Señor…si yo ofendiera a la gente,
dame valor para disculparme

Capítulo 4
- Simulación -
Parte I

4.1 Definición del archivo de simulación

Para probar y evaluar el sistema diseñado, se ha construido un conjunto de archivos de datos a partir de un conjunto de datos privados de financiación para el otorgamiento de créditos al consumo. Este archivo general consta de 6692 individuos y contiene la información requerida para el otorgamiento de financiamiento en función de un conjunto de variables socioeconómicas.

Se seleccionaron de este archivo un conjunto de N0 individuos que representan el archivo de individuos donantes y un conjunto de N1 individuos que representan el archivo de los individuos receptores. La forma de seleccionar estos N0 individuos donantes y N1 individuos receptores se hizo en dos etapas. En la primera etapa de simulación, se escogieron de manera aleatoria un 50% de los individuos para cada conjunto. En la segunda etapa, la selección de los individuos se hizo de acuerdo al cumplimiento de una restricción sobre una variable (estado marital del cliente). En este caso, el número de individuos donantes y receptores no es necesariamente el mismo.

Las variables comunes, también llamadas variables activas, se etiquetaron como: C1, C2, . . . , C13. Todas estas variables son de tipo continuo. Las variables que son empleadas como variables específicas o variables ilustrativas son: Dictamen, Puntos y Financiación. Con las variables Dictamen, Puntos y Financiación recodificadas, se construyó un conjunto de variables categóricas con el fin de tener variables específicas categóricas en las diferentes simulaciones.

Tomando en cuenta las relaciones que existen entre las variables comunes y las variables específicas, se construyó una base de datos para la ejecución del sistema GRAFT, creando de esta manera distintos escenarios.

Se dispone por lo tanto de los valores observados de las variables específicas, que permiten comparar los valores imputados (validación con los valores reales observados).

Tabla 4.1 Descripción del archivo original.

Variable	Tipo	Comunes	Específicas	
			Continuas	Nominales
Dictamen	N		√	√
Puntos	C		√	√
Financiación	C		√	√
C1	C	√		
C2	C	√		
C3	C	√		
C4	C	√		
C5	C	√		
C6	C	√		
C7	C	√		
C8	C	√		
C9	C	√		
C10	C	√		
C11	C	√		
C12	C	√		
C13	C	√		

En la tabla anterior se muestra la estructura de los datos empleada para el desarrollo de las simulaciones.

4.2 Construcción de la base de datos (Simulación I)

El procedimiento para la construcción de la base de datos ha sido el siguiente:

- A partir del archivo original, se han seleccionado de manera aleatoria al 50 % de los individuos que actuarán como individuos donantes y el resto como individuos receptores.

- Con los individuos donantes se ha construido un modelo explicativo (lineal multivariante) a partir de sus coordenadas factoriales[1], para generar los valores de las variables específicas.

[1] el cálculo de las coordenadas factoriales se ha hecho considerando a los individuos receptores como individuos suplementarios

- Usando el modelo construido en el paso anterior, se han generado los valores de las variables específicas en los receptores. Estos valores serán utilizados para validar los resultados de las imputaciones.

- A los valores generados para las variables específicas, se les ha agregado una componente aleatoria ~ N (**0**, **Σ**).

  Σ representa la matriz de varianza-covarianza de los errores de estimación en el modelo. Esta matriz se ha magnificado por 9 y forma lo que en este contexto se llama Base 2 y se ha reducido por 9 dando lugar a la Base 0. Todo esto con la finalidad de tener tres archivos de datos para la simulación (Base0, Base1, Base2) con fluctuación aleatoria creciente.

  Matrices de varianzas-covarianzas residuales para las variables específicas en las tres bases:

Base 0

	Variable 1	Variable 2	Variable 3
Variable 1	0.00234	26.84252	-0.00020
Variable 2	26.84252	2402770.70	-6.43618
Variable 3	-0.00020	-6.43618	0.00038

Base 1

	Variable 1	Variable 2	Variable 3
Variable 1	0.02110	241.58271	-0.00181
Variable 2	241.58271	21624936.00	-57.92564
Variable 3	-0.00181	-57.92564	0.00343

Base 2

	Variable 1	Variable 2	Variable 3
Variable 1	0.18987	2174.24440	-0.01625
Variable 2	2174.24440	194624420.00	-521.33078
Variable 3	-0.01625	-521.33078	0.03085

Se han seleccionado 9 coordenadas factoriales para la construcción del modelo, que representan un porcentaje acumulado de la variabilidad explicada de 90 % como se puede ver en la tabla 4.2

Tabla 4.2 Tabla de valores propios

Número	Valor propio	Porcentaje de la Variabilidad explicada	Porcentaje acumulado
1	2.7872	21.44	21.44
2	2.0468	15.74	37.18
3	1.5048	11.58	48.76
4	1.1204	8.62	57.38
5	1.0883	8.37	65.75
6	0.9614	7.40	73.15
7	0.8977	6.91	80.05
8	0.7712	5.93	85.98
9	0.6201	4.77	90.75
10	0.4442	3.42	94.17
11	0.3698	2.84	97.01
12	0.2806	2.16	99.17
13	0.1075	0.83	100.00

Traza de la matriz: 13.0

Se presentan a continuación algunos resultados de este proceso de construcción de la base de datos.

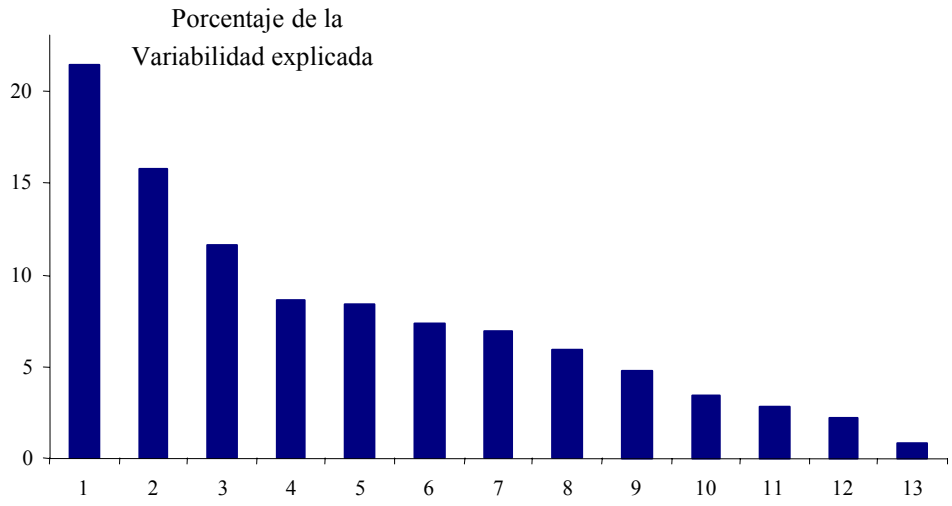


Figura 4.1 Histograma de valores propios

Se desea determinar si las variables comunes tienen un buen poder predictivo de las variables específicas. Para ello se realiza un análisis de componentes principales de las variables comunes proyectando en suplementario las variables específicas.

El siguiente reporte presenta las correlaciones entre las variables específicas y los factores.

Base 0 :

VARIABLES	CORRELATIONS VARIABLE-FACTEUR								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
C2 - Dictamen	0.04	0.52	0.22	0.13	-0.37	0.47	0.40	-0.31	0.02
C3 - Puntos	0.18	0.43	0.19	0.01	-0.01	0.34	0.26	-0.74	0.03
C4 - Financiacion	0.26	-0.62	-0.11	0.39	0.53	-0.10	0.12	0.11	0.10

Se presentan también los Valores-Test para las variables categóricas

MODALITES	VALEURS-TEST								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
4 . Dictamen_d									
AA_1 - C5=1	-2.5	-23.8	-12.9	-4.2	16.0	-18.3	-17.8	14.4	0.4
AA_2 - C5=2	2.5	23.8	12.9	4.2	-16.0	18.3	17.8	-14.4	-0.4
5 . Puntos_d									
AB_1 - C6=1	2.0	-13.9	-2.7	-1.4	1.5	-12.6	-13.8	31.2	-1.2
AB_2 - C6=2	-11.4	-8.7	-8.2	-2.2	1.0	-3.0	0.7	4.9	3.0
AB_3 - C6=3	-2.6	3.4	-1.5	0.8	1.0	1.8	3.7	-5.3	-0.8
AB_4 - C6=4	1.2	6.7	2.2	2.0	-3.2	4.5	3.1	-8.3	-2.1
AB_5 - C6=5	6.1	8.1	7.8	4.5	-0.7	6.7	5.3	-16.7	-1.8
AB_6 - C6=6	14.0	17.8	10.9	-2.5	-1.3	11.1	6.5	-24.7	2.1
6 . Financiacion_d									
AC_1 - C7=1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AC_2 - C7=2	-3.3	16.7	7.7	-15.4	-18.9	2.6	-4.2	-3.5	-6.1
AC_3 - C7=3	-10.4	20.9	1.4	-14.8	-17.1	1.9	-8.1	-2.7	-1.2
AC_4 - C7=4	-1.8	1.2	0.0	5.9	2.1	2.3	7.6	-0.7	0.9
AC_5 - C7=5	9.0	-24.6	-5.3	14.0	21.0	-4.4	1.2	4.3	2.9
AC_6 - C7=6	14.5	-12.3	1.6	0.5	3.2	-3.9	-0.7	1.9	0.3

Se puede ver en el reporte anterior que la capacidad predictiva es intermedia, tal como acontece en muchas situaciones reales. Esto se puede apreciar en los valores Test y las correlaciones Variable-Factor que se muestran.

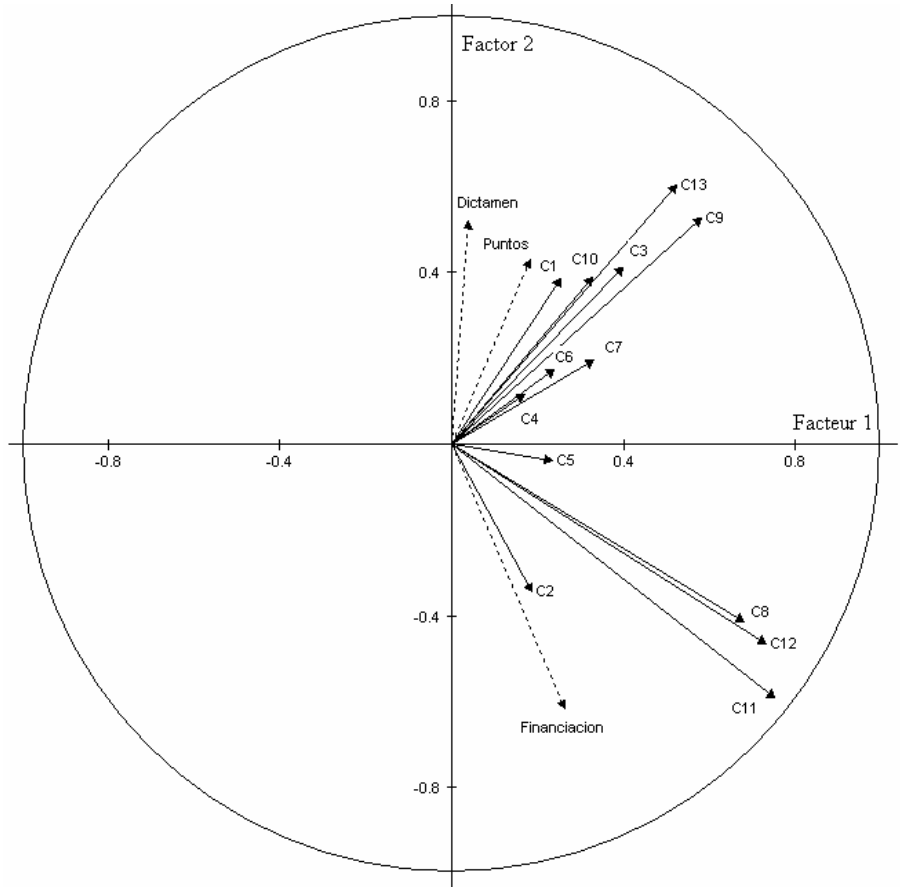
La figura 4.2 de las variables sobre el plano, es la mejor aproximación de los ángulos originales que forman las variables entre sí. Es la mejor representación plana de la matriz de correlaciones entre variables. La contribución de cada variable en la inercia de un eje, es la parte de inercia del eje debida a la variable en cuestión[2]. Indica cuales variables son responsables de la formación de cada eje. La suma de las contribuciones a un eje es igual a 1 (100 en porcentaje).

La posición de una variable ilustrativa en un plano factorial permite visualizar la relación de la variable con el conjunto de variables activas por intermedio de los ejes factoriales. La posición de una variable continua ilustrativa en un plano factorial es asimilable a una regresión visual. No tiene sentido calcular las contribuciones de las variables ilustrativas a la inercia de los ejes, puesto que estas variables no han intervenido en su formación.

