

APLICACIÓN DEL ANÁLISIS CLUSTER **EN EL ÁMBITO ECONÓMICO**

UN ESTUDIO DE LOS BARRIOS DE

DONOSTIA

Mariano Jiménez López

Amaya Zárraga Castro

Marian Zubia Zubiaurre

- Este trabajo ha sido financiado por los proyectos de investigación PB95-0396 y PB98-0149 de la Dirección General de Enseñanza Superior del Ministerio de Educación y Cultura y por los proyectos UPV 038.321-HC236/97 y UPV 038.321-HA041/99 de la Universidad del País Vasco (UPV/EHU).

APLICACIÓN DEL ANÁLISIS CLUSTER EN EL ÁMBITO ECONÓMICO

UN ESTUDIO DE LOS BARRIOS DE DONOSTIA

Resumen

Con este trabajo tratamos de estudiar la calidad de vida de Donostia, en el sentido de que queremos estudiar la relación que puede existir entre los habitantes de las viviendas y la estructura y utilización de las mismas. Intentaremos buscar un vínculo entre tipo de habitante, tipo de vivienda y áreas homogéneas de Donostia, utilizando diversas técnicas de Análisis Multivariante, como por ejemplo el Análisis de Componentes Principales, que es una técnica de reducción de la dimensionalidad de un conjunto de variables continuas, para poder recoger en pocos factores o componentes, la mayor parte de la información de esas variables. A posteriori, aplicaremos el Análisis Cluster y también el Análisis Fuzzy Clustering para formar los grupos de barrios homogéneos, donde la principal diferencia radica en que el método Fuzzy permite diversos grados de pertenencia de un barrio a diversos grupos, mientras que el método Cluster es menos flexible, en el sentido de que un barrio pertenece o no a un grupo.

PALABRAS CLAVE: Análisis Cluster, Análisis Fuzzy, Análisis Multivariante, viviendas, condiciones socioeconómicas

INDICE

1.INTRODUCCION.....	pág. 4
2.INDICADORES Y UNIDADES DEL ESTUDIO.....	pág. 5
3.METODOLOGIA.....	pág. 7
4.RESULTADOS.....	pág. 12
5.ANALISIS FUZZY.....	pág. 24
6.BIBLIOGRAFIA.....	pág. 29

1.INTRODUCCIÓN

Donostia nació en el siglo XII en el reinado de Sancho el Sabio (1150-1194) consiguiendo la capitalidad de Gipuzkoa en 1854. Desde el punto de vista de urbanización de la ciudad, La Concha ha ejercido funciones de epicentro de la ciudad. Antiguamente la ciudad se asentaba en torno al monasterio de San Sebastián el Antiguo, donde hoy día se encuentra el Palacio de Miramar. Pero por el dinamismo que adquirió por sus actividades, sobre todo a partir del derribo de las fortificaciones (1864), pronto necesitó una expansión de la superficie urbanizada, ya que Donostia constituía uno de los puntos clave para el turismo y para el comercio, así como para el proceso de industrialización, creándose algunas empresas, potenciando las comunicaciones y al fin y al cabo, organizando las actividades económicas y culturales de la provincia.

Para la expansión se comenzó ocupando los terrenos con mejores condiciones, es decir los que hoy día constituyen el centro de la ciudad, y poco a poco se expandieron hacia las zonas que componen hoy día los barrios de Miraconcha, Aiete e Igeldo. Zonas privilegiadas por el paisaje de la bahía de la Concha, la cercanía al Palacio de Miramar y a la playa. Por el precio que adquirieron estos terrenos, fueron los burgueses los que se instalaron allí, y en la actualidad se pueden observar todavía palacetes, villas y viviendas de gran calidad.

Pero también aumentó el proletariado, y se inició la ocupación de más terrenos llanos, aunque más alejados de la ciudad, y por tanto, con un precio más asequible. Fueron las zonas que ahora constituyen los barrios de Gros, El Antiguo, Egia, Amara Viejo, Añorga, Loyola y Martutene. Aunque comenzaron como barrios obreros, algunos de estos barrios cambiaron ese carácter y pudieron acomodar también a las clases altas sin dificultad. Es el caso de Gros y El Antiguo, ya que el primero por ejemplo, estaba situado cerca del centro y tenía la estación de ferrocarril y la carretera hacia Francia cerca, así como magníficas vistas al mar. Así mismo, El Antiguo también se transformó, sobre todo porque estaba cerca del Palacio de Miramar y de la playa, pasaban por esta zona las carreteras Madrid- Irún y la de Lasarte, y además había acceso o paso hacia Aiete, Igeldo, Benta Berri, es decir a zonas donde, con un precio más asequible, se estaba instalando la burguesía.

Los demás barrios en cambio, siguieron con su carácter rural y obrero, porque estaban bastante lejos del centro y se tardó en su saneamiento.

A partir de la Guerra Civil y hasta los años 60, el crecimiento de Donostia se limitó a continuar las líneas de acción ya creadas. Y actualmente está compuesta principalmente por 19 barrios: Aiete, Alza, Amara Berri, Amara Viejo, Añorga, Antiguo, Bidebieta, Centro, Egia, Gros, Herrera, Ibaeta, Igeldo, Intxaurreondo, Loyola, Martutene, Miraconcha, Parte Vieja y Ulía.

Quizás hoy en día no podamos observar esa diferencia sociológica y económica tan patente a finales del siglo pasado y a comienzos del siglo XX entre burgueses y la clase obrera. Pero con este estudio intentaremos buscar cierta homogeneidad actual entre estos barrios, teniendo en cuenta las características actuales de las viviendas y las condiciones socioeconómicas de los habitantes de las mismas, e intentaremos estudiar la situación de estos barrios actualmente, y si hoy día podemos encontrar esa diferenciación tan clara entre los barrios que hubo antaño.

En este estudio intentaremos analizar si es posible encontrar áreas homogéneas en Donostia, es decir, zonas de la ciudad en las que algunas características de las viviendas sean comunes y analizaremos si coinciden también las principales características de las personas que habitan en ellas. Así mismo, estudiaremos si se mantiene el vínculo pero en el sentido contrario, es decir si personas con características comunes habitan en viviendas de características similares.

2.INDICADORES Y UNIDADES DEL ESTUDIO

Tal y como lo hemos mencionado en el apartado anterior, Donostia actualmente está dividida en 19 barrios y son éstos las unidades de análisis de nuestro estudio. Para cada una de estas unidades, hemos obtenido información de 37 indicadores o variables, que recogen información tanto sobre la estructura y equipamiento de las viviendas, como de la situación socioeconómica de los habitantes de las mismas.

Los datos que hemos utilizado en el análisis, los hemos obtenido de las estadísticas de población y vivienda de 1991 y de 1996 proporcionadas por el EUSTAT, por lo que agradecemos su colaboración y ayuda. De toda la información que nos ha ofrecido el Eustat, hemos rescatado aquella que nos interesaba y a

continuación, para tener una homogeneidad en los datos, hemos tipificado los indicadores, obteniendo las 37 variables mencionadas en el párrafo anterior. Estas variables son: X1: % familias monoparentales, X2: % familias de un componente, X3: % familias de 6 o más componentes, X4: % de la población con profesión liberal, X5: % de la población con profesión autónoma, X6: % de empresarios, directores y profesiones superiores e intermedias, X7: % de empleados y trabajadores, X8: % de la población buscando su primer empleo, X9: % de la población jubilada, X10: % de amas de casa, X11: % de estudiantes, X12: % de la población mayor de 65 años, X13: % de la población menor de 14 años, X14: % de extranjeros residentes, X15: % de extranjeros no residentes, X16: % de divorciados, viudos y separados, X17: % de la población con estudios medios, X18: % graduados universitarios, X19: % de asistencia a guarderías, X20: % de viviendas propias (censo 91), X21: % de viviendas arrendadas (censo 91), X22: número medio de habitaciones de las viviendas, X23: % de viviendas con 1-2 habitaciones, X24: % de viviendas con 3-5 habitaciones, X25: % de viviendas con 6 o más habitaciones, X26: número medio de habitantes de las viviendas, X27: superficie media de las viviendas, X28: % de viviendas con calefacción, X29: % de viviendas sin calefacción, X30: % de viviendas construidas antes de 1900, X31: % de viviendas construidas entre 1901-1940, X32: % de viviendas construidas entre 1941-1950, X33: % de viviendas construidas entre 1951-1960, X34: % de viviendas construidas entre 1961-1970, X35: % de viviendas construidas entre 1971-1980, X36: % de viviendas construidas entre 1981-1990, X37: % de viviendas construidas después de 1991.

Como podemos observar, algunas de estas variables recogen información sobre la situación socioeconómica de los habitantes (X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15, X16, X17, X18, X19), otras en cambio recogen las necesidades espaciales de las familias y en qué grado están cubiertas, así como el confort que puedan tener (X1, X2, X3, X22, X26, X27, X28, X29). Y por último, tenemos aquellas variables que recogen las características básicas de las viviendas: el año de construcción, el número de habitaciones, si es propia o arrendada...

Salvo la información referente a viviendas propias o arrendadas (censo de 1991), el resto de la información se ha obtenido del censo de la población y vivienda de 1996.

3. METODOLOGÍA DEL ANÁLISIS

Como el objetivo de nuestro estudio es agrupar los individuos o unidades, en nuestro caso barrios de Donostia, sabemos que el análisis que vamos a realizar dependerá de las variables que tomemos y de las observaciones de las mismas, y que solamente servirá para los datos considerados. El principal problema es la extensión de los datos, y por consiguiente la dificultad a la hora de interpretar y a la hora de formar grupos. Dentro del conjunto de las técnicas multivariantes, hemos optado por aplicar un Análisis de Componentes Principales para intentar recoger en pocos factores la mayor variabilidad e información posible de las variables originales, y poder así facilitar el análisis.

