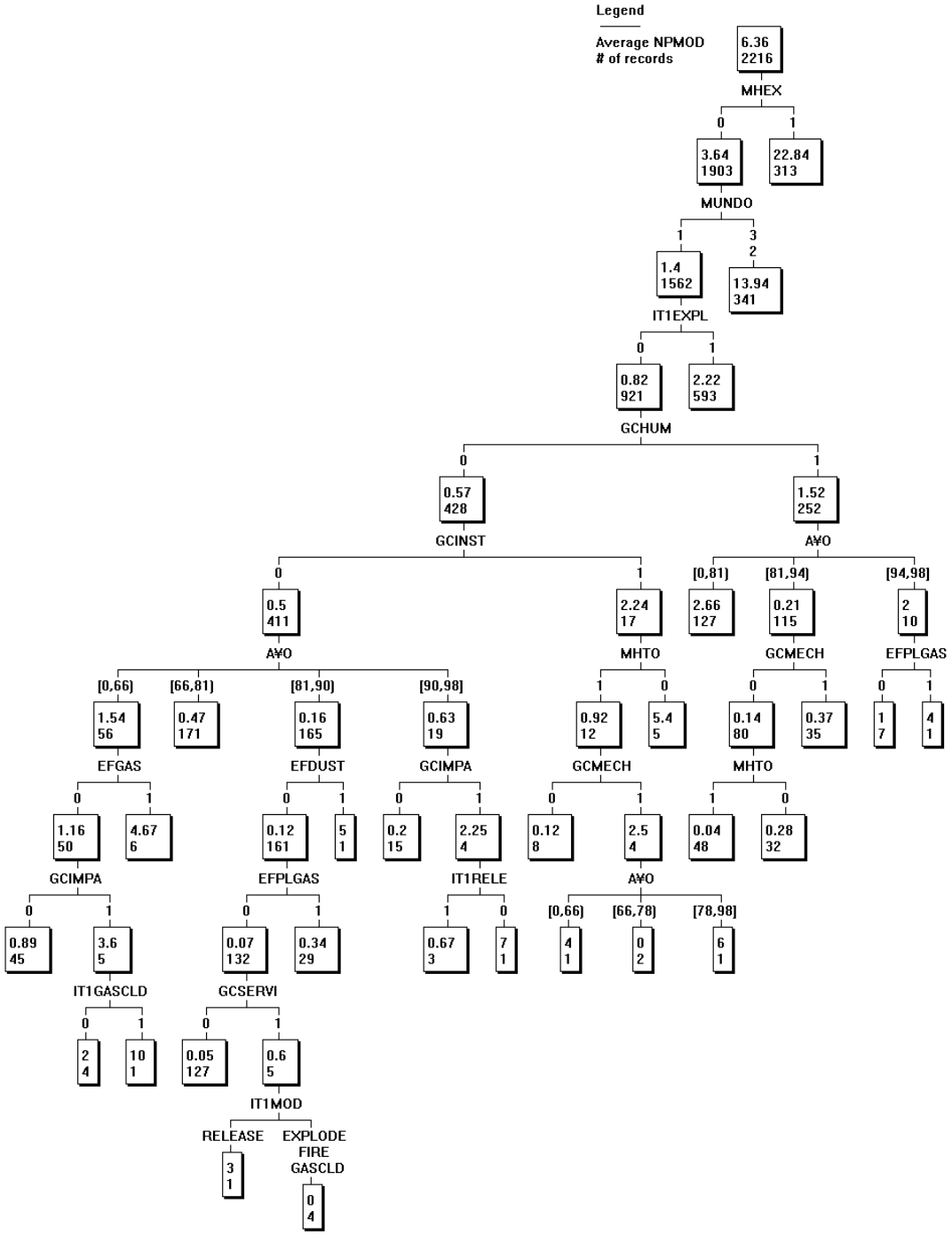


Fig. 5.4. Clasificación mediante el programa KS (3).