Tabla 4.3 Correlaciones de las variables activas con los factores

Variable	Eje 1	Eje 2	Eje 3	Eje 4	Eje 5	Eje 6	Eje 7	Eje 8	Eje 9
C1	0.25	0.38	0.54	0.37	-0.27	-0.07	0.1	-0.25	-0.29
C2	0.18	-0.34	-0.07	0.64	0.54	-0.07	0.29	-0.01	0.12
C3	0.4	0.41	0.53	0.3	-0.16	-0.12	-0.13	0.14	0.00
C4	0.17	0.12	0.27	-0.46	0.03	-0.38	0.71	0.14	0.07
C5	0.23	-0.04	0.15	-0.24	0.41	-0.65	-0.45	-0.26	0.02
C6	0.24	0.17	0.47	-0.16	0.46	0.31	-0.19	0.53	-0.01
C7	0.33	0.2	0.24	-0.28	0.32	0.52	0.11	-0.55	0.08
C8	0.68	-0.41	0.00	-0.29	-0.31	0.06	-0.12	0.06	-0.11
C9	0.58	0.53	-0.35	0.03	-0.13	-0.01	-0.02	0.00	0.34
C10	0.33	0.39	-0.53	-0.06	0.28	-0.01	0.1	0.05	-0.59
C11	0.75	-0.59	-0.03	0.12	0.01	0.01	0.04	0.03	-0.01
C12	0.73	-0.47	-0.01	0.01	-0.13	0.05	0.02	-0.01	-0.02
C13	0.52	0.6	-0.39	0.07	-0.01	-0.04	-0.04	0.06	0.19

[2] Aluja Banet, T.; Morineau, A. (1999)

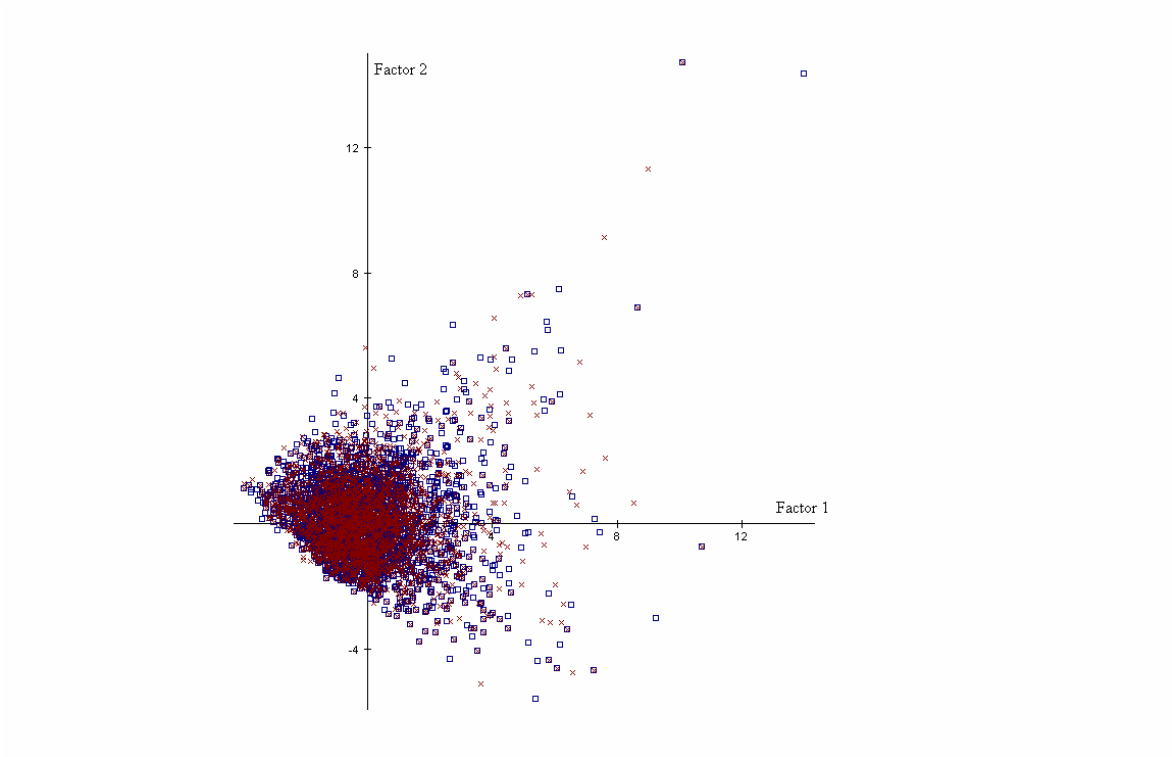


____Variables comunes (activas) _ _ _Variables específicas (ilustrativas)
Figura 4.2 Correlaciones de las variables activas e ilustrativas con los factores

La figura 4.3 es una representación gráfica de los individuos en un plano de los primeros dos ejes factoriales. Los individuos donantes (activos) y los receptores (suplementarios) se colocan sobre el mismo plano factorial. Es una representación del espacio común. Es la representación en un plano de los individuos, que mejor aproxima las distancias existentes entre ellas.

Se puede ver en esta gráfica la semejanza de los grupos o muestras en las variables comunes. Se puede observar también la existencia de algunos puntos muy alejados.

En la figura 4.4 se muestran las relaciones entre las variables específicas (gráficas de dispersión). Se puede ver también la forma de la distribución probabilística de las variables (histogramas de frecuencia). Son elementos que se pueden emplear en la etapa de preproceso.



□ Individuos activos (donantes) x Individuos ilustrativos (receptores)

Figura 4.3 Espacio común de los individuos en el primer plano factorial

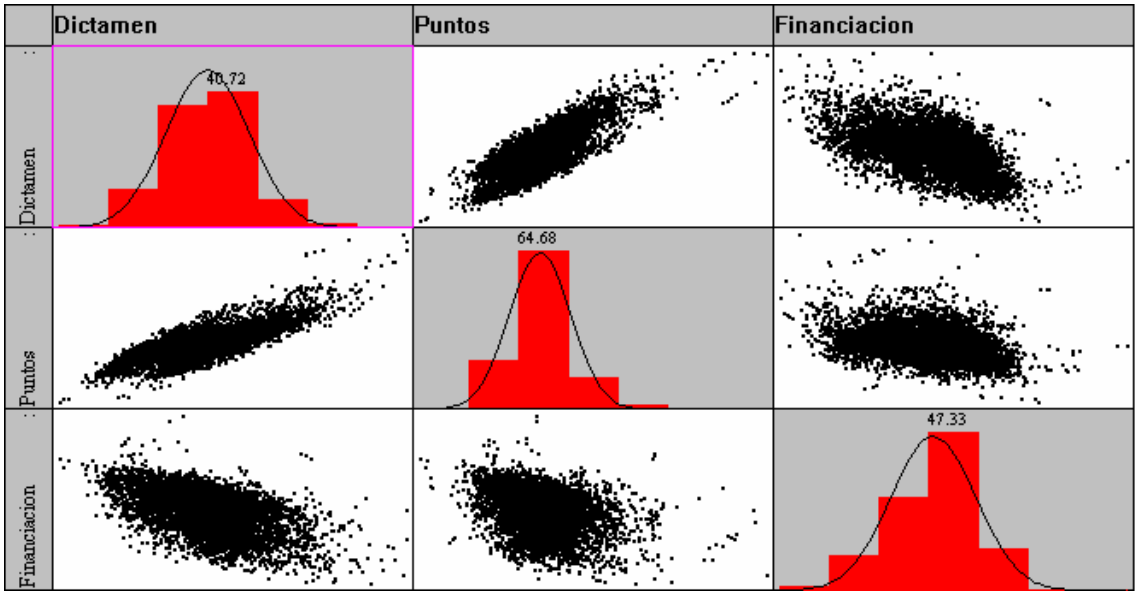


Figura 4.4 Características de las variables específicas continuas (Base 0)

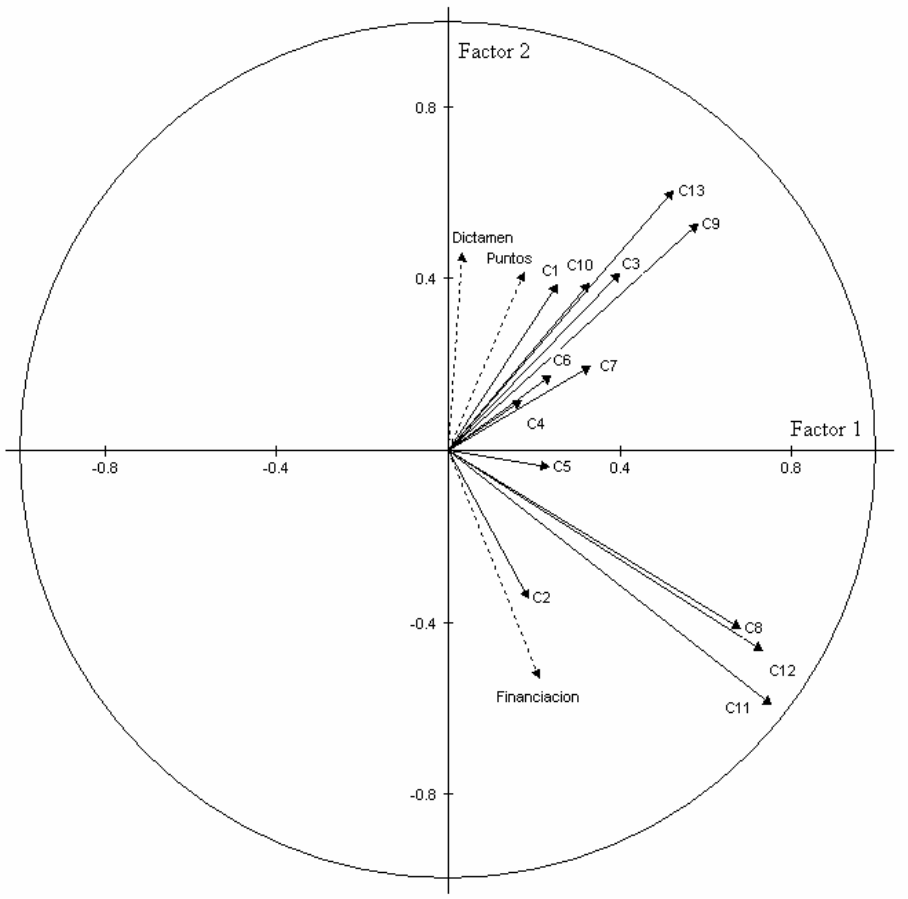
En el caso de la Base 1 se tienen los siguientes resultados:

Base 1:

VARIABLES	CORRELATIONS VARIABLE-FACTEUR								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
C2 - Dictamen	0.03	0.46	0.21	0.11	-0.31	0.42	0.34	-0.26	0.01
C3 - Puntos	0.18	0.41	0.18	0.01	0.00	0.33	0.24	-0.70	0.02
C4 - Financiacion	0.21	-0.53	-0.10	0.32	0.43	-0.07	0.11	0.09	0.09

MODALITES	VALEURS-TEST								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
4 . Dictamen_d									
AA_1 - C5=1	-2.2	-19.8	-10.5	-4.2	13.9	-17.4	-15.1	11.4	0.0
AA_2 - C5=2	2.2	19.8	10.5	4.2	-13.9	17.5	15.1	-11.4	0.0
5 . Puntos_d									
AB_1 - C6=1	-0.1	-14.2	-3.6	-1.7	0.7	-12.2	-12.2	28.9	-0.7
AB_2 - C6=2	-8.0	-7.2	-6.8	-1.0	1.6	-2.4	-0.4	3.8	2.7
AB_3 - C6=3	-3.5	3.4	-1.6	-0.8	0.6	0.1	3.9	-4.1	-0.9
AB_4 - C6=4	0.0	5.7	0.9	2.2	-1.5	4.4	3.2	-6.9	-1.9
AB_5 - C6=5	4.6	7.0	8.3	3.2	-2.1	7.4	4.7	-14.5	-1.6
AB_6 - C6=6	13.9	17.4	9.8	-0.8	-0.7	10.8	6.7	-25.0	1.8
6 . Financiacion_d									
AC_1 - C7=1	0.8	2.1	1.6	-1.9	-2.9	0.9	0.1	-0.9	-0.7
AC_2 - C7=2	-2.4	19.2	5.2	-13.6	-17.6	1.7	-4.8	-2.2	-6.5
AC_3 - C7=3	-8.2	15.0	1.2	-11.1	-12.7	1.2	-5.9	-2.7	0.4
AC_4 - C7=4	-2.0	-1.1	-0.3	4.0	3.8	1.3	4.0	-0.9	-0.1
AC_5 - C7=5	4.8	-16.2	-3.2	12.2	14.1	-1.7	4.1	3.7	2.3
AC_6 - C7=6	11.3	-14.3	-1.3	3.0	7.1	-3.5	-0.6	1.9	2.7

Nuevamente se puede ver en el reporte anterior que la capacidad predictiva es intermedia. Esto se puede apreciar en los valores Test y las correlaciones Variable-Factor que se muestran.