El Análisis de Componentes Principales (ACP), es una técnica del Análisis Multivariante, que nos permite la reducción de la dimensionalidad de un conjunto de variables continuas, intentando mantener la máxima información posible. De todas formas, en nuestro caso el ACP es una etapa intermedia, ya que en un análisis posterior, utilizaremos las coordenadas de los factores que nos proporciona este análisis para calcular sus distancias y así, proceder a la agrupación de los individuos, en nuestro caso barrios.

Una vez dibujada la nube de puntos, bien la de los individuos o bien la de las variables, la idea principal del ACP es buscar un plano o recta, conocida como primer plano factorial, sobre el que se proyecta la nube de puntos original, de modo que la configuración obtenida es lo máximo posible parecida a la configuración de la nube original, es decir el conjunto de distancias medida en el plano factorial para todas las parejas de puntos, se aproxime lo máximo posible a las distancias reales medidas en el espacio original. O en términos de inercia del punto o el producto del peso de un punto por el cuadrado de su distancia al centro de gravedad, consiste en maximizar la inercia proyectada.

Tras encontrar el primer plano o recta, se intenta buscar un subespacio de dimensión dos, y después de dimensión tres, y después de dimensión cuatro,... procurando que la inercia proyectada sobre estos subespacios sea máxima. El plano de dimensión dos, contiene la recta de inercia proyectada máxima, y por tanto, si hemos encontrado la dirección de la inercia máxima, tendremos que encontrar otra recta, ortogonal a la primera, y tal que defina el plano de inercia máxima. El de

dimensión tres, por tanto, consistiría en hallar otra recta ortogonal al plano de dimensión dos.

Como la inercia es aditiva en direcciones ortogonales, el ACP consiste en descomponer la inercia total en un número de componentes aditivas, tantas como dimensión tenga el espacio original, de modo que las inercias son cada vez menores en las direcciones ortogonales sucesivas.

El ACP definido con la elección de la distancia euclídea entre los individuos o puntos fila, está ligado con el ACP sobre los puntos variable con la distancia definida a partir de la correlación entre las variables.

En nuestro estudio, tenemos 37 variables y por tanto, hemos calculado su matriz de correlaciones y aplicado el Análisis de Componentes Principales. Estos son algunos de los resultados obtenidos:

VALOR PROPIO:	25,1498	4,0028	2,5807	1,7658	1,1851	1,1338
	0,4062	0,3057	0,1838	0,1280	0,0635	0,0419
	0,0142	0,0098	0,0048	0,0044	0,0009	0,0189
% DE VARIANZA:	67,97%	10,82%	6,97%	4,77%	3,20%	3,06%
	1,10%	0,83%	0,50%	0,35%	0,17%	0,11%
	0,03%	0,01%	0,01		0,05%	0,04%
% VAR.ACUMUL.:	67,97%	78,79%	85,77%	90,54%	93,74%	96,81%
	97,90%	98,73%	99,23%	99,57%	99,74%	99,86%
	99,95%	99,97%	99,99%	100,00	99,91%	

Como podemos observar, tenemos seis valores propios superiores a la unidad. Según estos resultados, mediante estos valores propios podemos obtener vectores propios que determinan 6 factores, con los que somos capaces de recoger el 96,81 % de la variación o inercia. Así mismo, recogemos también el porcentaje de la varianza explicada por cada eje, donde el primero recoge casi el 68% de la variación.

Ya hemos comentado que en nuestro estudio, el ACP es una etapa intermedia, ya que a continuación, usaremos las coordenadas o factores como nuevas variables y les aplicaremos un Análisis Cluster, tal que nos permita la subdivisión de los 19 barrios de Donostia en un número reducido de grupos. Es decir, tras realizar el ACP, en lugar de 37 variables tenemos 6 factores, y por tanto, para cada uno de los 19 barrios tenemos observaciones de cada uno de los factores, a los que aplicaremos el Análisis Cluster.

Tanto para el Análisis Factorial como para el Análisis Cluster, hemos utilizado el programa informático "DYANE, Diseño y Análisis de Encuestas" diseñado por Dr. Miguel Santesmases Mestre, que permite una amplia gama de tipos de análisis.

Con el Análisis Cluster tratamos de agrupar una colección de observaciones, de modo que el grado de asociación sea alto entre los miembros del grupo, y bajo entre miembros de grupos diferentes. Es decir, agrupamos objetos o individuos similares y buscamos patrones que puedan explicar o caracterizar esos grupos.

En primer lugar, hay que especificar una medida de similitud entre las observaciones y proceder a formar grupos. Para ello, hemos optado por métodos jerárquicos aglomerativos, construyendo un árbol en el que se representan las relaciones entre las unidades y desplazándonos desde las ramas del árbol hasta la raíz, formando en cada paso los grupos y finalizando con un único grupo en el que se recogen todos los individuos o barrios. También podríamos haber optado por métodos no jerárquicos, pero a diferencia del método anterior, el número de grupos final está especificado con antelación, y como nuestro objetivo al fin y al cabo es especificar cuántas zonas homogéneas de la ciudad podemos tener, y cuáles son sus características principales, no podemos establecer de antemano cuántas agrupaciones finales tendremos. El procedimiento de ambos métodos es similar, ya que consiste en comenzar con tantos grupos como individuos o entidades y a continuación, calculando la matriz de similitud, ir agrupando individuos y grupos.

Por sencillez, hemos escogido la distancia euclídea entre las coordenadas del ACP como medida de similitud entre dos individuos o barrios, agrupando barrios con menor distancia, y el método encadenamiento completo para formar los grupos, donde la medida de similitud entre los candidatos o grupos y el grupo viene determinado por la mayor de las distancias entre el candidato o integrantes del primero y cada uno de las unidades que forman el otro grupo. Es decir, la distancia euclídea entre dos barrios i e i' lo calculamos con la siguiente fórmula:

$$d(i, i')^2 = \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

donde los barrios i e i' están situados en un espacio $p=6$ dimensional. Si hemos comenzado con tantos grupos como barrios ($n=19$), una vez que hemos agrupado los dos individuos más cercanos, tendremos $n-1$ grupos, y tendremos que volver a calcular de nuevo la matriz de similitud entre este grupo y el resto de los individuos para volver a juntar dos individuos o añadir uno al grupo creado anteriormente. Así, iremos reduciendo el número de grupos hasta conseguir uno único con todos los individuos.

De esta forma y con las observaciones de los 19 barrios de Donostia que disponíamos, hemos obtenido el árbol y las ramas que comentábamos anteriormente, hasta conseguir un único grupo. Es decir, hemos aplicado el Análisis Cluster en las observaciones de los factores del Análisis de Componentes Principales, y hemos recogido el árbol resultante en el Gráfico 1 adjunto, donde los valores que aparecen en el gráfico recogen las distancias entre dos barrios o grupos de barrios.

El principal problema que puede surgir es cómo hacer el corte para diferencias los grupos creados, ya que si cortamos a una distancia mayor, menos grupos de barrios tendremos pero puede que no podamos caracterizar ese grupo tan claramente. De este modo, si hacemos el corte en la distancia 2,60 por ejemplo, podemos ver que se crean 8 grupos de barrios, es decir, pasamos de 19 grupos, cada uno compuesto por un barrio, a 8 grupos. El más amplio es el primero, en el que los integrantes son Herrera, Martutene, Ulia, Aiete, Ibaeta, Añorga y Bidebieta. En el segundo encontramos los barrios Amara Viejo, Loyola, Antiguo y Egia, mientras que en el tercero, Igeldo y Miraconcha, y en el cuarto, los barrios Centro y Parte Vieja. Por último, nos quedarán cuatro barrios que no se agrupan, los barrios Amara Berri, Alza, Intxaurreondo y Gros.

Tal y como hemos mencionado anteriormente, el primer y segundo grupo son muy amplios, en el sentido de que el primero está compuesto por 7 barrios y el segundo por 4. En estos dos casos, nos resultará bastante complicado buscar variables que caractericen a tantos barrios a la vez, por lo que resultará conveniente hacer un corte a una distancia menor, por ejemplo a la distancia 1,5 con el fin de proceder y facilitar la identificación de las principales características de los barrios que componen estos grupos. De esta forma, tendríamos 12 grupos en total, ya que los dos primeros grupos se desglosan en 5, es decir, Herrera, Martutene y Ulia sería el primer grupo, Aiete, Ibaeta y Añorga sería el segundo, Bidebieta el tercero, Amara Viejo y Loyola el cuarto, y por último el quinto grupo estaría compuesto por los barrios El Antiguo y Egia. El resto queda igual que en la partición anterior.