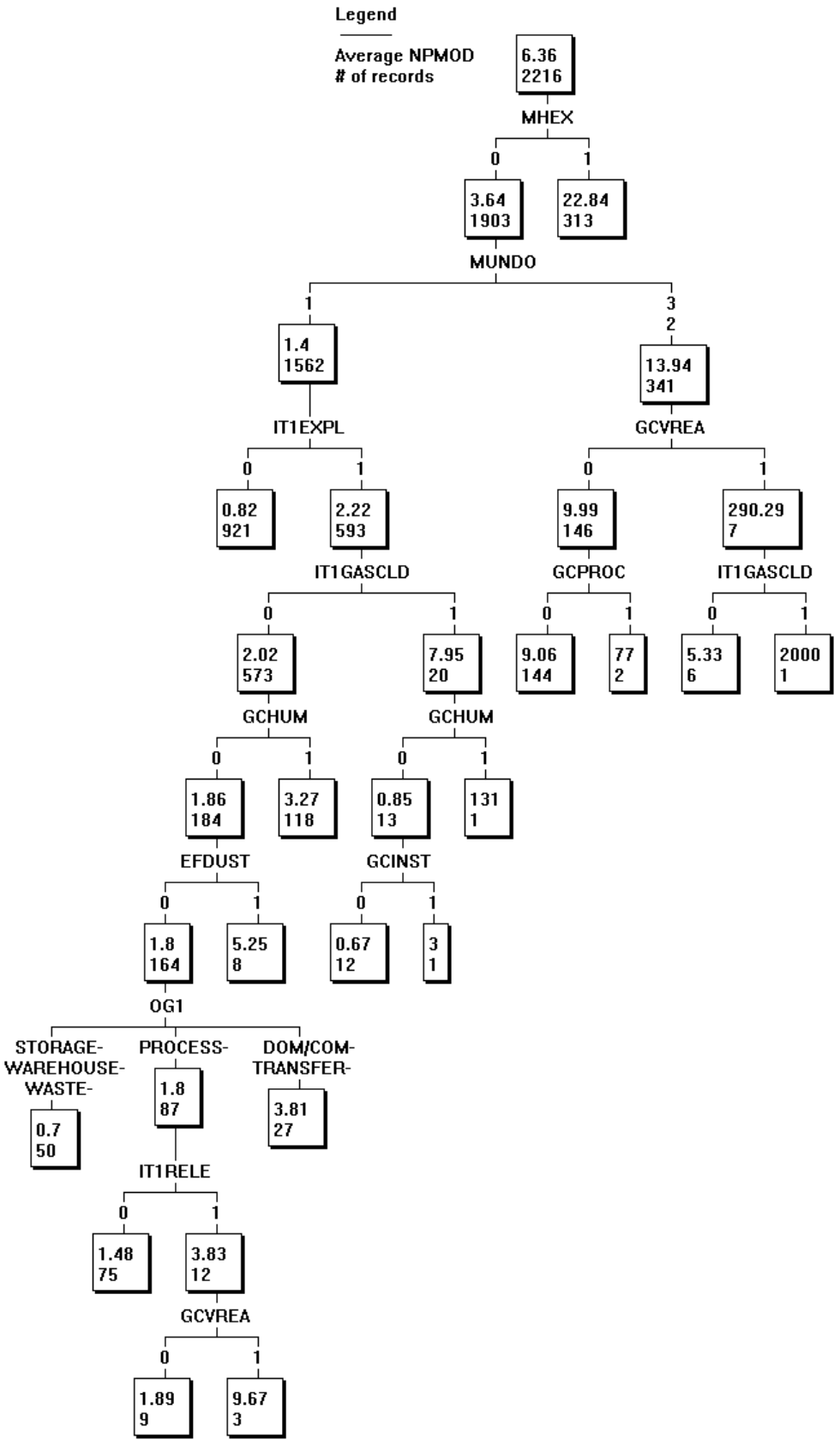
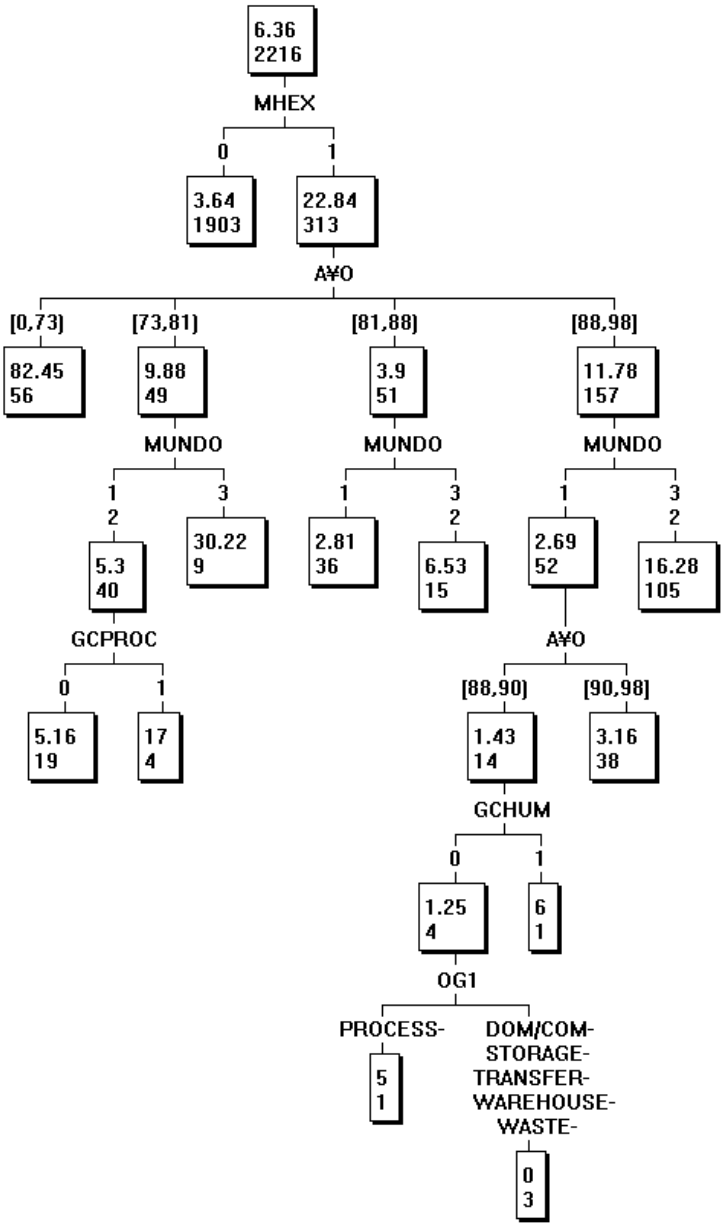