____Variables activas (comunes) _ _ _Variables ilustrativas (específicas)

Figura 4.5 Correlaciones de las variables activas e ilustrativas con los factores (Base 1)

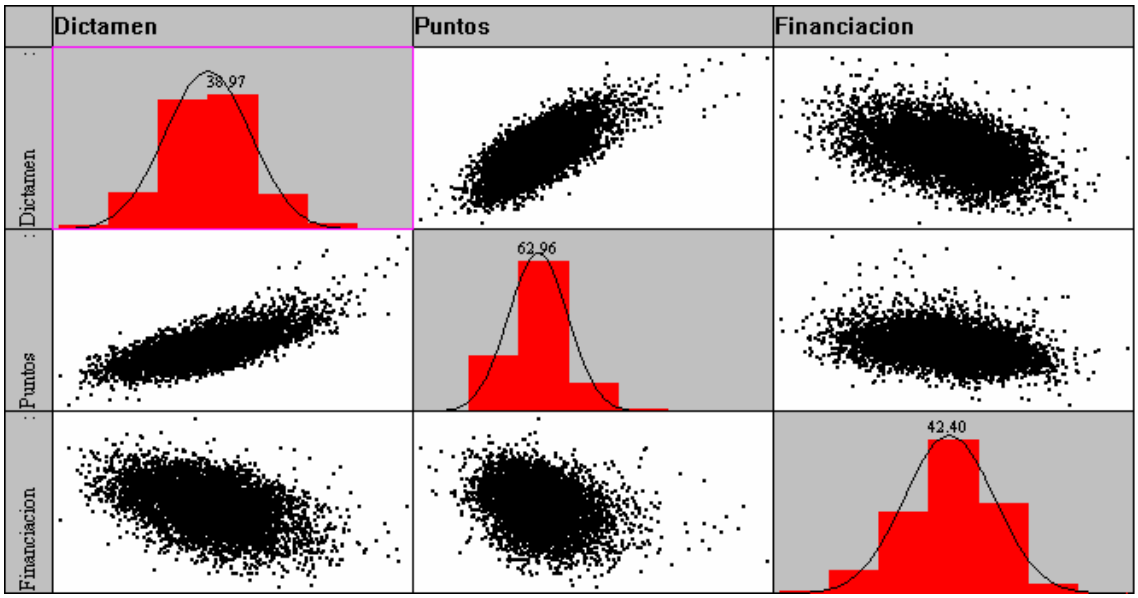


Figura 4.6 Características de las variables específicas (Base 1)

En el caso de la Base 2 se tienen los siguientes resultados:

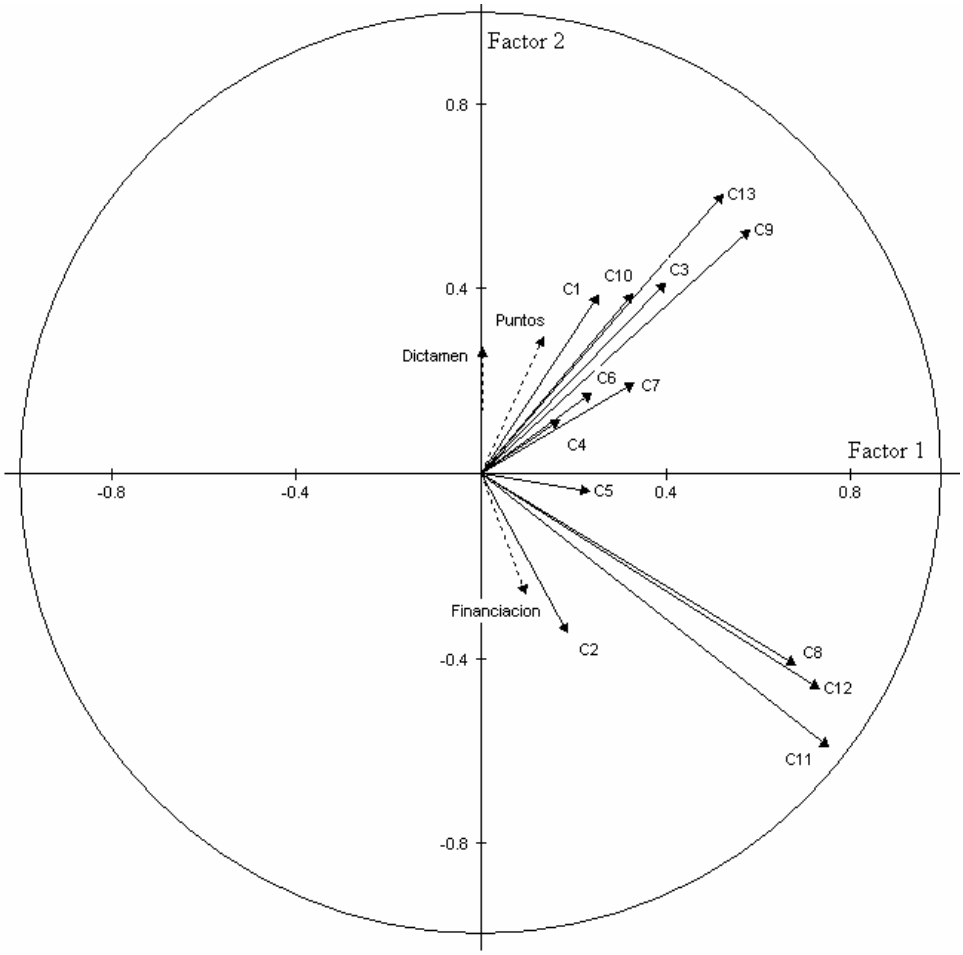
Base 2:

VARIABLES	CORRELATIONS VARIABLE-FACTEUR								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
C2 - Dictamen	0.00	0.27	0.10	0.07	-0.17	0.23	0.20	-0.17	0.02
C3 - Puntos	0.13	0.29	0.12	0.01	0.00	0.23	0.19	-0.53	0.03
C4 - Financiacion	0.10	-0.26	-0.03	0.18	0.22	-0.06	0.07	0.05	0.02

MODALITES	VALEURS-TEST								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
4 . Dictamen_d									
AA_1 - C5=1	-1.1	-12.3	-4.5	-3.5	7.5	-10.6	-9.0	7.2	-0.5
AA_2 - C5=2	1.1	12.3	4.5	3.5	-7.5	10.6	9.0	-7.2	0.5
5 . Puntos_d									
AB_1 - C6=1	-2.4	-11.3	-2.9	-2.2	1.5	-7.2	-8.7	22.2	-0.4
AB_2 - C6=2	-5.8	-2.0	-2.2	-0.3	0.3	-2.2	0.4	2.0	0.3
AB_3 - C6=3	-0.1	-0.7	-1.7	1.2	0.0	0.8	1.5	-2.0	-1.0
AB_4 - C6=4	0.8	1.8	-1.5	0.7	-1.8	0.3	0.8	-3.4	0.8
AB_5 - C6=5	1.2	4.5	2.2	1.8	-1.6	2.3	3.2	-6.9	-1.9
AB_6 - C6=6	8.8	13.0	7.3	-0.1	0.6	9.5	6.5	-21.9	2.2
6 . Financiacion_d									
AC_1 - C7=1	-2.8	6.5	1.0	-4.9	-5.5	1.7	-2.8	-1.4	-0.2
AC_2 - C7=2	-3.7	9.1	0.7	-5.4	-9.9	0.8	-3.3	-2.6	-0.9
AC_3 - C7=3	-0.4	3.7	-0.1	-2.7	-1.7	2.2	-0.4	-1.1	-0.7
AC_4 - C7=4	0.5	-0.5	0.5	-1.5	0.2	1.3	1.9	1.4	-0.
AC_5 - C7=5	-0.4	-3.1	-1.0	4.3	3.2	-1.1	1.2	0.2	1.1
AC_6 - C7=6	4.6	-10.9	-0.5	6.8	9.5	-3.5	1.6	2.4	1.3

Se puede observar que la Base 2 indica un bajo poder predictivo de las variables específicas, augurando una fusión de datos problemática.

En la figura 4.5, se puede observar una ligera variación en la variable específica Financiación con respecto a la observada para la Base 0. La variación resulta más notoria en la gráfica 4.7 (Base 2). Vale la pena recordar que la Base 0 contiene la matriz de varianzas-covarianzas reducida y la Base 2, aumentada.



____Variables activas (comunes) _ _ _Variables ilustrativas (específicas)

Figura 4.7 Correlaciones de las variables activas e ilustrativas con los factores (Base 2)

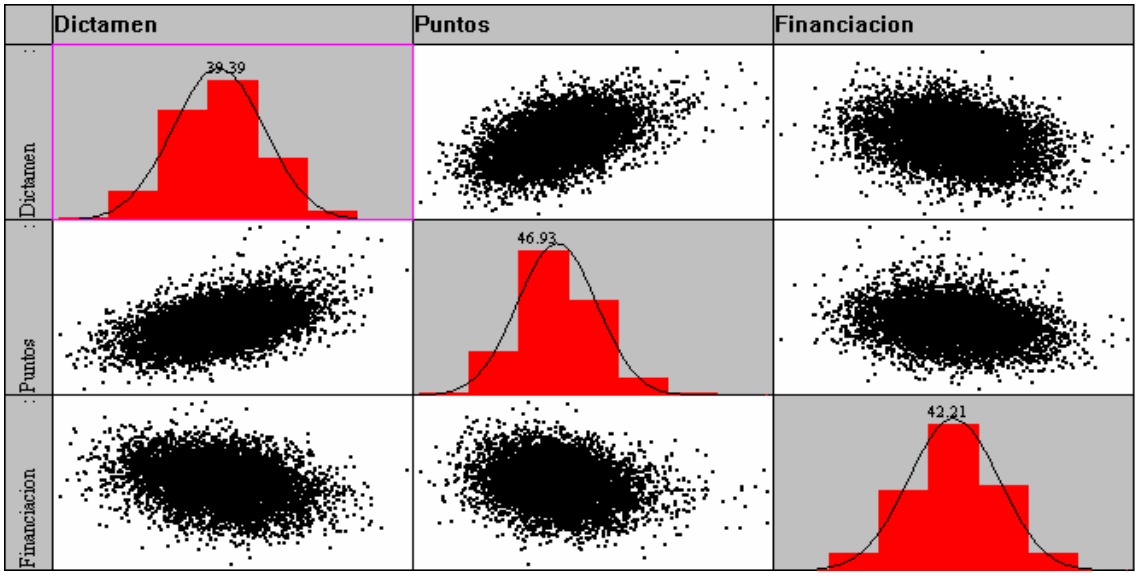


Figura 4.8 Características de las variables específicas (Base 2)

4.3 Ensayos (Simulación I)

En la parte experimental, los ensayos, representan la ejecución del sistema GRAFT para las distintas opciones de imputación. Con el fin de evaluar el comportamiento de los distintos métodos implantados en el sistema, se harán ejecuciones con diferentes valores para los parámetros requeridos en cada caso. Los resultados de los distintos ensayos se presentarán mediante reportes diseñados para su análisis (cf. 3.5). Algunas gráficas construidas con estos resultados serán presentadas para facilitar la interpretación de los resultados. El mismo procedimiento de experimentación se hará para cada uno de los archivos construidos para la investigación (Base0, Base1, Base2).
Se presentan 4 gráficas con la siguiente información para el análisis:

- Gráfica 1: Se muestra en esta gráfica el nivel de significancia para una prueba de hipótesis de igualdad de medias y desviación estándar en el caso de las variables continuas y una prueba $\chi 2$ en el caso de las variables categóricas (cf. 3.5.1). Se gráfica el nivel de significancia promedio -Global Stats. (ASL)-. El valor de referencia para estas gráficas es el valor obtenido por la imputación empleando la media (moda en el caso categórico) en el caso determinístico y las extracciones aleatorias globales en el caso estocástico.

- Gráfica 2: En esta gráfica se presentan los resultados relativos a la exactitud de la imputación para las variables continuas y categóricas empleando los indicadores - Tau y Tau' – (cf. 3.3).

- Gráfica 3: En esta gráfica se muestran los resultados de la homogeneidad que existe entre las variables continuas imputadas, las variables categóricas imputadas y la homogeneidad que existe entre las variables continuas y categóricas imputadas. Se gráfica el promedio de la diferencia en valor absoluto de las correlaciones de cada par de variables medidas en los donantes como en los receptores –Internal Homogeneity (ACD) – (cf. 3.5.2).

- Gráfica 4: Esta gráfica es similar a la gráfica 3, aplicada a la medición de homogeneidad entre variables comunes y específicas –External Homogeneity (ACD) -.

Los resultados que a continuación se presentan corresponden a los métodos:

- T1DM Take One Deterministic Multivariate
- T1SM Take One Stochastic Multivariate
- TKDM Take K Deterministic Multivariate
- TKSM Take K Stochastic Multivariate

Se ha decidido reportar solo estos métodos, debido a que los métodos univariantes tienen la característica de no preservar las relaciones entre las variables específicas. No significa esto que estos métodos no tengan validez, pues si de antemano se sabe que las variables específicas son independientes, estos métodos podrían ser empleados.

Se desea observar el comportamiento de los resultados obtenidos con el método T1DM haciendo variar el porcentaje de repetición de los individuos donantes (parámetro P).
De la misma forma, se desea observar el comportamiento de los resultados obtenidos con los métodos TKDM y TKSM haciendo variar el número de vecinos empleados para la imputación (parámetro K).