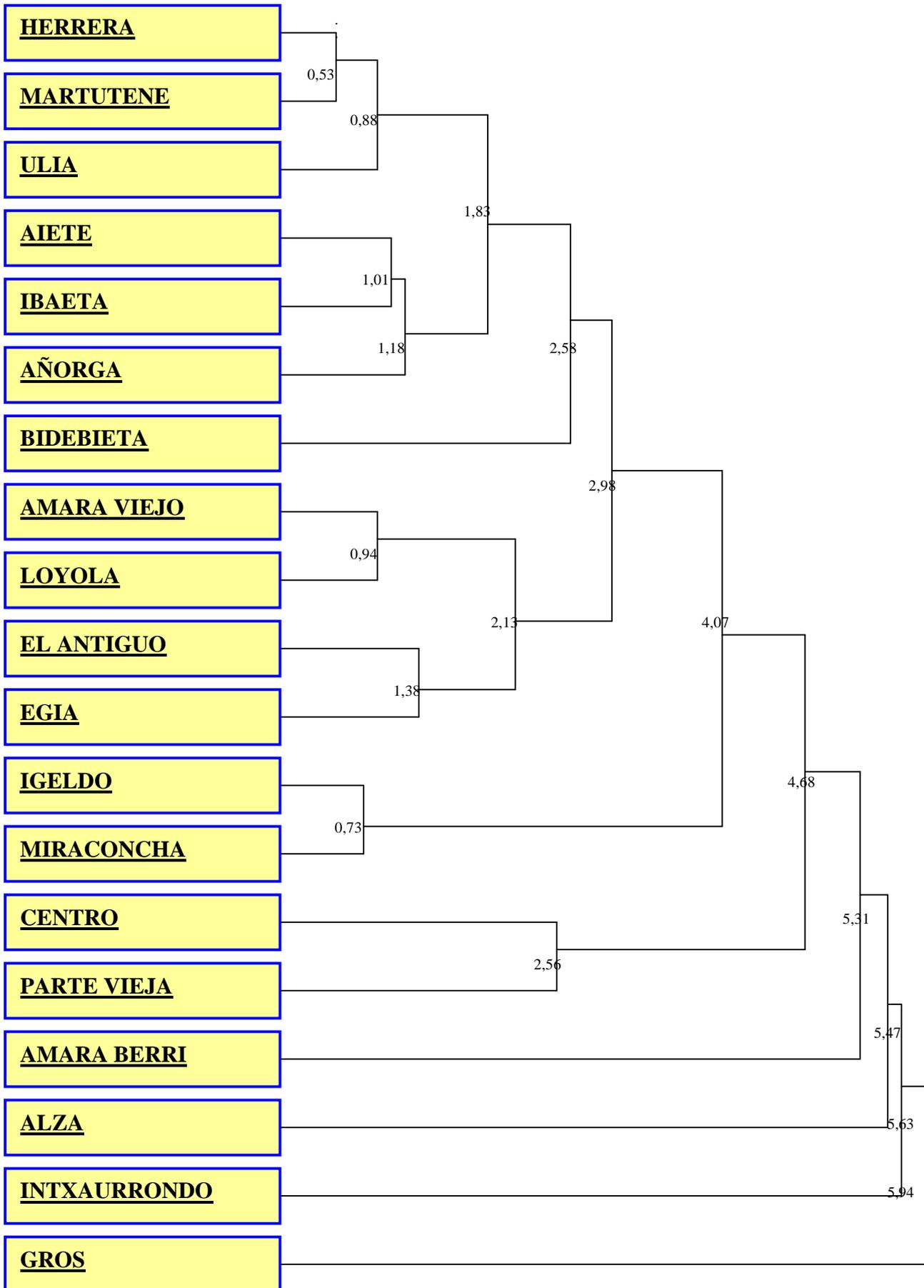


GRAFICO 1

4. RESULTADOS

Una vez que tenemos especificados los grupos, lo que nos interesa es determinar si podemos caracterizar cada grupo o zona de Donostia por el tipo de vivienda, así como por la tipología de los habitantes. Es decir, si podemos identificar cada zona por unas características específicas de esa zona y que no lo son de las demás.

Sea X es la variable medida en el municipio de Donostia, tal que sigue una distribución Normal de media m_x y varianza σ_x^2 , y sea X_i la variable en el grupo i tal que sigue una distribución Normal de media m_i y varianza σ_i^2 .

$$H_0: m_i = m_x$$

$$H_a: m_i \neq m_x$$

Bajo la $H_0 \Rightarrow$

$$\bar{X}_i \in N \left(m_x, S_i^2 = \frac{\sigma_i^2}{n_i} \left(\frac{n - n_i}{n - 1} \right) \right)$$

Es decir, bajo la hipótesis nula tenemos que la media muestral de la variable en el grupo i sigue también una distribución Normal, con media igual a la media poblacional de esa variable, y una varianza igual a S_i^2 , una varianza que obtendríamos en el caso de una extracción sin reemplazamiento de los n elementos o barrios.

Podemos encontrar una región crítica que nos permita rechazar o no la hipótesis nula. Es decir, si θ^* es un estimador de θ con una distribución $N(0,1)$, sabemos que

$$\Pr \{ -t_{\alpha/2} \leq \theta^* \leq t_{\alpha/2} \} = 1 - \alpha$$

Donde α es el nivel de significación y $t_{\alpha/2}$ es el cuantil $\alpha/2$ de la distribución Normal Simple. Por tanto, con un nivel de significación del 5%,

$$\Pr \{ -t_{0,025} \leq \theta^* \leq t_{0,025} \} = 1 - 0,05 = 0,95$$

$$\Pr \{ -1,96 \leq \theta^* \leq 1,96 \} = 0,95$$

Es decir, la región crítica será: $RC = (-\infty, -1,96) \cup (1,96, \infty)$, y por tanto, si el valor del estadístico cae en la región crítica, se rechazará la hipótesis nula a un nivel de significación del 5%.

En nuestro caso, en lugar del estimador θ^* , bajo la hipótesis nula tenemos el siguiente estadístico Z_i :

$$Z_i = \frac{\bar{X}_i - m_x}{\sqrt{\frac{\sigma_i^2}{n_i} \left(\frac{n - n_i}{n - 1} \right)}}$$

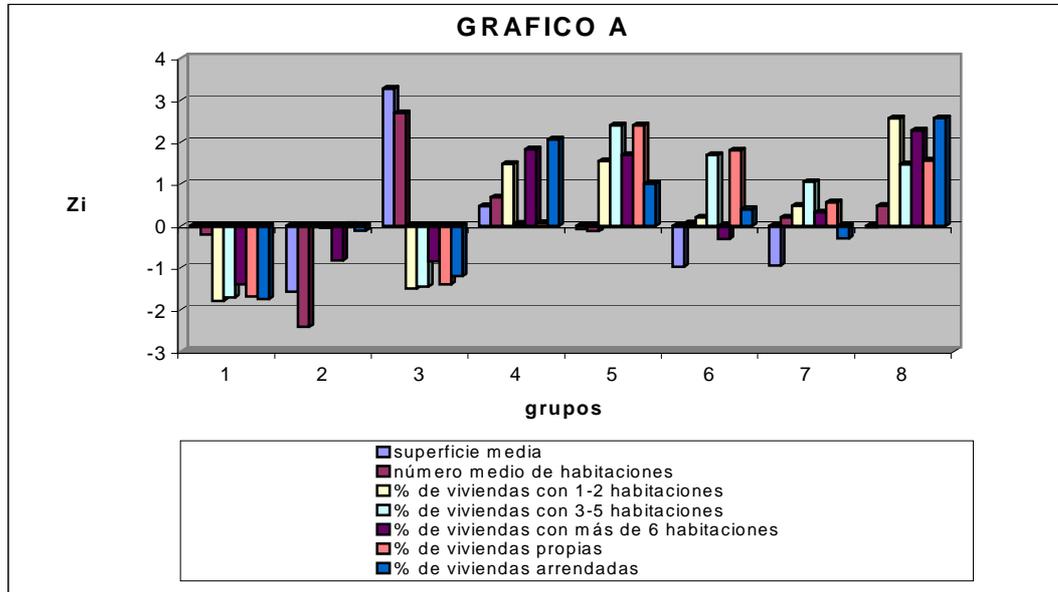
que también sigue una distribución $N(0, 1)$ y por consiguiente, con un nivel de significación del 5%, y procediendo de la misma manera, obtendremos la misma región crítica. De modo que si el valor del estadístico Z_i , en valor absoluto, supera el valor 1,96, consideraremos que esa variable es significativa y que es una variable a tener en cuenta al caracterizar ese grupo, ya que rechazaremos la hipótesis nula. En el caso contrario, es decir si no rechazamos la hipótesis nula, esa variable en media no es significativa ya que en ese grupo sigue el comportamiento medio de la variable en toda la población. De esta forma, discriminaremos las variables e identificaremos cuáles recogen realmente las principales características de la zona.

El siguiente paso será identificar las variables que caracterizan esos grupos, y de esta forma poder definir el tipo de vivienda y habitantes característicos de cada uno de los grupos. Para ello, procedemos a calcular el valor del estadístico Z_i descrito anteriormente y a rescatar aquellos cuyos valores se acerquen en valor absoluto a 1,96.

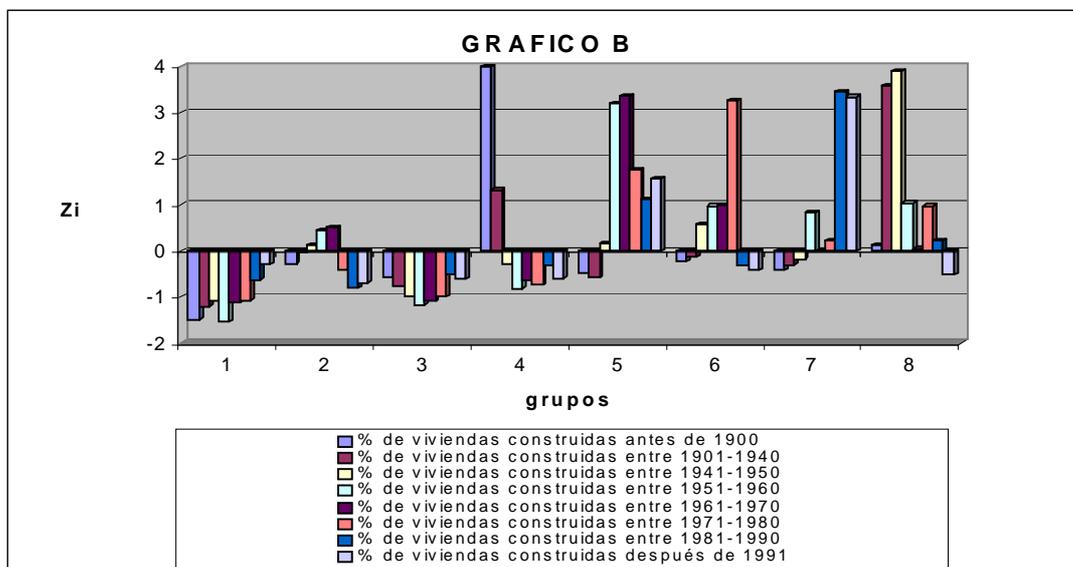
Hemos podido observar, que de las 37 variables sobre todo destacan unas pocas a la hora de caracterizar los grupos, es decir, destacan el año de construcción, tamaño de la vivienda y número de habitantes. Por ello, en los tres gráficos siguientes hemos recogido los valores de los estadísticos Z de estas variables para los ocho grupos, con el fin de ver las similitudes y diferencias entre ellos.