Fig. 5.5. Clasificación mediante el programa KS (y 4).



El porcentaje de la varianza explicada es del 56,7%. Esto significa que con las terminaciones del árbol se predice de media un 56,7% mejor el número de muertos que si se utiliza la media del nodo raíz.

El principal inconveniente de esta técnica es que su desarrollo (el "crecimiento del árbol") está fuertemente influenciado por los datos originales. Ello implica que en la medida en la que en los datos de partida se incluyan circunstancias especiales, éstas afectarán a las futuras posibilidades predictivas del método, desviándolo notoriamente. Esta circunstancia se hace evidente al aplicar la clasificación propuesta a un conjunto de 36 accidentes no utilizados en el desarrollo del árbol y para los que se obtiene un 49% de aciertos (frente al 58,5% indicado para los datos usados en la creación del

árbol de clasificación). De todo ello se desprende que la clasificación obtenida en los últimos niveles de los árboles presentados no es representativa. Una posible ampliación de la técnica planteada es determinar hasta qué nivel del árbol la clasificación es estadísticamente significativa. Asimismo, también resultaría enormemente útil poder establecer los márgenes de confianza para las medias de cada categoría.

5.7. CONCLUSIONES

En el presente capítulo se han analizado los accidentes con sustancias peligrosas mediante un enfoque multivariante y utilizando varias técnicas estadísticas. Como se expone en la introducción el objetivo final es obtener una herramienta predictiva que permitiera estimar o pronosticar la gravedad de un evento en función de ciertas variables o, cuanto menos, evidenciar qué factores son realmente influyentes en la gravedad de estos accidentes.

Una vez aplicadas las tres técnicas presentadas (análisis de conglomerados combinado con un análisis de correspondencias múltiples, análisis logit y análisis mediante árboles de fallos) es evidente la dificultad de utilizar estos métodos (con los datos disponibles) con fines predictivos dada la escasa representatividad de las clasificaciones obtenidas

Las tres metodologías ponen de manifiesto que:

1. Los factores registrados en MHIDAS no son suficientes ni suficientemente representativos como para que sean utilizados para estimar la gravedad de los accidentes. Existen muchas otras variables no registradas con una influencia mucho más relevante en la gravedad.

2. Los factores más significativos son, para los tres métodos, la presencia de sustancias explosivas, el nivel de desarrollo del país donde sucede el evento, la antigüedad del mismo y la ocurrencia de explosiones o incendios, frente a derrames o fugas sin más consecuencias. Estos resultados coinciden con los obtenidos con la utilización del análisis de regresión múltiple que fueron objeto de un completo estudio en un Proyecto Final de Carrera de la Diplomatura de Estadística de la UPC.

3. Lo expuesto en el punto anterior permite confirmar lo que el juicio experto ya tiene por evidente; un análisis detallado de las clasificaciones obtenidas no permite extraer otras conclusiones más allá de las expuestas.

4. Sin embargo, pese a lo anterior, se pone claramente de manifiesto que el conjunto de accidentes analizado no puede ser tratado como un conjunto homogéneo sino que está integrado por subconjuntos claramente diferenciados entre sí y que deben ser tenidos en cuenta a la hora de extraer conclusiones o desarrollar estudios globales. Esta constatación es utilizada en el Capítulo 6 de esta memoria para desarrollar tres estudios concretos, dos de los cuales contemplan tipologías específicas de accidente (explosiones de sustancias inflamables en países desarrollados y de ocurrencia reciente, por ejemplo).