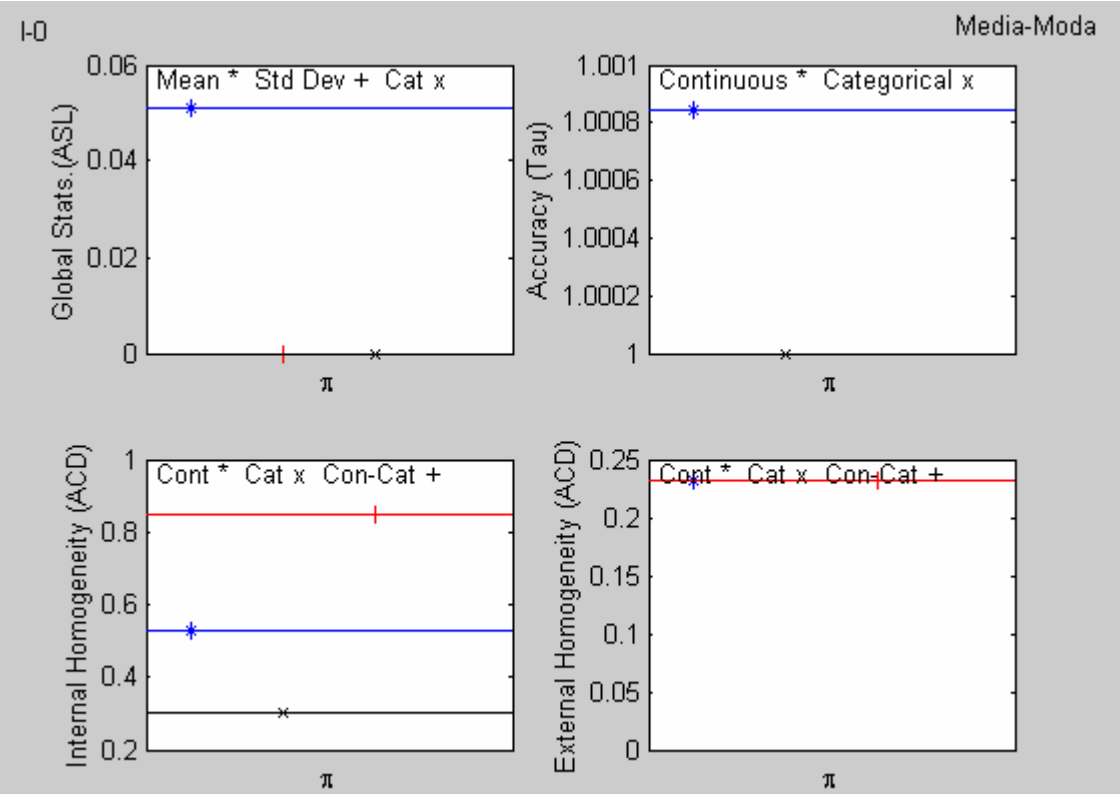


Figura 4.9a Valores de referencia ($\pi = 0$)

Omitir la gráficas de homogeneidad.

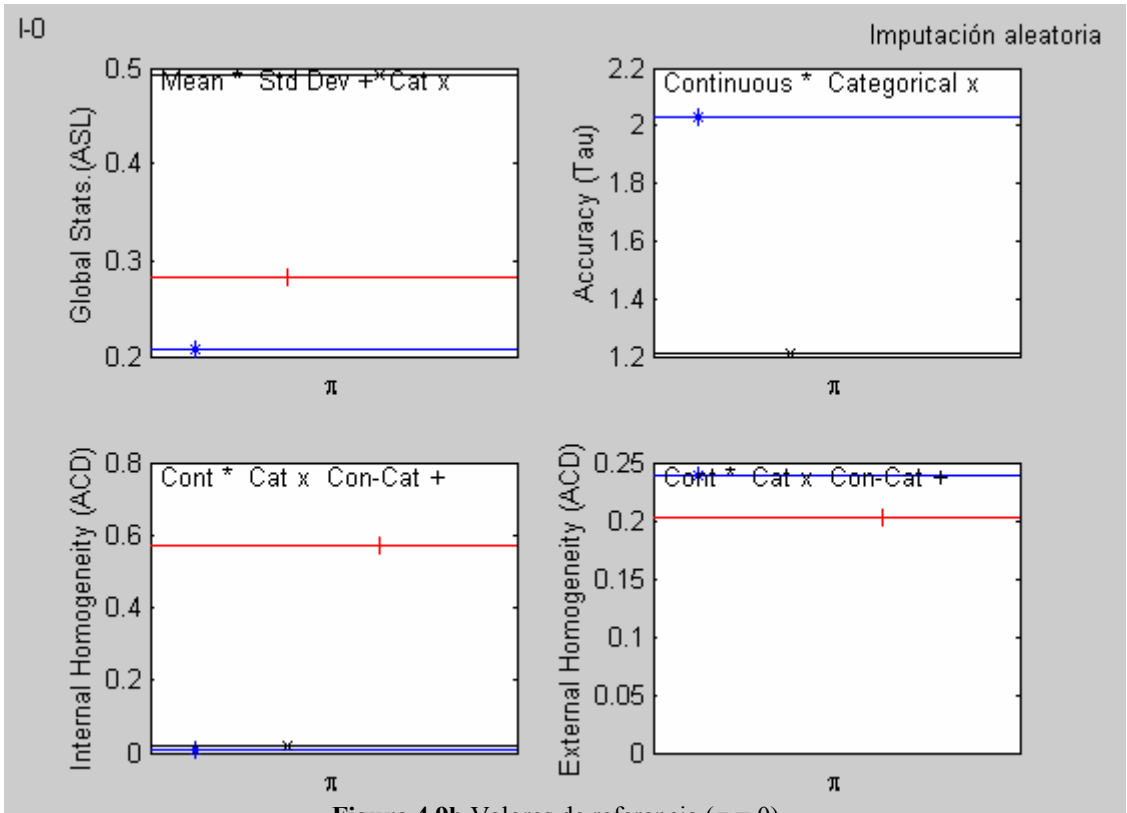


Figura 4.9b Valores de referencia ($\pi = 0$)

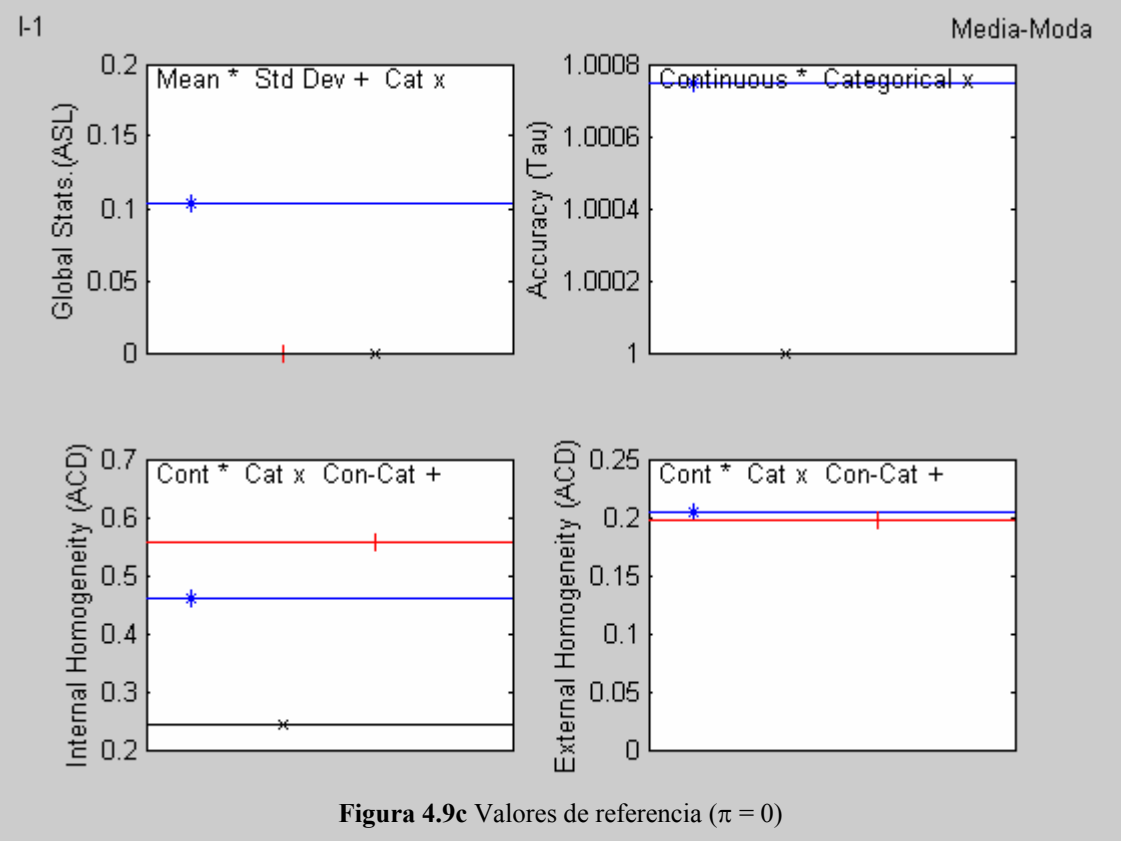


Figura 4.9c Valores de referencia ($\pi = 0$)

Omitir la gráficas de homogeneidad.

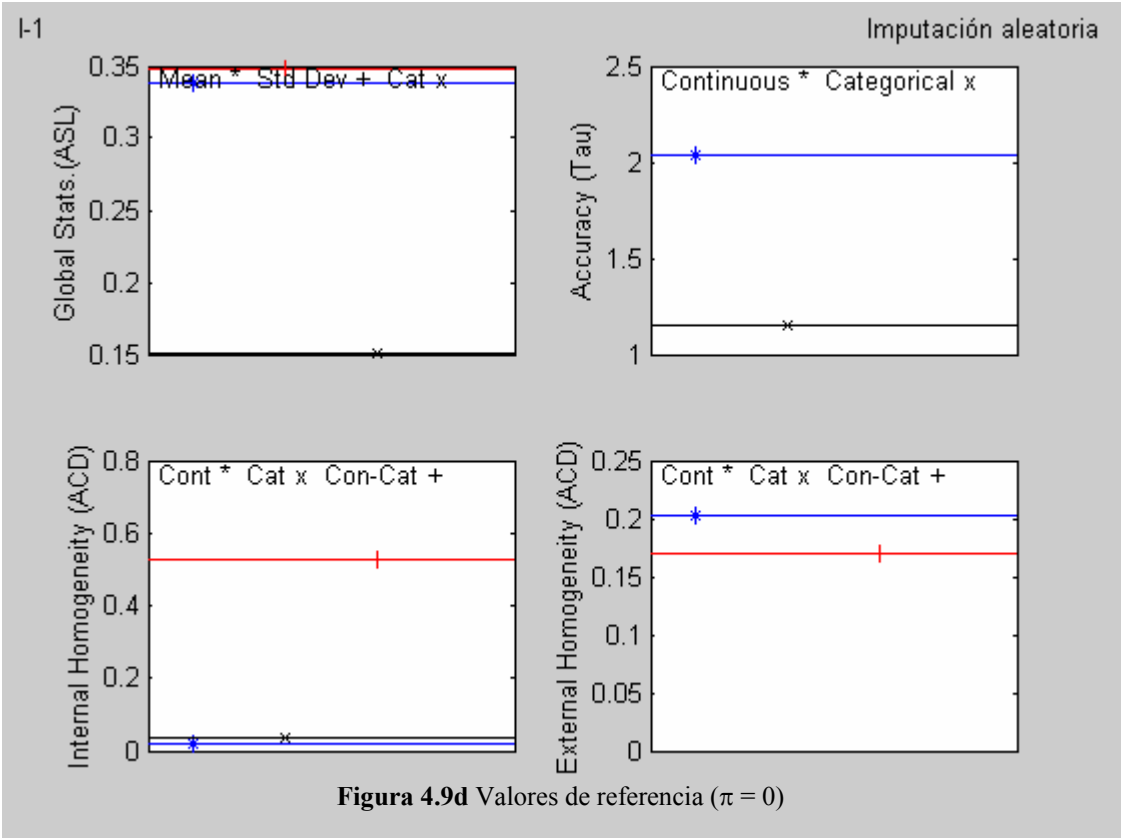


Figura 4.9d Valores de referencia ($\pi = 0$)

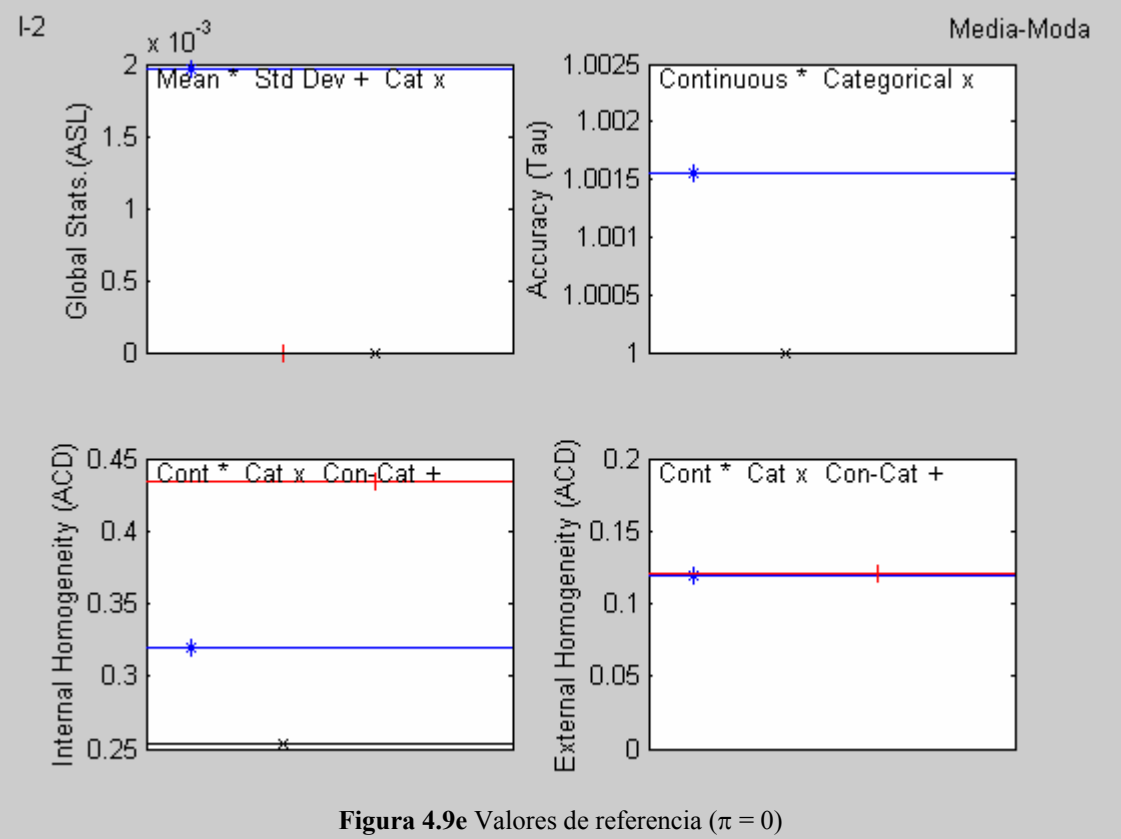


Figura 4.9e Valores de referencia ($\pi = 0$)

Omitir la gráficas de homogeneidad.