Por ejemplo, en el gráfico A se recogen las variables superficie media de las viviendas, número medio de habitaciones, % de viviendas con 1-2 habitaciones, % de viviendas con 3-5 habitaciones, % de viviendas con 6 o más habitaciones, % de viviendas propias y arrendadas, para los ocho grupos. Se recogen los valores de los

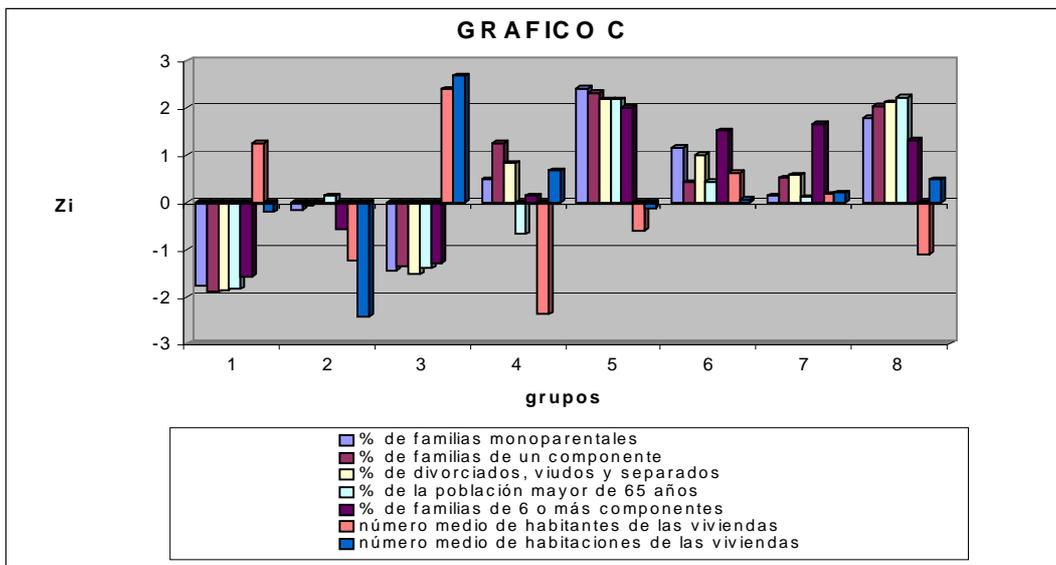
estadísticos Z de las variables que describen, sobre todo, el tamaño de la vivienda, así como el régimen de posesión de la misma. Podemos observar que considerando estas variables en conjunto, los grupos no se comportan de forma similar, ya que los gráficos correspondientes a cada grupo son bastante diferentes.



Siguiendo con las características de las viviendas, el gráfico B recoge los valores de los estadísticos Z de las variables que recogen las fechas de construcción, considerando también los ocho en un mismo gráfico. Como podemos observar en el gráfico, parece que existe cierta similitud entre los grupos 1 y 3, mientras que el comportamiento del resto de los grupos parece totalmente diferente. Lo estudiaremos con más detalle más adelante.

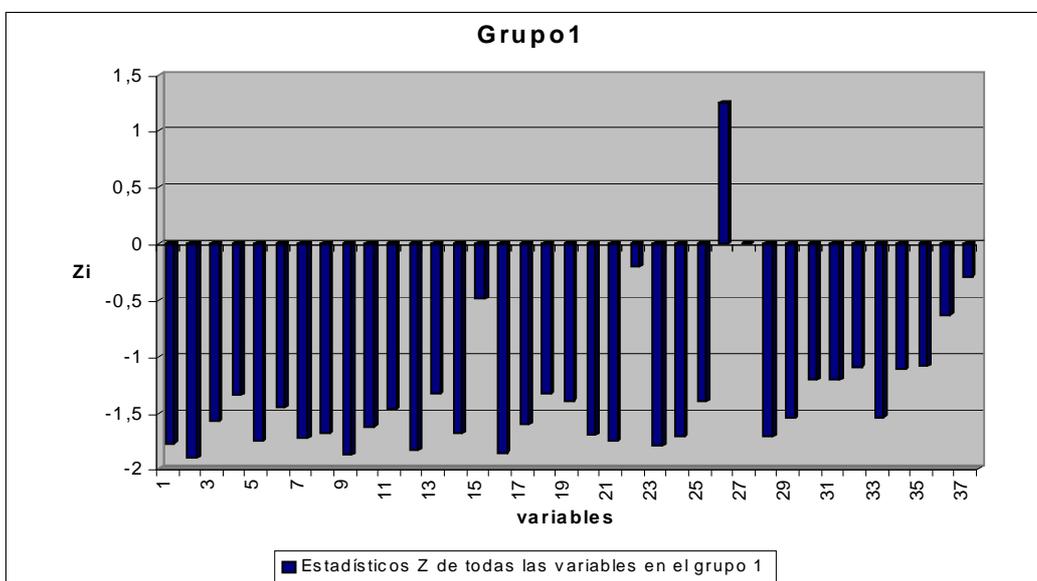


En el gráfico C, en cambio, se recogen las características de los habitantes de las viviendas de los barrios agrupados en ocho áreas, así como la densidad de ocupación de la vivienda. Es decir, variables como número medio de habitantes, % de familias de 1 componente, % de familias de 6 o más componentes, número medio de habitaciones, porcentaje de divorciados, porcentaje de mayores de 65 años y % de familias monoparentales. Como podemos observar, parece que el grupo 1 y el grupo 3 tienen cierta similitud en algunos de los indicadores, pero el resto de los grupos difieren totalmente en ellos.



A continuación procederemos a describir cada uno de los grupos por separado, intentando buscar las variables significativas en cada caso.

GRUPO 1: HERRERA, MARTUTENE, ULIA, AIETE, IBAETA, AÑORGA, BIDEBIETA



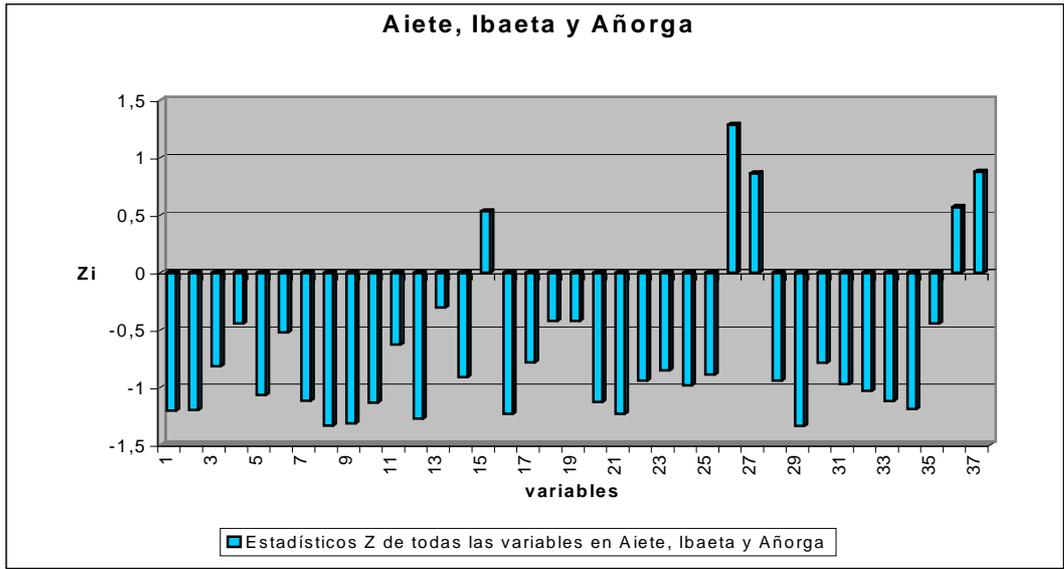
Este grupo es el que tiene más componentes y quizás por ello sea tan difícil encontrar variables que identifiquen este grupo.

Lo que sí podemos destacar es que casi todos los estadísticos Z son negativos, por lo que las medias de las variables en este grupo son inferiores a la media global. La excepción es la variable X26 (número medio habitantes), ya que es la única variable que en esta agrupación toma un valor positivo, aunque según nuestro contraste, tampoco podríamos considerarla significativa.

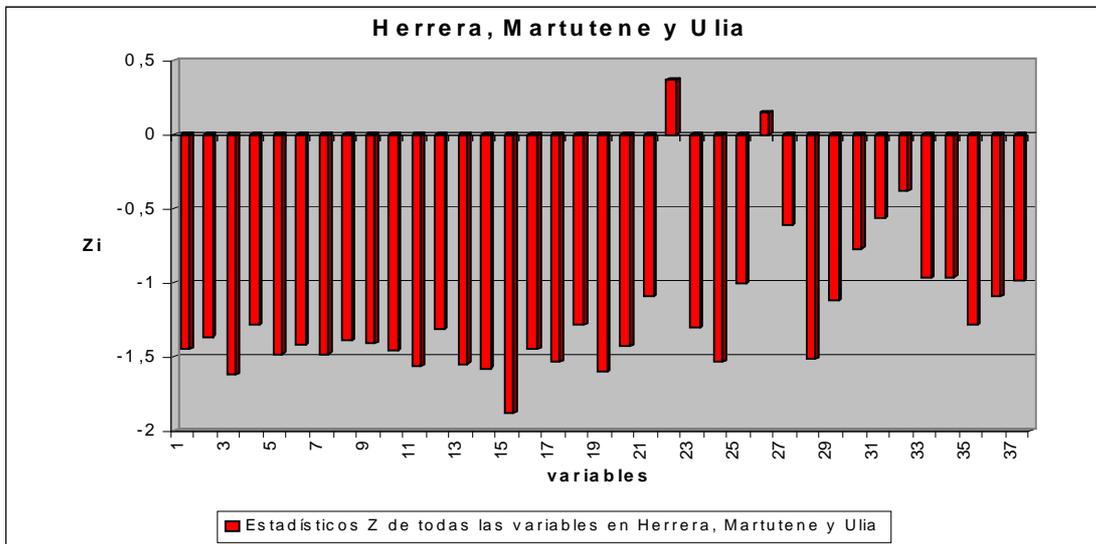
Como no podemos encontrar variables representativas del grupo, y como es tan difícil describirlo en conjunto, hemos intentado analizarlo desglosándolo en tres subgrupos para ver si así podemos rescatar las variables principales y para ver si así podemos estudiar las principales características de las viviendas de estas zonas de Donostia y sus habitantes. Es decir, tal y como hemos mencionado anteriormente, a la hora de hacer un corte en las distancias, optamos por una distancia menor obteniendo un número mayor de grupos, y por consiguiente, en este caso el primero se desglosa en tres grupos.