La falta de información complementaria sobre los accidentes es claramente evidente en lo que respecta a la densidad de población en la zona de influencia del evento y el porcentaje de población afectado por el mismo. Con la información disponible en

MHIDAS se clasificarían en el mismo grupo de "sin consecuencias mortales" tanto el incidente menos relevante como el accidente de Seveso ocurrido en 1976, por ejemplo.
Las metodologías estadísticas utilizadas en este capítulo han puesto de manifiesto su potencial indiscutible en este ámbito. Sin embargo debe mejorarse sustancialmente la calidad de los datos de partida con el fin de obtener unos resultados más representativos con su aplicación. En la Tabla 5.23 se indican algunas variables que se consideran indispensables para obtener una mejora sustancial en los resultados. Todas ellas han aparecido como relevantes en el Capítulo 2 (índices de riesgo) o en el Capítulo 3 (bases de datos y otras referencias bibliográficas) y algunas están previstas en la base de datos MARS en cuya futura cumplimentación cabe poner esperanzas (Anexo 1).

Tabla 5.23. Factores determinantes de las consecuencias de los accidentes no recogidos en MHIDAS.

FACTOR	DESCRIPCIÓN
Sobre la instalación	
Actividad industrial	Sector de actividad Capacidad productiva Equipos e instalaciones Grado de automatización
Ubicación y dimensiones	Dimensiones de la planta Ubicación (en polígono industrial, entorno urbano,...) Distancia a núcleos urbanos
Medidas de prevención	Sistemas de control del proceso Sistemas de alivio, paralización programada, etc.
Medidas de protección	Sistemas de rociadores Columnas hidrantes exteriores Equipos de primera y segunda intervención Posibilidades de contención del efecto dominó
Sobre el proceso	
Equipo causante	Instalación origen: reactor, separador, columna de destilación,... Volumen Material de construcción
Condiciones de proceso	Temperatura, presión, volumen, agitación,...
Sensibilidad del proceso	Posibles desviaciones agravantes Sensibilidad a impurezas, catalizadores,...
Sobre las sustancias	
Propiedades físico químicas	Densidad, viscosidad, calor de combustión Estado físico Toxicidad Reactividad con otras sustancias. Energías de reacción
Cantidad interviniente en cada suceso	Cantidad de cada sustancia participante en cada suceso del accidente (cantidad derramada, cantidad incendiada,...)

Tabla 5.23. (Cont.) Factores determinantes de las consecuencias de los accidentes no recogidos en MHIDAS.

Condiciones de almacenamiento	Forma de empaquetamiento Ubicación del almacenamiento Tamaños de pila Posibilidades de separación de sustancias reactivas entre sí
Sobre el evento	
Secuencia de ocurrencia	Secuencia en la que se desarrollan los eventos
Especificación de los sucesos	Especificación de cada uno de ellos: Tamaño, duración, ...
Actuación de emergencia	
Sobre el entorno	
Densidad de población en el entorno	Población en la zona de influencia del evento Personal de planta/personal de intervención/público en general
Facilidades para la evacuación	Vías de acceso, transportes internos de la planta y externos
Condiciones ambientales	Temperatura y humedad Viento: velocidad y dirección
Sobre las consecuencias	
Nº de muertos	Especificación de la causa (onda expansiva, quemaduras, asfixia,...) Especificación de su ubicación (posición, distancia al foco,...) Condición (personal de planta, equipo de intervención, personal no formado,...) Edad Equipos de protección individual
Nº de heridos	Severidad de las lesiones: Hospitalizados, no hospitalizados, con o sin baja, control médico periódico Especificación de la causa (onda expansiva, quemaduras, asfixia,...) Especificación de su ubicación (posición, distancia al foco,...) Condición (personal de planta, equipo de intervención, personal no formado,...) Edad Equipos de protección individual
Nº de evacuados	Especificación de su ubicación (posición, distancia al foco,...) Condición (personal de planta, equipo de intervención, personal no formado,...) Edad Equipos de protección individual
Flora y fauna	Nº de individuos afectados. Tipología Extensión afectada
Contaminación de cauces naturales de agua	Cantidad de m^3 afectados Tipo de curso de agua afectado (lago, embalse, río,...)

Tabla 5.23. (Cont.) Factores determinantes de las consecuencias de los accidentes no recogidos en MHIDAS.

Contaminación de suelos	Metros cúbicos afectados, superficie afectada Efectos sobre el suelo
Contaminación del aire	Formación de nube tóxica: Tamaño, duración, dispersión Zonas afectadas
Daños materiales	En la instalación / En el entorno Propiedad pública (viales, accesos...) o privada Valoración económica. % del total expuesto al evento
Pérdidas por paralización	Cuantificación económica de las pérdidas a consecuencia de la paralización de la planta

En lo relativo a la calidad de los datos necesarios para el análisis, es importante poder disponer de una muestra no sesgada (muchos accidentes sin muertos y sin consecuencias externas pasan desapercibidos y no son registrados) y cumplimentada con criterios homogéneos e información técnicamente contrastada.