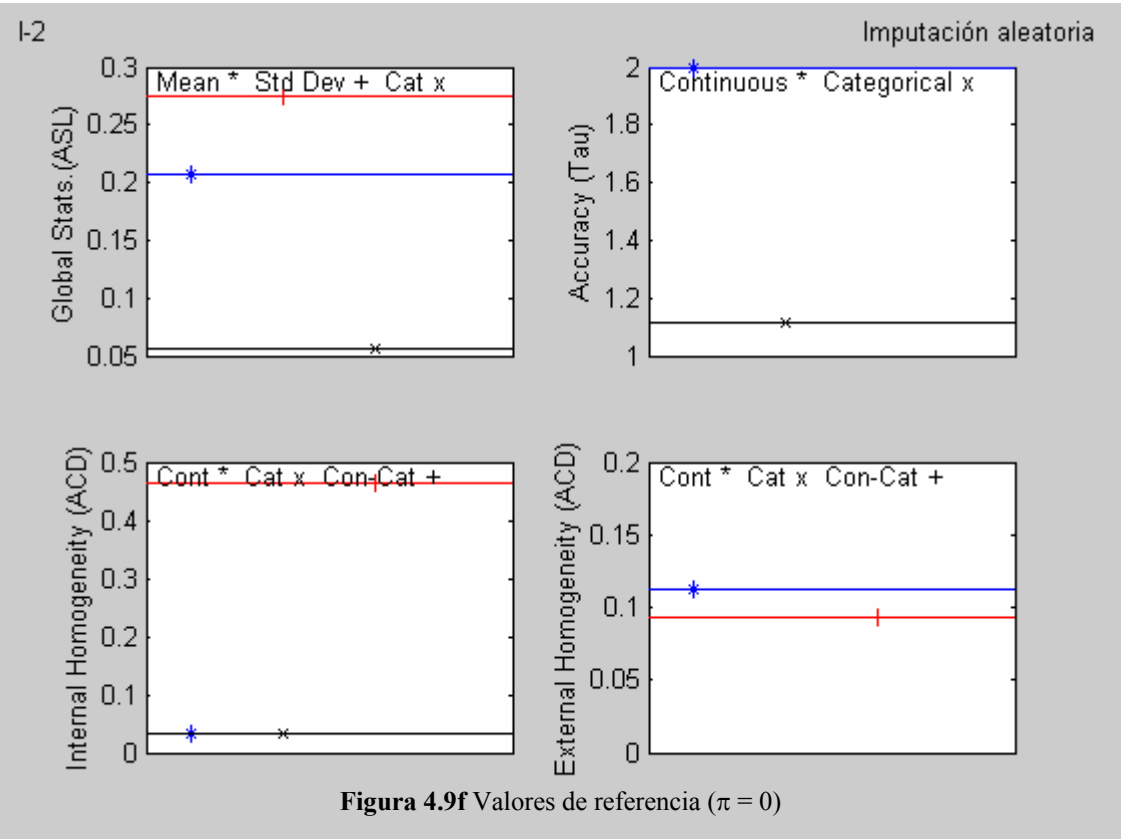


Figura 4.9f Valores de referencia ($\pi = 0$)

4.3.1 Imputación T1DM

Ensayo 1.1.

Procedimiento de imputación:	T1DM (Take One Deterministic Multivariate)
Parámetro:	P (probabilidad de repetición de los donantes)
Valor del Parámetro:	P = 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0
Archivos:	Base0, Base 1, Base 2

Los valores de referencia (imputación empleando la media –moda-) se pueden observar en las figuras 4.9a, 4.9c y 4.9e.

4.3.1.1 Resultados:

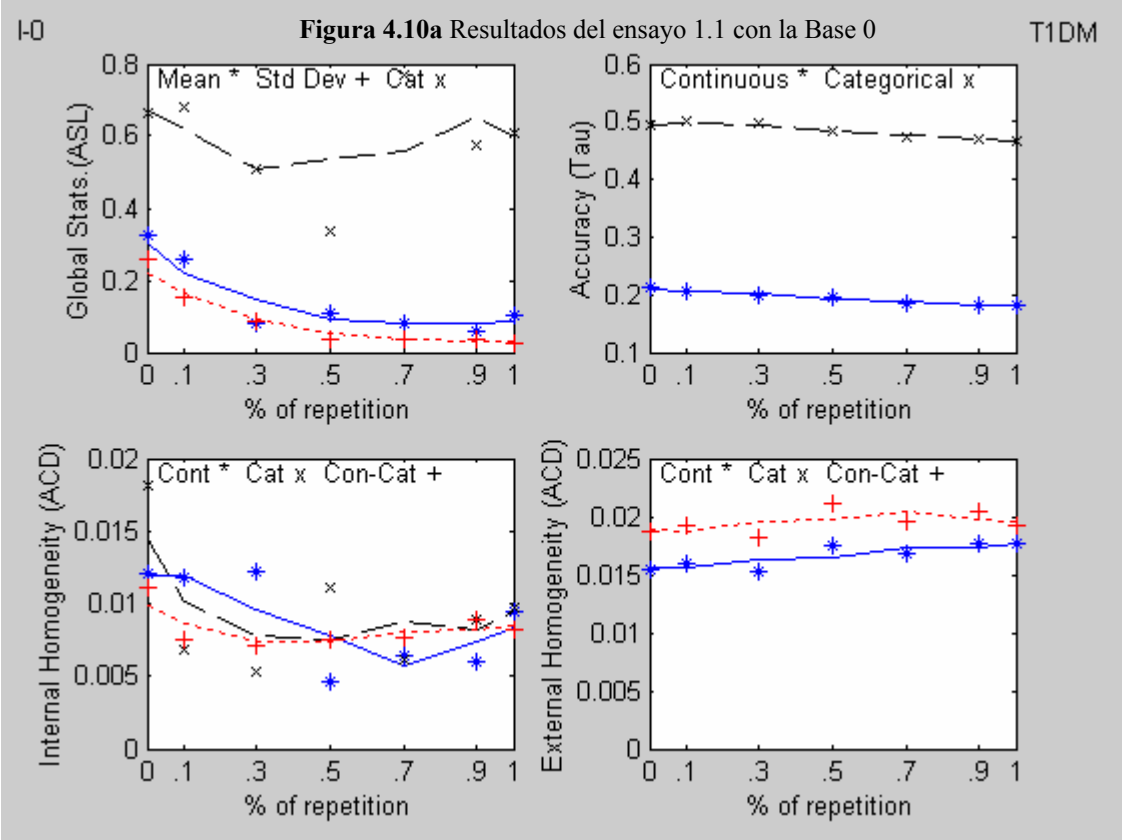


Figura 4.10a Resultados del ensayo 1.1 con la Base 0

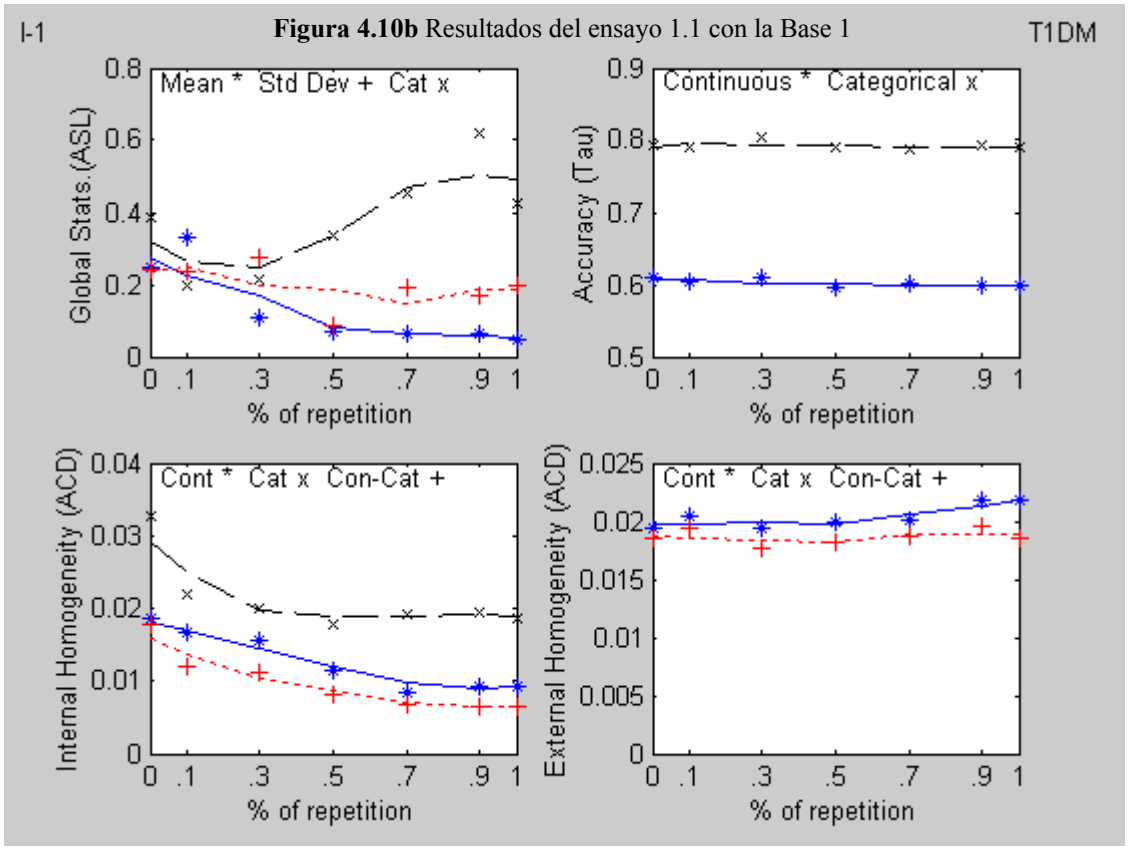


Figura 4.10b Resultados del ensayo 1.1 con la Base 1

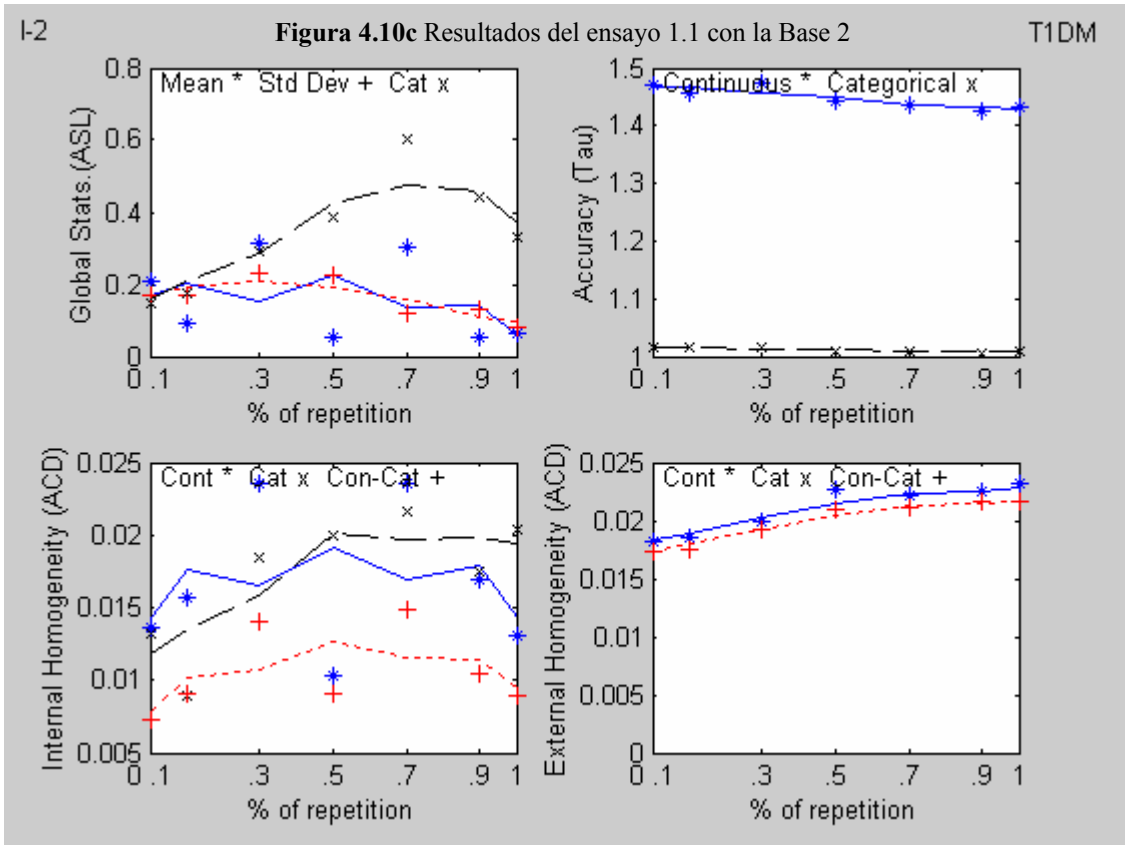


Figura 4.10c Resultados del ensayo 1.1 con la Base 2

4.3.1.2 Análisis de los resultados

- ASL.- Para los casos de la varianza residual normal y reducida, se reproduce bien la media y la desviación estándar para valores de P = 0, con una tendencia a empeorar con el incremento de P. En el caso del ruido aleatorio amplificado, los resultados son ligeramente diferentes; se reproduce mejor la desviación estándar. En todos los casos, la reproducción categórica presenta buenos resultados.