Observando el árbol de agrupación, podríamos intentar describir este grupo estudiando los siguientes tres subgrupos, Herrera, Martutene y Ulia por una parte, Aiete, Ibaeta y Añorga por otra, y por último Bidebieta.

Desglosando el grupo y calculando de nuevo los estadísticos, tampoco hallamos variables identificativas de los mismos, aunque sí podemos observar algunos cambios. Por ejemplo, en el subgrupo, Aiete, Ibaeta y Añorga, la media de las viviendas construidas a partir de 1981 es superior a la media global, mientras que no ocurre lo mismo en los demás grupos. Además, también podemos observar, que en estos tres barrios la media de la variable superficie media es superior a la media global, pero la media del número de habitaciones, aunque cercano a la media global, sigue siendo inferior. Mencionar que en estos tres barrios, así como en el resto de los barrios de este grupo, la media del número de habitantes es ligeramente superior a la media global. Por tanto, podríamos decir que en estos tres barrios se han construido muchas viviendas desde 1980, que son viviendas grandes y que no están aglomeradas.

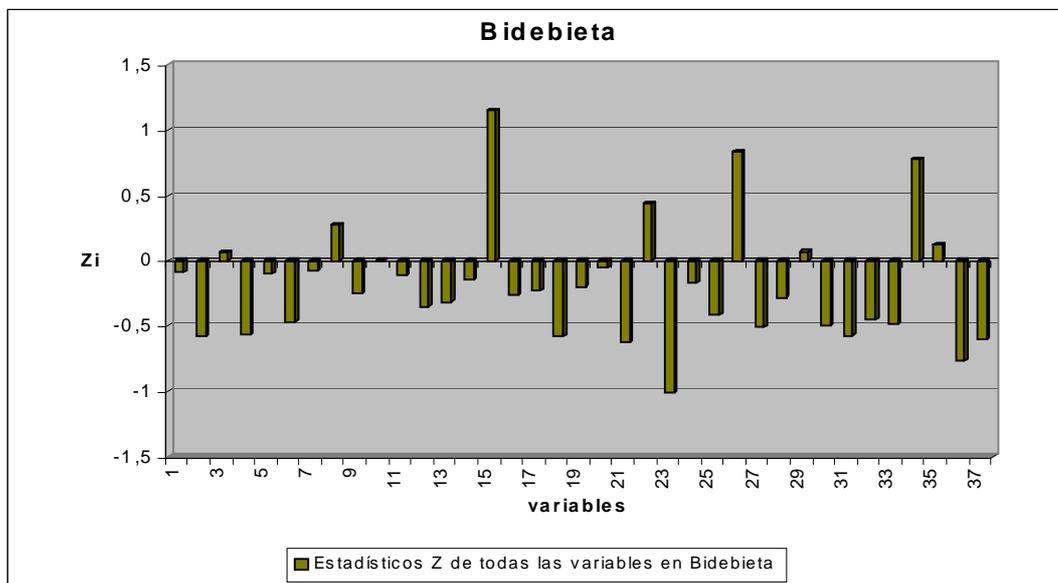


En cuanto al subgrupo Herrera, Martutene y Uliia, no podemos destacar ninguna variable, pero comentar que son viviendas con superficie en media inferior que la media global, pero que el número medio de habitaciones en media es superior, así como el número medio de habitantes. Parece que en estos barrios las viviendas son más pequeñas que en los barrios del otro subgrupo y además están más aglomeradas.

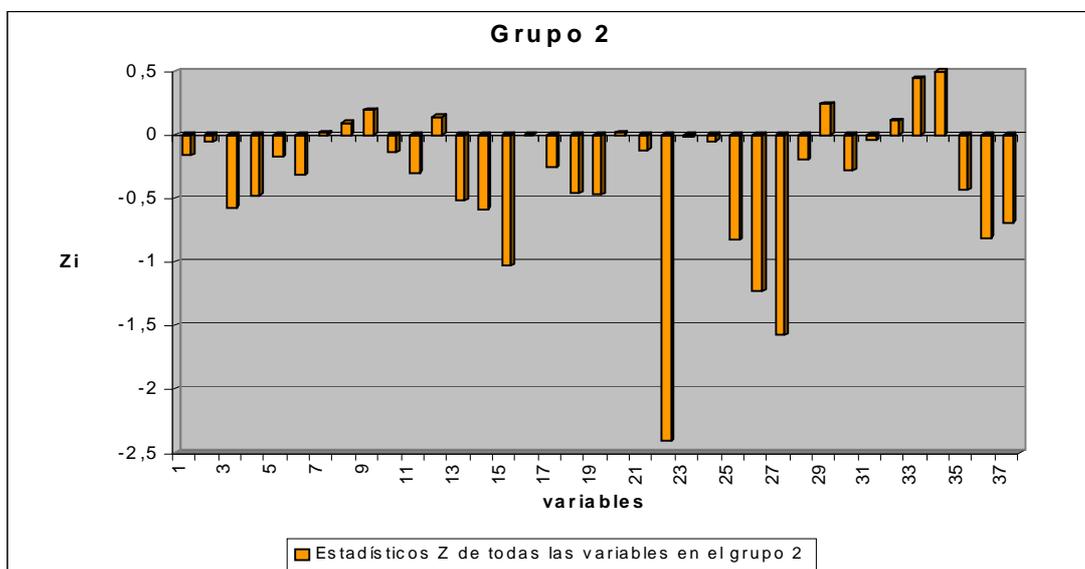


En Bidebieta, a diferencia del primer subgrupo, la media de las viviendas construidas entre 1961 y 1980 es superior a la media global, y en media tienen una superficie inferior, pero el número medio de habitaciones y de habitantes es superior al número medio global. Además, si tenemos en cuenta que también es superior la

media de las familias de 6 o más componentes, podríamos decir que al igual que las viviendas del subgrupo anterior, son viviendas con gran densidad de habitantes.



GRUPO 2: AMARA VIEJO, LOYOLA, EL ANTIGUO, EGIA



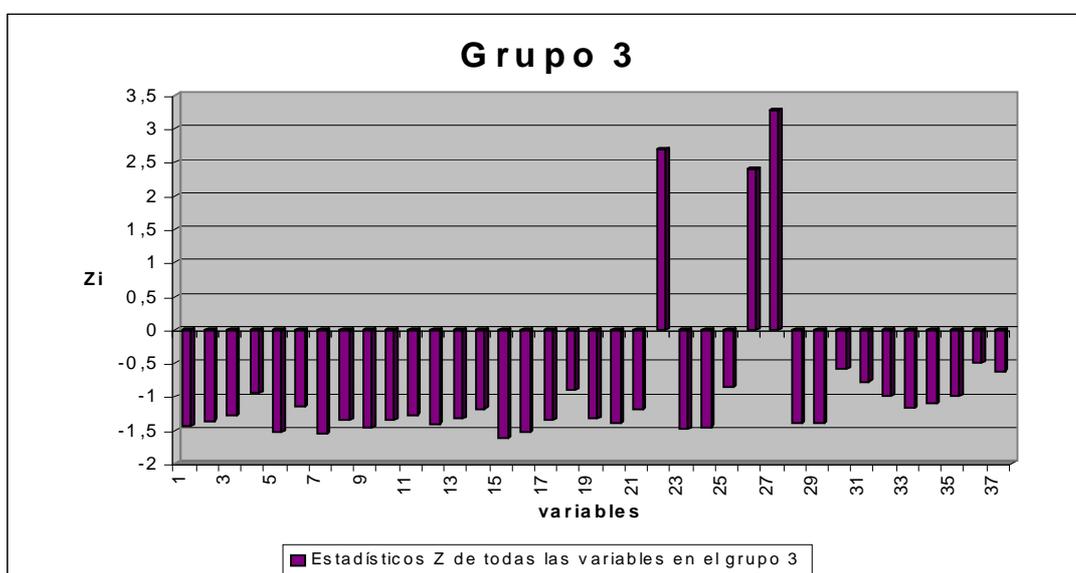
En este grupo cabe destacar que la mayoría de los estadísticos de las variables son cercanos a cero, por lo que en este grupo el comportamiento de la mayoría de las variables es el comportamiento medio de esas variables en toda la población. El único indicador en valor absoluto superior a 1,96 es el número medio de habitaciones, pero también son de destacar la superficie media y el número medio

de habitantes. Todos ellos toman valores negativos por lo que la media de dichas variables en este grupo es inferior al global.

Por tanto, podríamos definir este grupo como zona en la que predominan viviendas de poca superficie y de pocas habitaciones, y que no están aglomeradas, en el sentido de que en ellas habitan pocas personas.

De todas formas, hay que mencionar que el barrio El Antiguo ha variado mucho desde 1996, ya que se han construido muchas viviendas, entre ellas viviendas de protección oficial, y puede que incorporando esta información, El Antiguo no estuviese en este grupo.