- Tau.- Para los casos de la varianza residual normal y reducida, los valores obtenidos del índice (continuo y categórico) son buenos y no parecen depender del parámetro P. En el caso del ruido aleatorio amplificado, la imputación por la media, comparada con este método resulta mejor opción.

- ACD (Interna).- Los valores obtenidos del índice son buenos y en los casos de la varianza residual normal y reducida, se observa una tendencia decreciente con el parámetro P. En el caso del ruido aleatorio amplificado, esta tendencia no se observa.

- ACD (Externa).- Los valores obtenidos del índice son buenos y en los casos de la varianza residual normal y reducida, parece no tener efecto el parámetro P. En el caso del ruido aleatorio amplificado, se observa una ligera tendencia creciente con el parámetro P.

4.3.2 Imputación T1SM

Ensayo 1.2.

Procedimiento de imputación: T1SM (Take One Stochastic Multivariate)

Parámetro: Ninguno

7 repeticiones

Archivos: Base0, Base 1, Base 2

Los valores de referencia (imputación aleatoria) se pueden observar en las figuras 4.9b, 4.9d y 4.9f.

4.3.2.1 Resultados:

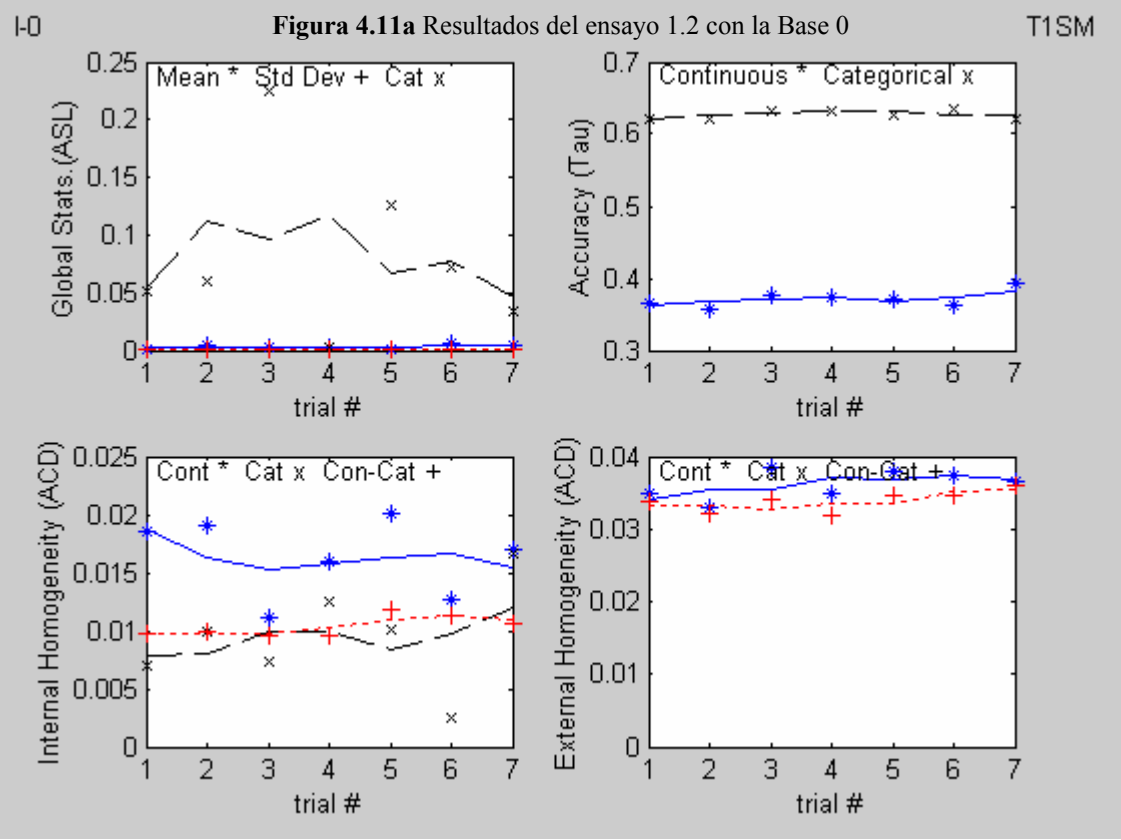


Figura 4.11a Resultados del ensayo 1.2 con la Base 0

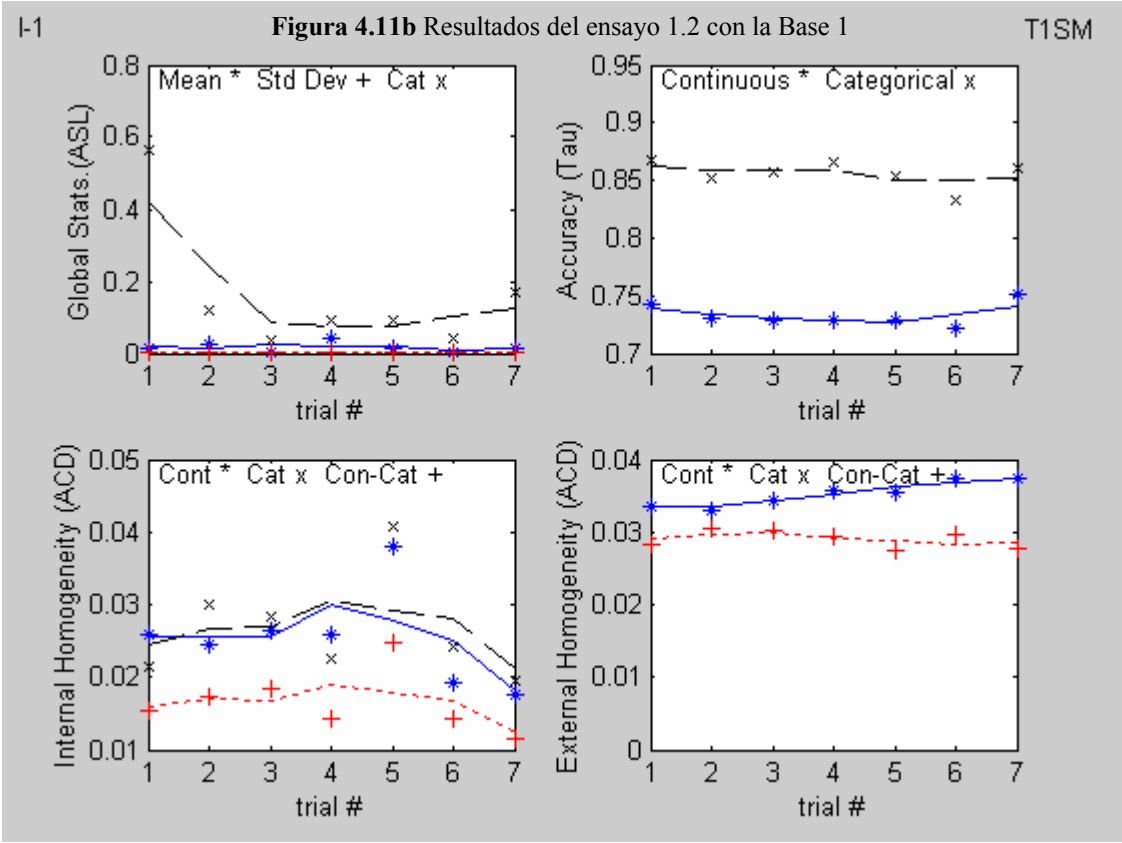


Figura 4.11b Resultados del ensayo 1.2 con la Base 1

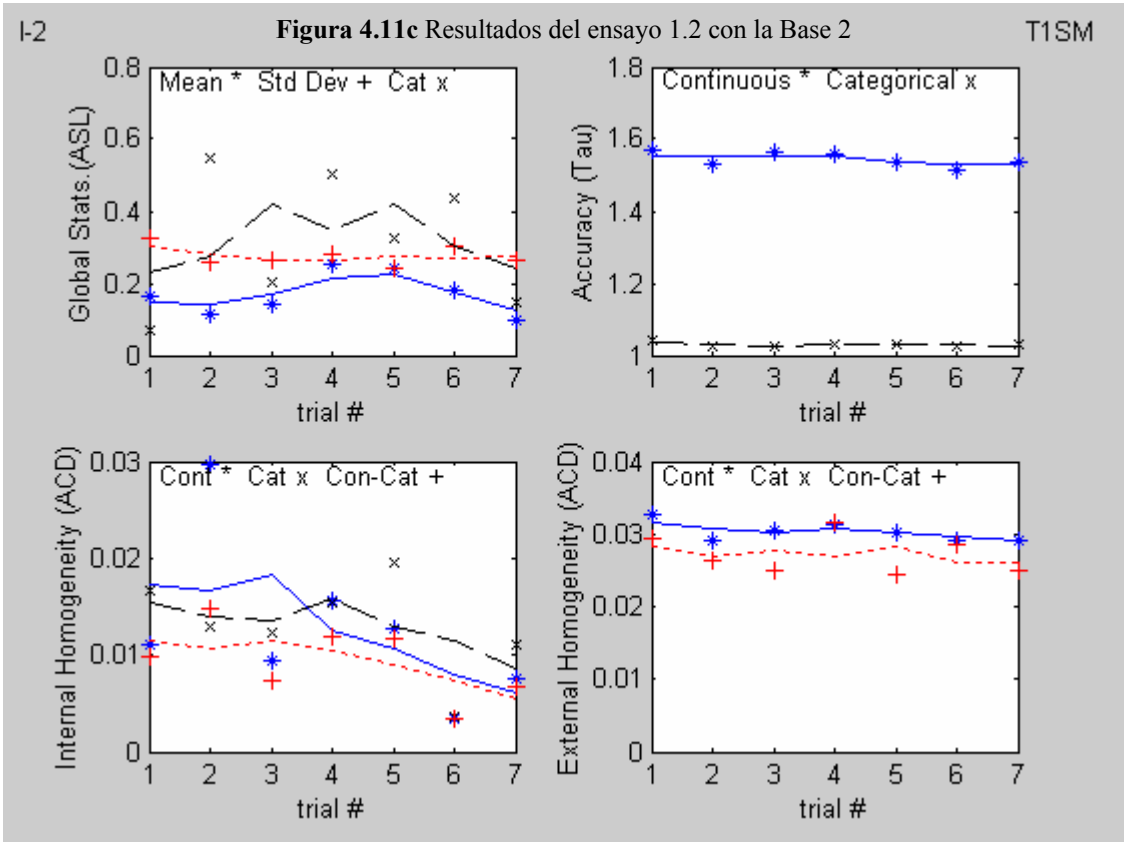


Figura 4.11c Resultados del ensayo 1.2 con la Base 2

4.3.2.2 Análisis de los resultados

- ASL.- Para los casos de la varianza residual normal y reducida, no se reproducen bien la media ni la desviación estándar. En el caso del ruido aleatorio amplificado, los resultados son diferentes; se reproducen bien la media y la desviación estándar. En todos los casos, la reproducción categórica presenta mejores resultados.

- Tau.- Para los casos de la varianza residual normal y reducida, los valores obtenidos del índice (continuo y categórico) son buenos en promedio. En el caso del ruido aleatorio amplificado, la imputación por la media, comparada con este método resulta mejor opción.

- ACD (Interna).- No parece tener influencia el incremento de fluctuación aleatoria en los resultados. Los resultados obtenidos con el método T1DM son ligeramente mejores.

- ACD (Externa).- No parece tener influencia el incremento de fluctuación aleatoria en los resultados. Nuevamente se observa que los resultados obtenidos con el método T1DM son ligeramente mejores.

4.3.3 Imputación TKDM

Ensayo 1.3.

Procedimiento de imputación:	TKDM (Take K Deterministic Multivariate)
Parámetro:	K (número de vecinos)
Valor del Parámetro:	K = 5, 10, 15, 20, 25, 30, 35, 40
Archivos:	Base0, Base 1, Base 2

Los valores de referencia (imputación empleando la media –moda-) se pueden observar en las figuras 4.9a, 4.9c y 4.9e.