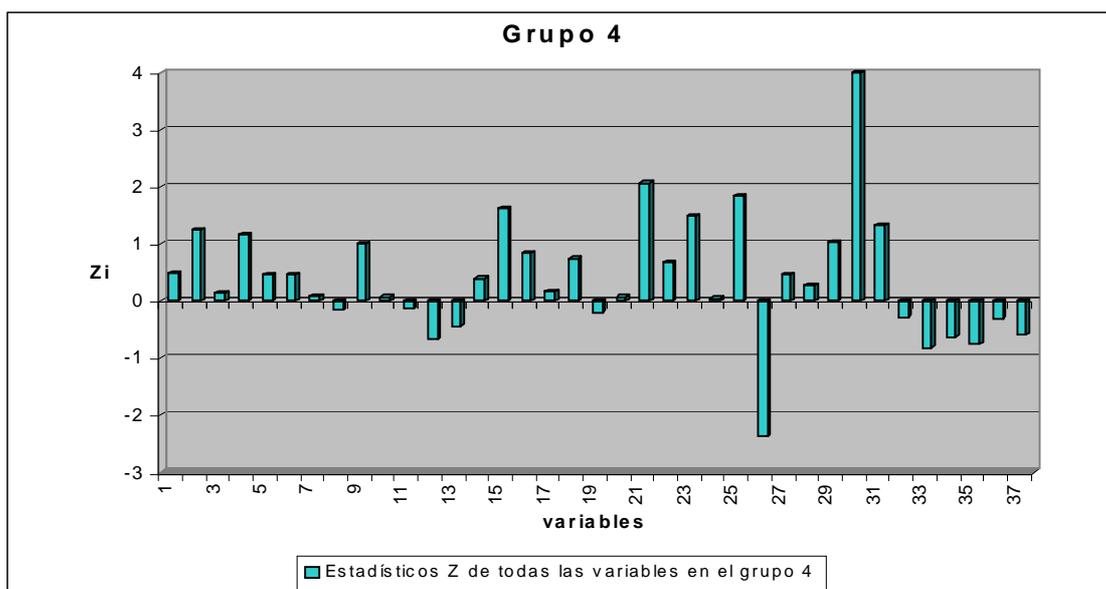
GRUPO 3: IGELDO, MIRACONCHA



Al contrario que en el grupo dos, en este grupo cabe destacar que son viviendas grandes y de muchas habitaciones, pero las familias no son de un componente ni de más de seis, familias de tamaño intermedio pero superior al tamaño medio.

En realidad podríamos esperar este resultado, ya que estas dos zonas siguen siendo un lugar privilegiado en la ciudad. Son barrios cercanos al centro y a la playa, y en general, con grandes vistas. Siguen siendo barrios más bien para familias con disposición económica alta, y las viviendas principalmente responden a las necesidades de estas familias.

GRUPO 4: *CENTRO, PARTE VIEJA*



En este grupo tenemos los barrios Centro y Parte Vieja. En estos barrios están ubicadas viviendas construidas antes de 1900 y a principios del siglo y que son viviendas o bien pequeñas (1-2 habitaciones) o bien grandes (6 o más habitaciones) pero principalmente son viviendas arrendadas.

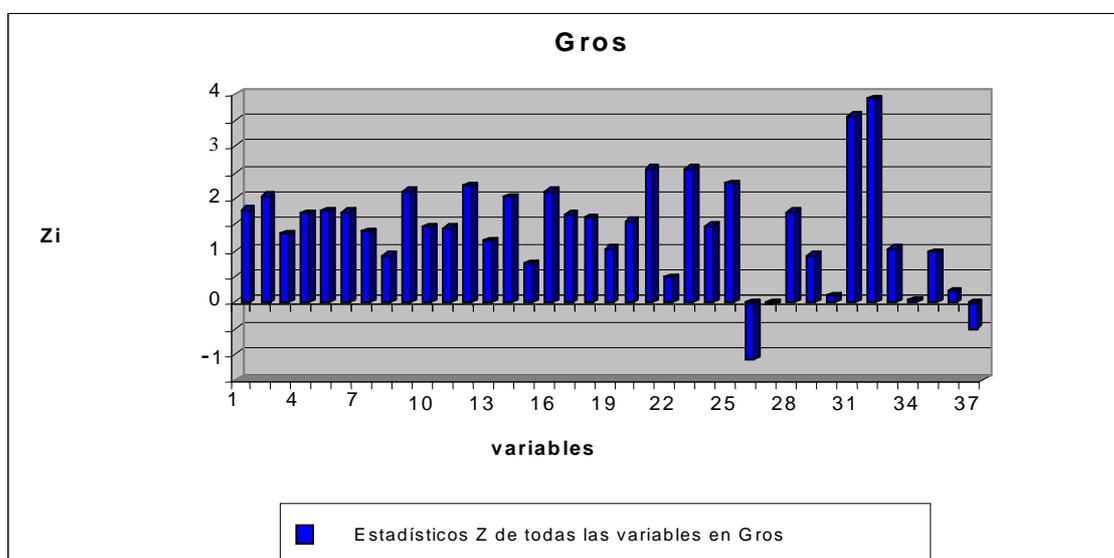
Ya lo comentábamos en la introducción, cuando decíamos que la expansión urbanística de Donostia comenzó sobre todo en el Centro y posteriormente se pasó a la periferia de la ciudad. Está claro que en estos dos barrios encontraremos sobre todo viviendas antiguas y grandes, ya que en un primer instante se instaló aquí la burguesía. Pero hoy en día adquieren también otra connotación, pues sobre todo son viviendas arrendadas, y podemos pensar que ya no tienen el carácter de lugar de veraneo para la clase de alto standing.

Hay que observar además, que destaca el indicador número medio de habitantes, pero tomando un valor negativo, es decir que los barrios de este grupo están inversamente relacionados con esa variable, por lo que podríamos decir que en estas viviendas, grandes o pequeñas, viven pocos habitantes en cada una. Destacar también la variable familia de un componente, pero que ésta toma un valor positivo y bastante elevado, que podría corroborar la afirmación anterior.

GRUPO 5 y GRUPO 6: GROS Y AMARA BERRI

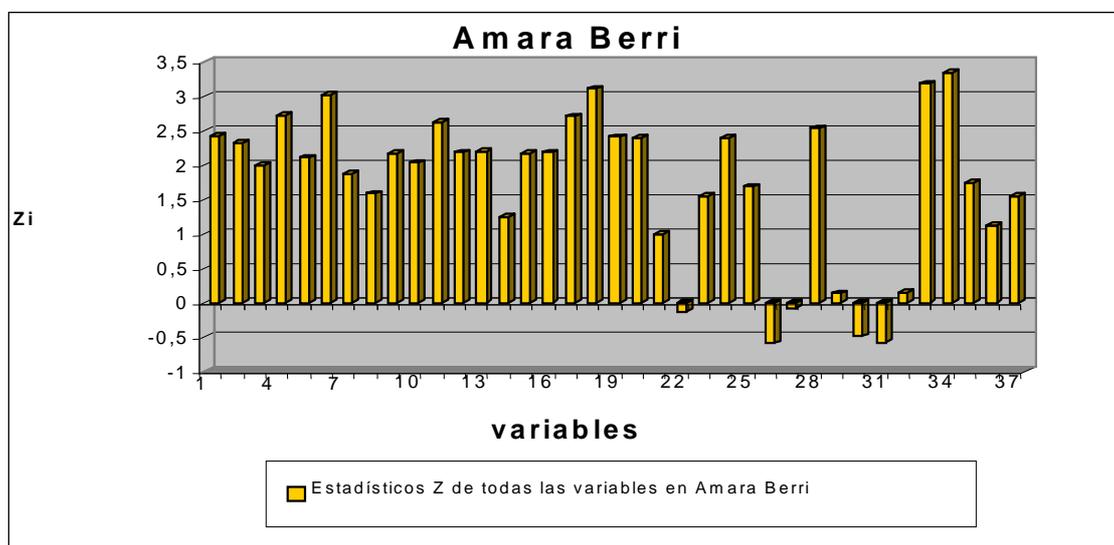
Los barrios Amara Berri y Gros se parecen en el sentido de que, en general, las personas residentes en estos barrios son personas mayores de 65 años, también residen personas con un trabajo de gran responsabilidad, es decir empresarios, directores, con profesiones superiores e intermedias, personas divorciadas, viudas y separadas y muchas familias monoparentales.

Aunque podríamos diferenciar dos tipos de familias diferentes, en el sentido de que en Gros residen muchas personas mayores de 65 años y jubilados, bien en viviendas pequeñas (de 1-2 habitaciones) y bien en viviendas de 6 o más habitaciones, pero si observamos el indicador que recoge el número medio de habitantes de estas viviendas, como toma un valor negativo, podríamos sospechar que son personas que con el paso del tiempo se han ido quedando solos en esas viviendas grandes o incluso que al quedarse solos, se han trasladado a viviendas más pequeñas. Puede ser que algunas viviendas pequeñas estén ocupadas por los miembros de las familias rotas que se han quedado solas, es decir, los divorciados.



A diferencia de Gros, en Amara Berri, las viviendas son de 3-5 habitaciones construidas a posteriori que en Gros, ya que en el primero destacan las variables X33 y X34, mientras que en el segundo destacan los indicadores X31 y X32. Además predominan las viviendas propias. Podemos decir que la tipología de las familias de Amara Berri se diferencia bastante si la comparamos con las de Gros, ya que en el primero, el porcentaje de personas con estudios medios y universitarios es muy

elevado y residen tanto familias grandes como pequeñas, quizás familias recientemente formadas y familias formadas desde hace tiempo, ya que además podemos destacar el gran porcentaje de asistencias a guarderías, así como personas menores de 14 años.



GRUPO 7 y GRUPO 8: ALZA E INTXAURRONDO

En cuanto a estos dos barrios, aunque los gráficos que recogen los estadísticos Z son diferentes, podemos encontrar algunas similitudes, por ejemplo, en cuanto al tipo de vivienda, vemos que tanto en Alza como en Intxaurreondo la superficie media de la vivienda y el número medio de habitaciones, sin considerarlas variables significativas, toman valores inferiores a la media. Es más, la media de la variable 6 o más habitaciones es claramente inferior a la media global, pero sin embargo el número de viviendas de 1-2 habitaciones y de 3-5 habitaciones es superior. Parece que en estos dos barrios las viviendas no son muy grandes, ni por la superficie ni considerando el número de habitaciones. Cabe mencionar que en Alza la media de viviendas en alquiler es superior a la media global, pero en Intxaurreondo no lo es. Por otra parte, la media de viviendas propias está muy por encima de la media global en Alza y en Intxaurreondo en cambio, aun siendo también superior, tampoco podemos considerarla como una variable significativa.

Siguiendo con la tipología de las viviendas, lo que sí es diferente es el periodo de construcción, ya que en Alza destaca el periodo comprendido entre 1971

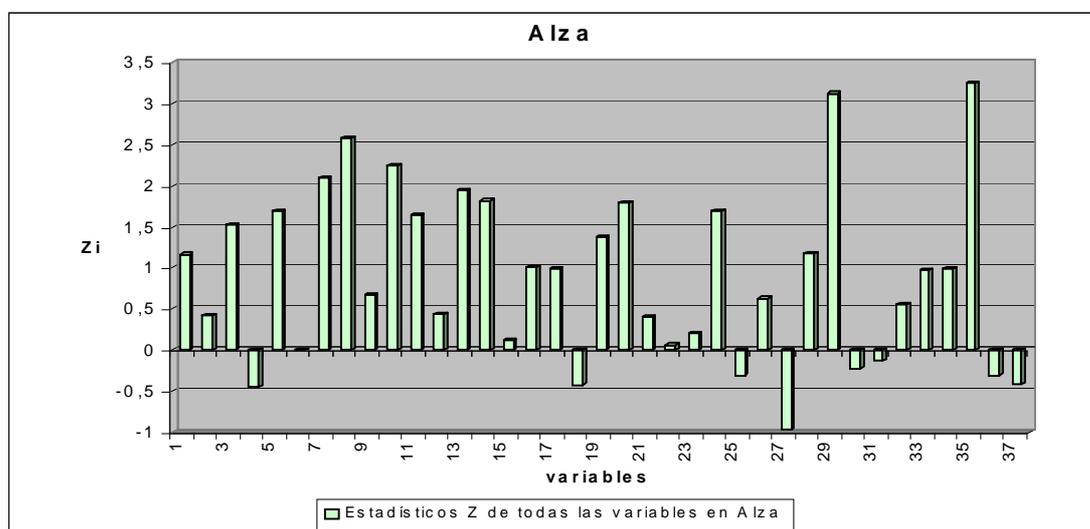
y 1980, mientras que en el periodo que destaca el barrio de Intxaurreondo, desde 1981 hasta hoy en día, Alza tiene una media inferior a la global.

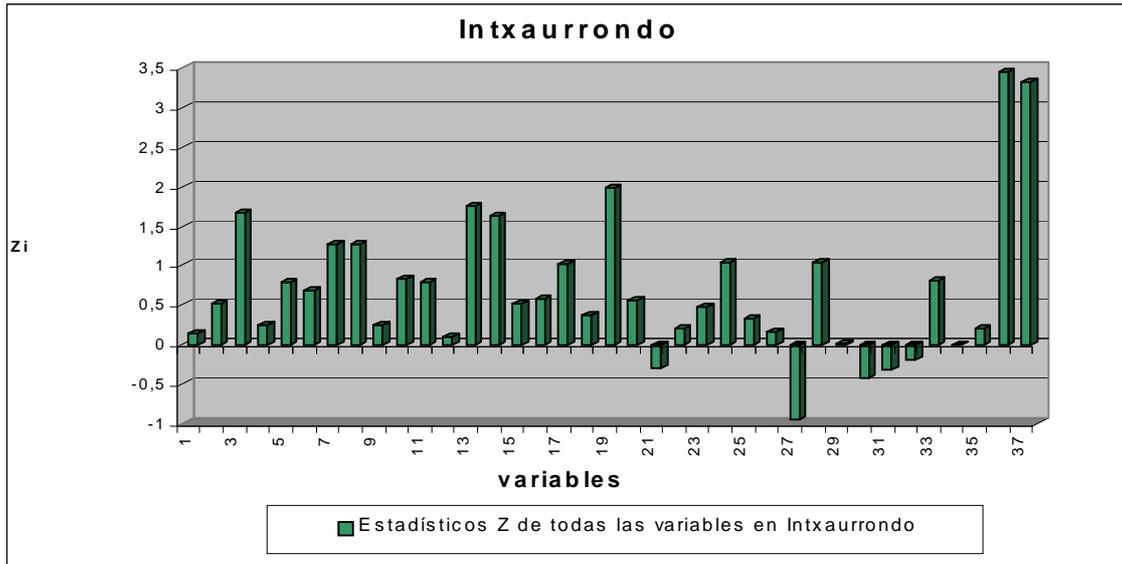
Por tanto, en Intxaurreondo y en Alza tenemos viviendas de reciente construcción (en Alza un poco más antiguas) y en régimen de propiedad, y en general, son viviendas de 3-5 habitaciones y de superficie algo menor que la media. Además, tenemos que mencionar que Intxaurreondo ha cambiado mucho en los últimos años, ya que se ha construido un centro comercial y muchas viviendas que podrían cambiar las características principales de este barrio, y por tanto la asignación del mismo a un grupo o a otro, ya que este estudio lo hemos realizado con datos hasta 1996, y la construcción se ha disparado en este barrio a partir de esta fecha.

En cuanto a los habitantes de estos dos barrios, vemos que no son familias de pocos componentes, ya que las variables X2, X1, X16 y X12, es decir, porcentaje de familias de un componente, porcentaje de familias monoparentales, porcentaje de divorciados, viudos y separados y porcentaje de la población mayor de 65 años, mantienen el comportamiento medio, mientras que destaca la variable X3, la variable de porcentaje de familias de 6 o más componentes. Así mismo, podemos destacar la variable que recoge el porcentaje de la población menor de 14 años en los dos barrios y el porcentaje de asistencia a guarderías en el barrio de Intxaurreondo.

Analizando las demás variables, destacan en Alza las variables porcentaje de amas de casa y buscando su primer empleo. Además, también es significativa la variable que recoge el porcentaje de empleados y trabajadores, así mismo, podemos considerar como variable destacable, la variable porcentaje de autónomos.

Los dos gráficos siguientes recogen los valores del estadístico Z de todas las variables en estos dos barrios.





5. ANALISIS FUZZY

En el estudio realizado hasta ahora, intentábamos buscar grupos de modo que formasen una partición tal que el grado de asociación fuese alto dentro del grupo y bajo entre individuos de diferentes grupos. Nuestra tarea ha consistido en dividir los 19 barrios de San Sebastián, caracterizados por 37 indicadores, en grupos o clusters homogéneos. Los barrios de un grupo son similares y los barrios de diferentes clusters, en cambio, son lo más desiguales posibles, donde además el número de grupos suele ser normalmente desconocido. Es decir, hemos realizado un Análisis Cluster clásico.

En el Análisis Cluster realizado, hemos utilizado la idea de que un barrio pertenece o no pertenece a un grupo, de forma que lo incorporábamos al grupo y posteriormente realizábamos un estudio para ver qué indicadores caracterizaban esos grupos. Pero, se podría hacer también un estudio con mayor flexibilidad, generalizando la idea de pertenencia o no pertenencia del barrio al grupo. Es decir, permitiendo diversos grados de asociación de un barrio a varios grupos.

Ahora la idea no es pertenecer o no a un grupo, sino que un individuo o barrio tiene un grado de compatibilidad con un grupo y también otro grado de compatibilidad con otros grupos diferentes, y por tanto podría incorporarse a uno o a otro. Con estos grados de compatibilidad, intentaremos hacer un estudio similar al que hicimos en los apartados anteriores, es decir, basándonos en estos grados de

compatibilidad intentaremos ver si existen de nuevo zonas homogéneas o grupos de barrios homogéneos de la ciudad. Sobre todo queremos ver si añadiendo flexibilidad cambian muchos los resultados que conseguimos con el análisis clásico. Es lo que se conoce como FUZZY CLUSTERING.

Entre los métodos de agrupamiento difuso, podemos encontrar el método c-mean y el método de agrupamiento basado en relaciones de equivalencia. Para hacer el estudio propuesto, vamos a optar por el segundo método, ya que en el primero hay que especificar el número de grupos finales, mientras que en el segundo método no.

La idea del método basado en relaciones de equivalencia, es buscar relaciones de compatibilidad sobre el conjunto de datos, relaciones que son reflexivas y simétricas. Por ejemplo podemos utilizar la siguiente fórmula general:

$$R(x_i, x_k) = 1 - \delta \left[\sum_{j=1}^p |x_{ij} - x_{kj}|^q \right]^{1/q}$$

Donde i y j son los diferentes barrios de San Sebastián, p recoge el número de indicadores (en nuestro estudio son 37 en total), δ es la inversa de la mayor distancia entre los barrios y $q \in \mathbb{R}_+$.

Al igual que en el Análisis Cluster, hemos optado por tomar la distancia euclídea para calcular la distancia entre dos barrios, es decir $q=2$. Hemos calculado las distancias entre todos los pares de barrios, y por tanto, hemos podido calcular también el valor de δ , es decir, la inversa de la mayor de ellas, en nuestro caso $1/18,36 = 0,05$. A continuación, hemos calculado la matriz R , una matriz de orden 19×19 , en la que se recogen los grados de compatibilidad entre todos los barrios.

A partir de esta matriz, hemos calculado las relaciones de equivalencia, que a diferencia de las relaciones de compatibilidad anteriores, tienen la propiedad de que son reflexivas, simétricas y además transitivas. Son una generalización de las relaciones de compatibilidad calculadas con la fórmula anterior. Para tal fin, hemos utilizado el siguiente teorema:

Teorema: Sea R una relación de compatibilidad fuzzy en un set universal X finito con $|X|=n$. Entonces, la aproximación max-min transitiva de R es la relación $R_{(n-1)}$.