4.3.3.1 Resultados:

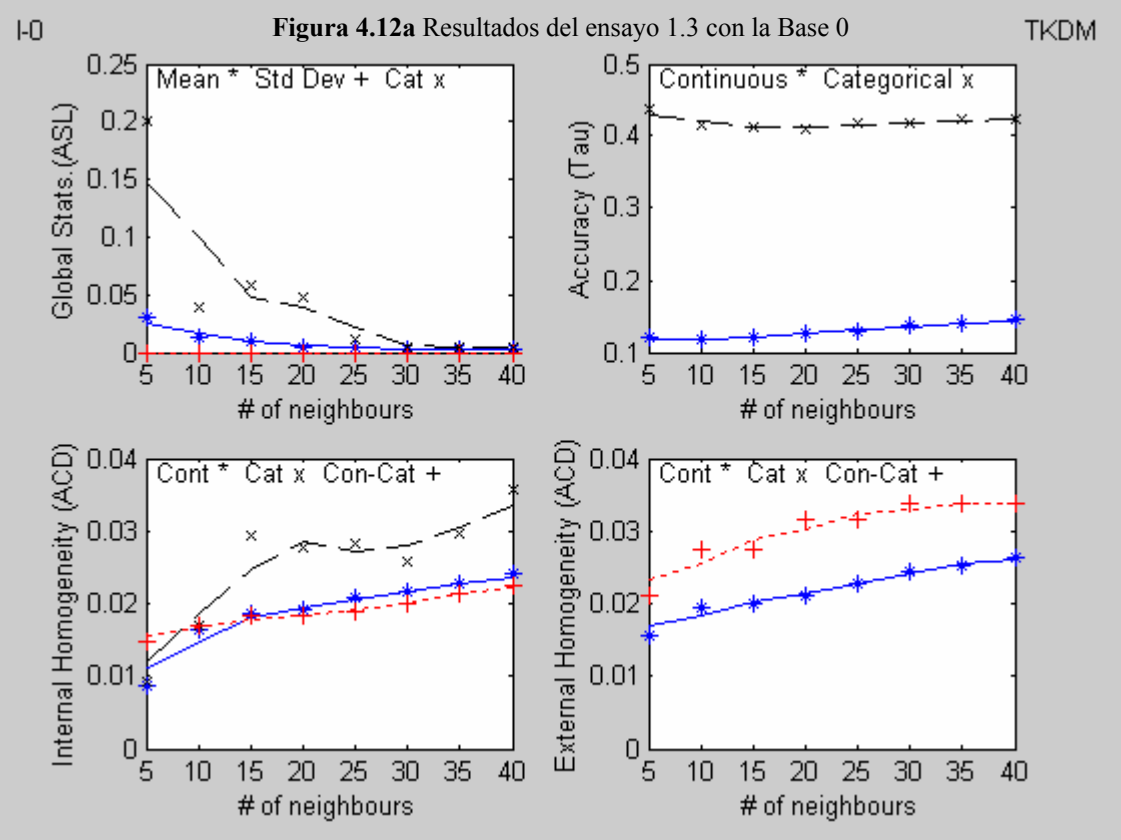
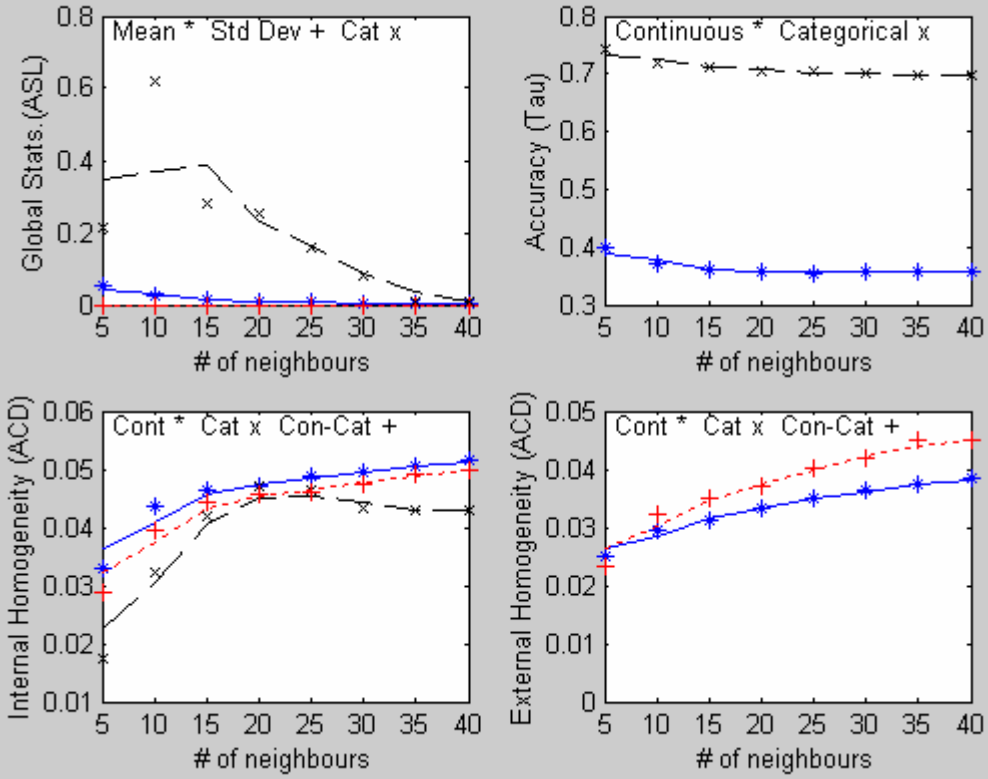
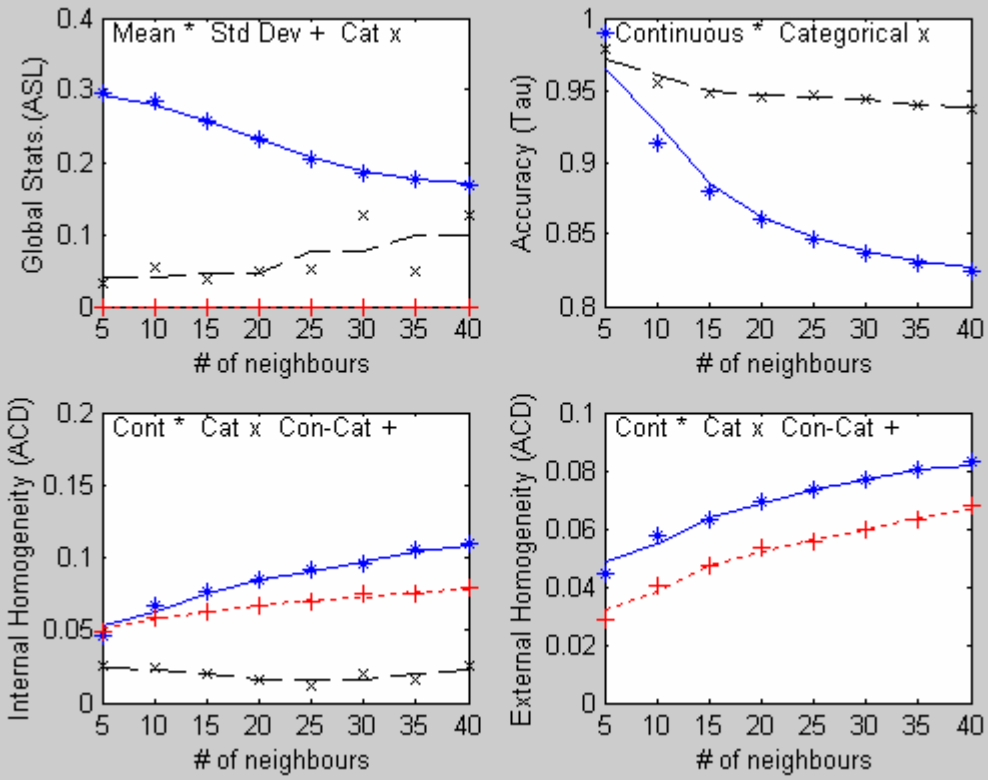


Figura 4.12a Resultados del ensayo 1.3 con la Base 0

I-1 **Figura 4.12b** Resultados del ensayo 1.3 con la Base 1 TKDM



I-2 **Figura 4.12c** Resultados del ensayo 1.3 con la Base 2 TKDM



4.3.3.2 Análisis de los resultados

- ASL.- Para los casos de la varianza residual normal y reducida, la mejor reproducción de la media se presenta para K = 5, con una tendencia a empeorar con el incremento de K. Se reproduce mal la desviación estándar. En el caso del ruido aleatorio amplificado, los resultados son diferentes; se reproduce mejor la media y presenta una tendencia decreciente con el valor de K. En todos los casos, la reproducción categórica presenta mejores resultados.

  Se obtienen los mejores resultados con un solo vecino (método T1DM). El hecho de tomar un número creciente de vecinos significa obtener estadísticos marginales cada vez más alejados de los reales.

- Tau.- Para los casos de la varianza residual normal y reducida, los valores obtenidos del índice (continuo y categórico) son buenos y no parecen depender del parámetro K. En el caso del ruido aleatorio amplificado, la imputación mejora incrementando el valor de K.

- ACD (Interna).- Los valores obtenidos del índice son buenos en todos los casos de la varianza residual; se observa una ligera tendencia creciente con el parámetro K. Los valores aumentan en general con la varianza residual.

- ACD (Externa).- Los valores obtenidos del índice son buenos en los casos de la varianza residual normal y reducida, no tanto así en el caso del ruido aleatorio amplificado. Se observa una ligera tendencia creciente con el parámetro K. Los valores aumentan en general con la varianza residual.

4.3.4 Imputación TKSM

Ensayo 1.4.

Procedimiento de imputación:	TKSM (Take K Stochastic Multivariate)
Parámetro:	K (número de vecinos)
Valor del Parámetro:	K = 5, 10, 15, 20, 25, 30, 35, 40
Archivos:	Base0, Base 1, Base 2

Los valores de referencia (imputación aleatoria) se pueden observar en las figuras 4.9b, 4.9d y 4.9f.

4.3.4.1 Resultados:

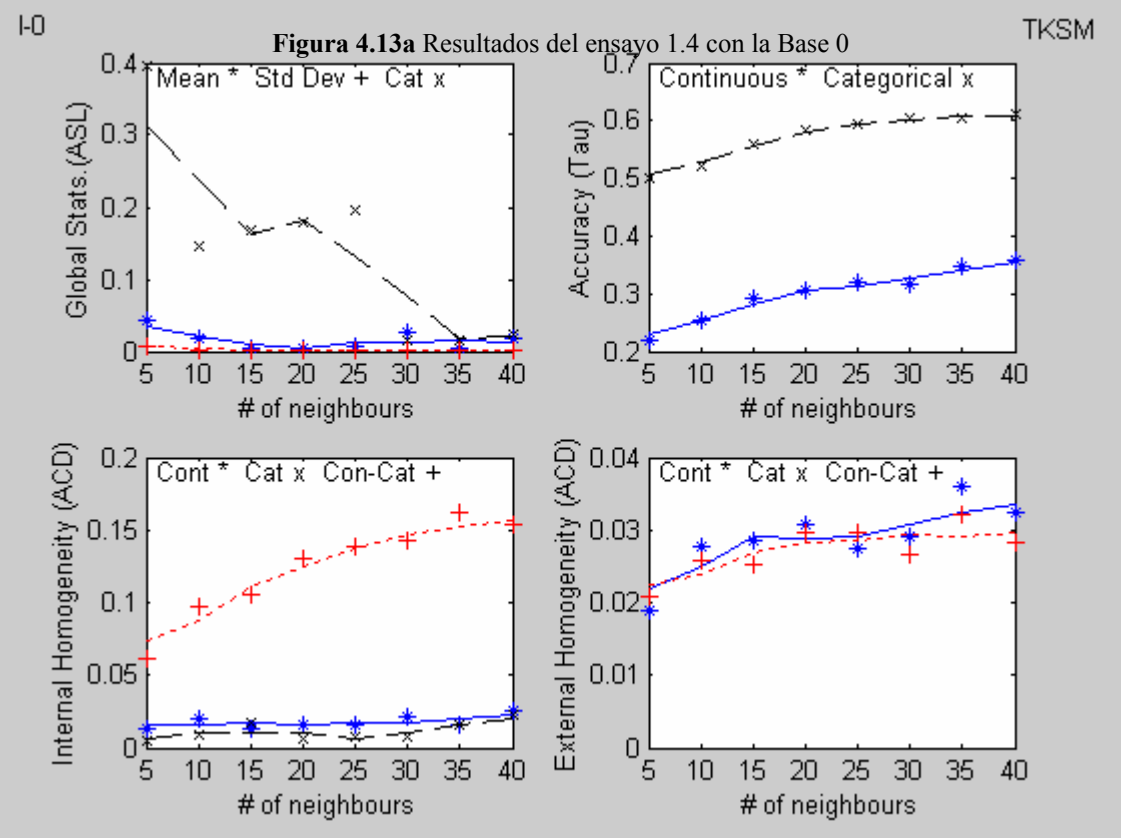
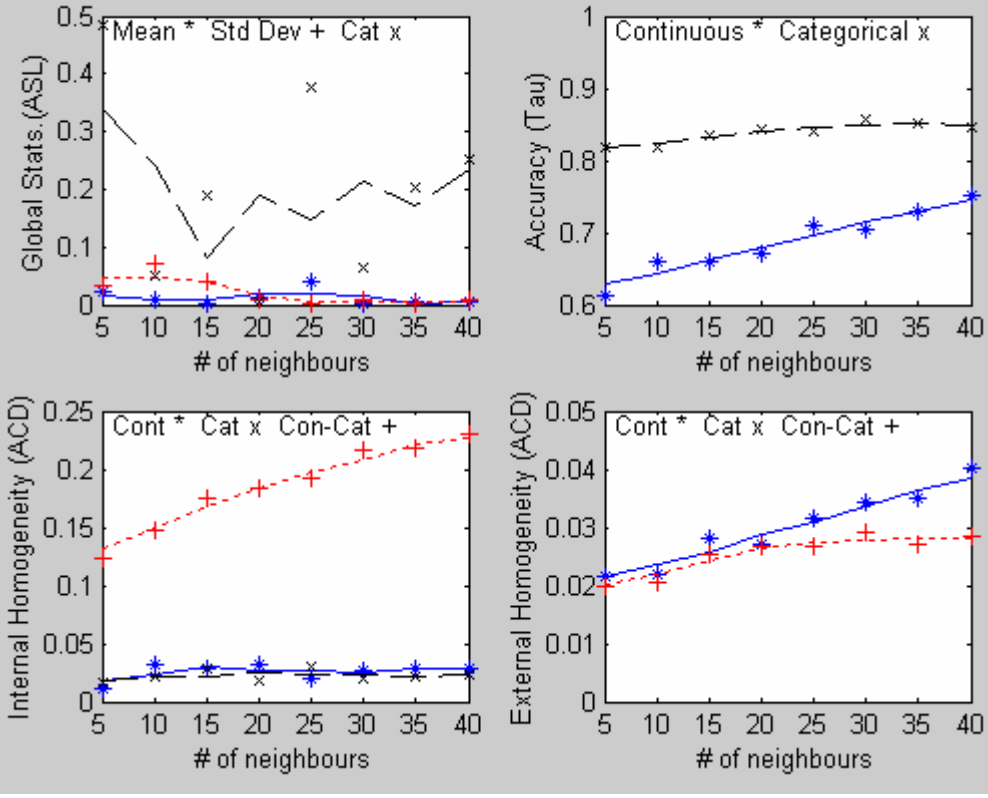
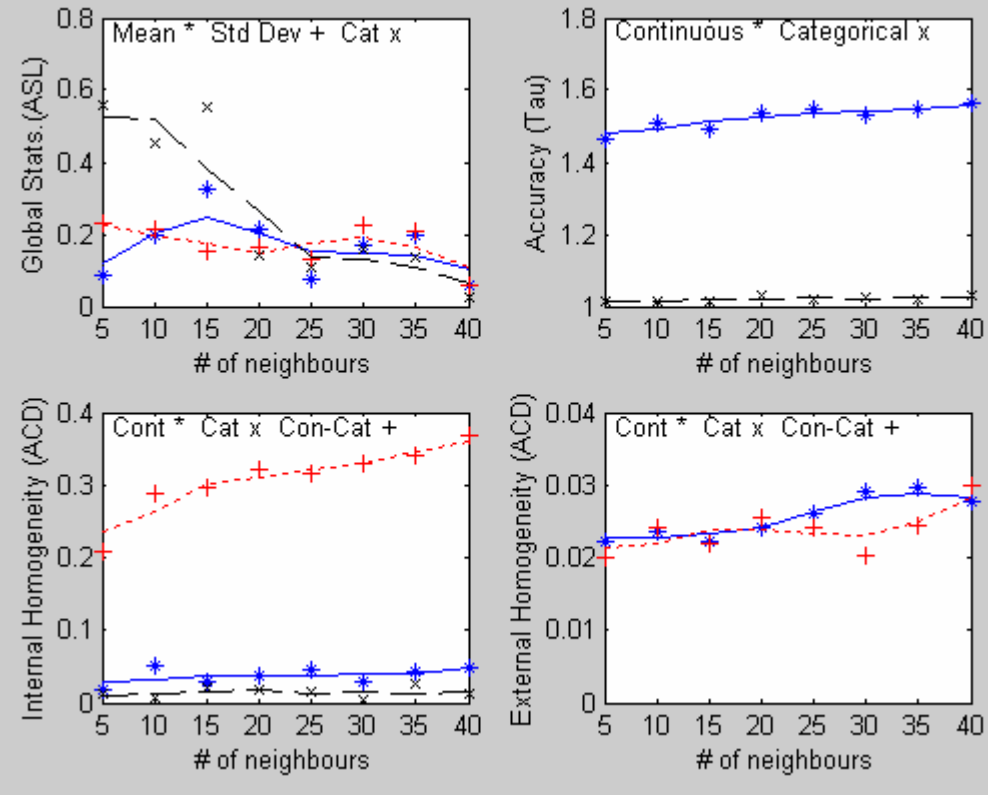