El teorema sugiere calcular la aproximación transitiva $R_T = R^{(n-1)}$, calculando la siguiente secuencia de relaciones:

$$R^{(2)} = R \circ R$$

$$R^{(4)} = R^{(2)} \circ R^{(2)}$$

$$R^{(8)} = R^{(4)} \circ R^{(4)}$$

$$R^{(2^k)} = R^{(2^{(k-1)})} \circ R^{(2^{(k-1)})}$$

Hasta que no se produzca una nueva relación o hasta que $2^k \geq n-1$. La composición max-min se calcula de la siguiente forma:

Sean $R_1(x,y)$ y $R_2(y,z)$ $(x,y) \in X \times Y$, $(y,z) \in Y \times Z$, dos relaciones fuzzy de compatibilidad.

$$\text{Entonces, } R_1 \circ R_2 = \{(x,z), \max_y \{ \min \{ \mu_{R_1}(x,y), \mu_{R_2}(y,z) \} \} / x \in X, y \in Y, z \in Z \}$$

Y donde $\mu_{R_1 \circ R_2}(x,y) = \max_{y \in Y} \{ \min \{ \mu_{R_1}(x,y), \mu_{R_2}(y,z) \} \} / \forall x \in X, \forall z \in Z$ es la función de equivalencia max-min transitiva.

Así, siguiendo el teorema anterior, podemos calcular la matriz $R^{(2^k)}$ de las relaciones de equivalencia. En nuestro caso, como el número de barrios es 19, tendremos que calcular las matrices $R^{(2)}$, $R^{(4)}$, $R^{(8)}$, $R^{(16)}$ y $R^{(32)}$. La matriz $R^{(32)}$ por tanto, recogerá las relaciones de equivalencia entre todos los barrios.

Una vez calculada la matriz de relaciones de equivalencias, tendremos que hacer cortes a diferentes niveles para estudiar los diferentes grupos que se van creando, y para poder ir formando grupos y completar también un árbol al igual que lo hicimos en el Análisis Cluster Clásico. Este árbol lo hemos recogido en el gráfico 2 adjunto a continuación, donde los valores que aparecen en el gráfico, recogen las relaciones de equivalencia entre dos barrios o grupos, y por tanto, mayor valor supone mayor relación y por tanto, antes se agruparán.

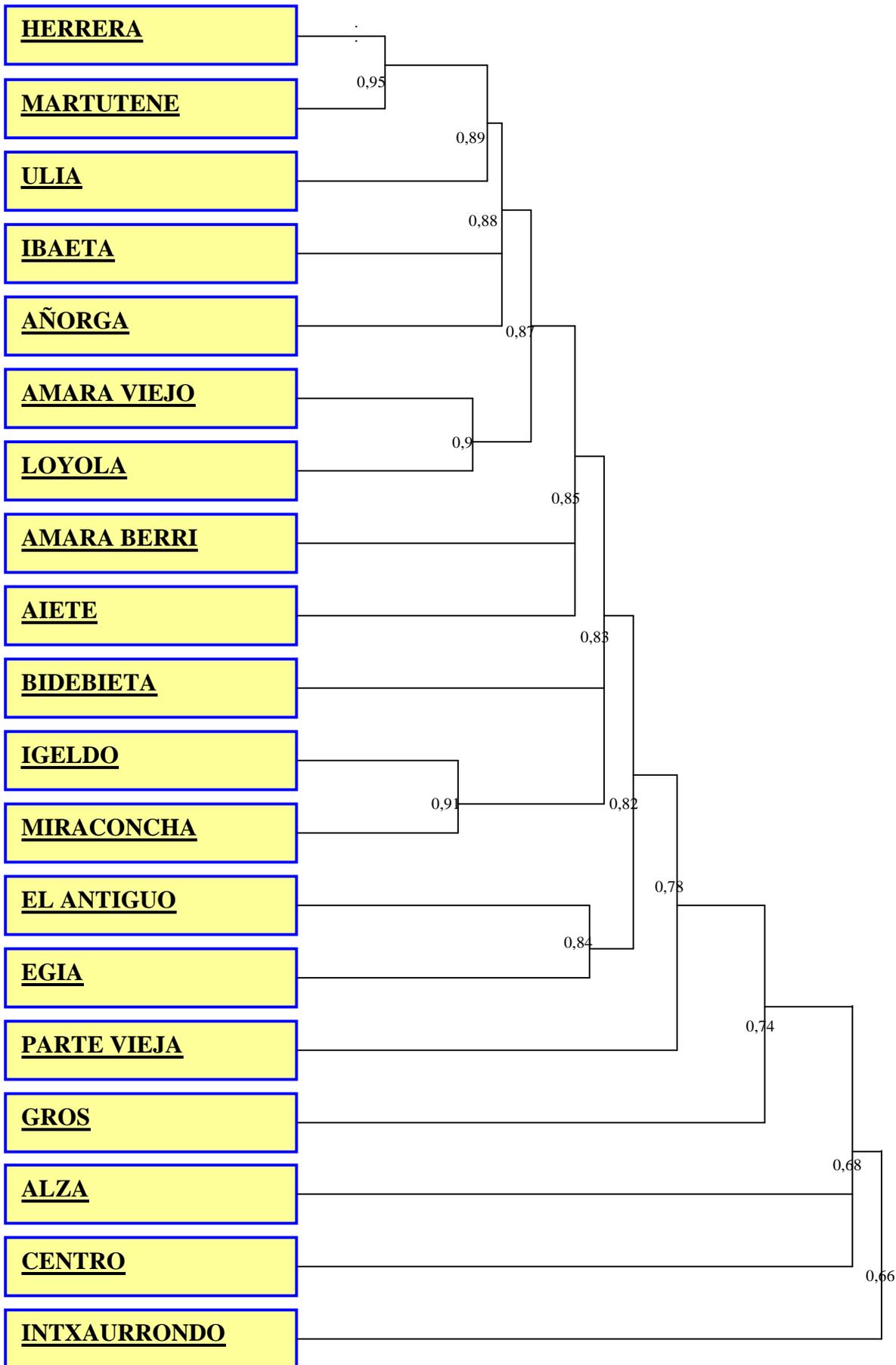


GRAFICO 2

Observando este gráfico, tenemos que mencionar que ya no están tan claros los grupos de barrios, aunque también podemos observar algunas de las similitudes que veíamos en el Análisis Cluster Clásico realizado en los apartados anteriores. Con el Análisis Fuzzy diferenciaríamos 4 grupos, aunque el primero vuelve a ser bastante amplio y podemos encontrar subgrupos de barrios que hay que considerar.

En un primer grupo incorporamos los barrios Herrera, Martutene, Ulia, Ibaeta, Añorga, Amara Viejo, Loyola, Amara Berri, Aiete y Bidebieta. Como podemos observar, salvo Amara Viejo, Loyola y Amara Berri, los demás barrios los incorporábamos al primer grupo que formamos en el Análisis Cluster Clásico, y se mantiene de nuevo la relación entre Herrera y Martutene, aunque ya no la de Aiete e Ibaeta. Amara Viejo y Loyola en el análisis anterior, los incorporábamos al grupo 2 conjunto con Antiguo y Egia, pero ahora, aunque mantienen la similitud que ya tenían en el estudio anterior, se diferencian de los barrios de Antiguo y Egia, que vuelven a mantener la relación entre ellas, pero que forman un grupo diferente al primero, es decir forman el grupo 3. El grupo 2, en cambio, vuelve a estar compuesto por los barrios Igeldo y Miracocha, que ya formaban también un grupo anteriormente.

Igual que en Análisis Clásico, al final del árbol nos quedan individuos o barrios que se van incorporando poco a poco al grupo que forman todos los demás barrios. En este caso, podemos encontrarnos a Parte Vieja, Gros, Alza, Centro e Intxaurre. Las novedades con respecto al estudio anterior son Parte Vieja y Centro, que en el estudio anterior los dos barrios formaban un grupo y que en éste, en cambio, se van incorporando al grupo general sin mantener la relación tan clara que se veía antes.

Como podemos observar, añadiendo cierta flexibilidad a la hora de formar los grupos, los resultados cambian un poco, pero lo importante es que vuelven a aparecer las relaciones entre barrios que comentábamos sin esa flexibilidad y por tanto también las características que resaltábamos en el estudio anterior.

6. BIBLIOGRAFIA

- C.M.Cuadras. *Métodos de Análisis Multivariante*. Editorial Universitaria Barcelona. 1981
- B.Escofier, J.Pages. *Análisis Factoriales Simples y Múltiples*. UPV-EHU. 1992.
- G.A.F.Seber. *Multivariate observations*. Wiley Series in probability and Mathematical Statistics. 1984
- W.R.Dillon, M.Goldstein. *Multivariate Analysis. Methods and Applications*. .Wiley Series in probability and Mathematical Statistics. 1984
- M.R.Anderberg. *Cluster Analysis for applications. Probability and Mathematical Statistics*. Academic Press. 1973.
- A.Kauffmann, J.Gil Aluja, A.Terceño. *Matemática para la economía y la gestión de empresas*. F.C. Ediciones foro Científico. 1994.
- A.Kauffmann, J.Gil Aluja. *Las matemáticas del azar y la incertidumbre*. Editorial centro de estudios Ramón Areces. 1990.
- H.J.Zimmermann. *Fuzzy set and theory and its applications*. Kluwer Academic Publishers. 1991.
- G.J.Klir, Bo Yuan. *Fuzzy sets and Fuzzy logic. Theory and applications*. Prentice Hall PTR. 1995.
- H.Bandemer, S.Gottwald. *Fuzzy Sets, Fuzzy Logic, Fuzzy methods*. Wiley. 1995.
- D.Dubois, H.Prade. *Fuzzy set and systems. Theory and applications*. Academic Press. 1980.
- M.J.Calvo. *Crecimiento y estructura urbana de San Sebastián*. Sociedad Gipuzcoana de ediciones y publicaciones. 1983.
- T.Aluja, A Morineau. *Aprender de los datos: El Análisis de Componentes Principales*. EUB. 1999
- E.Aurelli, B.Baldazzi. *Households and Territory the locations of the population of Rome in relation to housing supply characteristics*. Social Indicators Research, 44: 97-118. Kluwer Academic Publishers, 1998.