Figura 4.13a Resultados del ensayo 1.4 con la Base 0

I-1 **Figura 4.13b** Resultados del ensayo 1.4 con la Base 1 TKSM



I-2 **Figura 4.13c** Resultados del ensayo 1.4 con la Base 2 TKSM



4.3.4.2 Análisis de los resultados

- ASL Para los casos de la varianza residual normal y reducida, no se reproducen bien la media ni la desviación estándar. En el caso del ruido aleatorio amplificado, los resultados son diferentes; la reproducción de la media y la desviación estándar es mejor y no parece depender del número de vecinos. En todos los casos, la reproducción categórica presenta mejores resultados.

- Tau.- Para los casos de la varianza residual normal y reducida, los valores obtenidos del índice (continuo y categórico) son buenos y presentan una tendencia creciente con el parámetro K. En el caso del ruido aleatorio amplificado, los valores parecen no depender del valor de K. Los valores para el caso continuo son superiores a 1.

- ACD (Interna).- En todos los casos, los valores obtenidos son buenos y no parecen depender de la varianza residual.

- ACD (Externa).- En todos los casos, los valores obtenidos son buenos. Se observa una ligera tendencia creciente con el valor de K.

4.4 Construcción de la base de datos (Simulación II)

Ahora, la construcción de la base de datos se hará tomando en cuenta una característica específica. Se han seleccionado del archivo original como donantes, aquellos individuos con un valor específico en una variable (estado civil > 1). El resultado de esta selección fue un archivo de donantes con 5329 individuos y un archivo de receptores con 1633 individuos. El procedimiento de construcción de la base de datos ha sido el mismo que se empleó para construir la base para la simulación I (cf. 4.2).

Matrices de varianzas-covarianzas residuales para las variables específicas en las tres bases:

Base 0

0.00227	26.98513	-0.00018
26.98513	2438198.90	-6.90123
-0.00018	-6.90123	0.00037

Base 1

0.02039	242.86618	-0.00163
242.86618	21943790.00	-62.11103
-0.00163	-62.11103	0.00337

Base 2

0.18348	2185.79560	-0.01467
2185.79560	197494110.00	-558.99927
-0.01467	-558.99927	0.03036

Se han seleccionado 9 coordenadas factoriales para la construcción del modelo, que representan un porcentaje acumulado de la variabilidad explicada aproximado de 90 % como se puede ver en la tabla 4.4

Tabla 4.4 Tabla de valores propios

Número	Valor propio	Porcentaje de la Variabilidad explicada	Porcentaje acumulado
1	2.6962	20.74	20.74
2	1.9110	14.70	35.44
3	1.3705	10.54	45.98
4	1.1870	9.13	55.11
5	1.0743	8.26	63.38
6	1.0217	7.86	71.24
7	0.9246	7.11	78.35
8	0.7902	6.08	84.43
9	0.7043	5.42	89.84
10	0.5356	4.12	93.96
11	0.3269	2.51	96.48
12	0.3205	2.47	98.94
13	0.1372	1.06	100.00

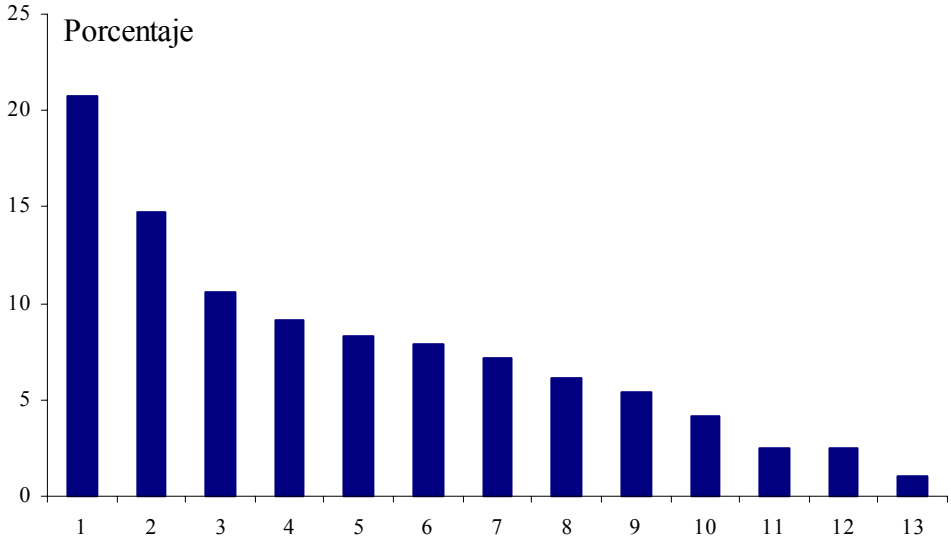


Figura 4.14 Histograma de valores propios

Se han seguido los mismos pasos empleados en la primera parte de los ensayos (cf. 4.2).

Se presentan a continuación algunos resultados de este proceso de construcción de la base de datos.

Base 0:

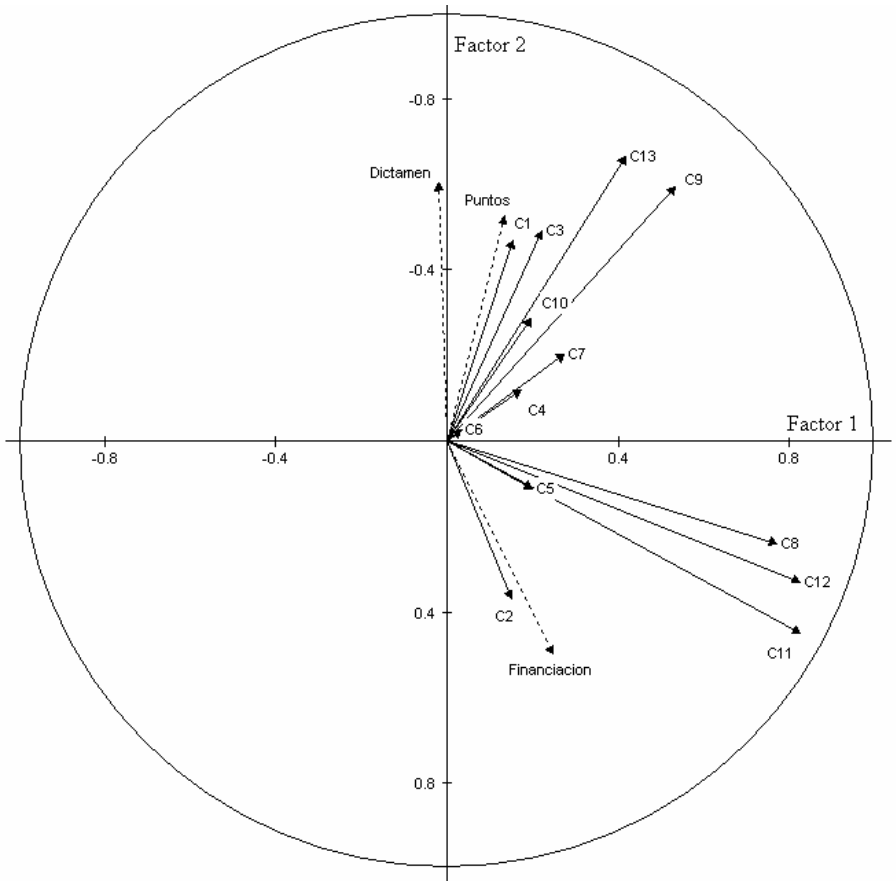
VARIABLES	CORRELATIONS VARIABLE-FACTEUR								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
C2 - Dictamen	-0.02	-0.60	-0.22	-0.06	0.10	-0.40	-0.53	-0.28	-0.13
C3 - Puntos	0.14	-0.53	-0.12	0.19	0.24	-0.13	-0.33	-0.64	-0.21
C4 - Financiacion	0.25	0.50	0.17	-0.04	0.70	0.15	0.12	0.28	-0.08

MODALITES	VALEURS-TEST								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
4 . Dictamen_d									
AA_1 - C5=1	1.7	32.7	13.7	4.4	-4.2	20.4	31.3	16.3	4.1
AA_2 - C5=2	-1.7	-32.7	-13.7	-4.4	4.2	-20.4	-31.3	-16.3	-4.1
5 . Puntos_d									
AB_1 - C6=1	0.4	21.7	2.5	-5.1	-12.4	6.4	22.1	33.6	10.4
AB_2 - C6=2	-8.9	10.9	8.0	-2.2	-3.3	0.0	-0.1	7.3	-2.2
AB_3 - C6=3	-4.3	-1.6	2.0	-0.6	2.6	-1.4	-4.9	-6.5	-1.7
AB_4 - C6=4	-0.4	-8.3	-3.5	-0.1	2.3	-0.5	-6.2	-11.9	0.5
AB_5 - C6=5	6.7	-15.2	-6.5	1.8	8.8	-2.5	-9.7	-18.3	-1.2
AB_6 - C6=6	14.2	-26.0	-9.7	11.8	9.7	-5.3	-11.0	-24.4	-10.7
6 . Financiacion_d									
AC_1 - C7=1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AC_2 - C7=2	-7.5	-21.3	-11.1	4.8	-33.8	-6.9	-3.4	-11.8	6.1
AC_3 - C7=3	-13.2	-17.9	-5.1	3.5	-28.1	-6.8	0.2	-7.7	4.3
AC_4 - C7=4	2.0	0.5	2.3	-6.0	6.8	2.4	-8.9	-2.7	-5.7
AC_5 - C7=5	10.4	24.4	8.7	-0.3	34.5	7.2	9.0	14.4	-1.6
AC_6 - C7=6	13.9	9.8	-2.9	3.9	4.8	-0.2	7.3	6.1	1.4

Se puede ver en el reporte anterior que la capacidad predictiva es intermedia, tal como acontece en muchas situaciones reales. Esto se puede apreciar en los valores Test y las correlaciones Variable-Factor que se muestran.

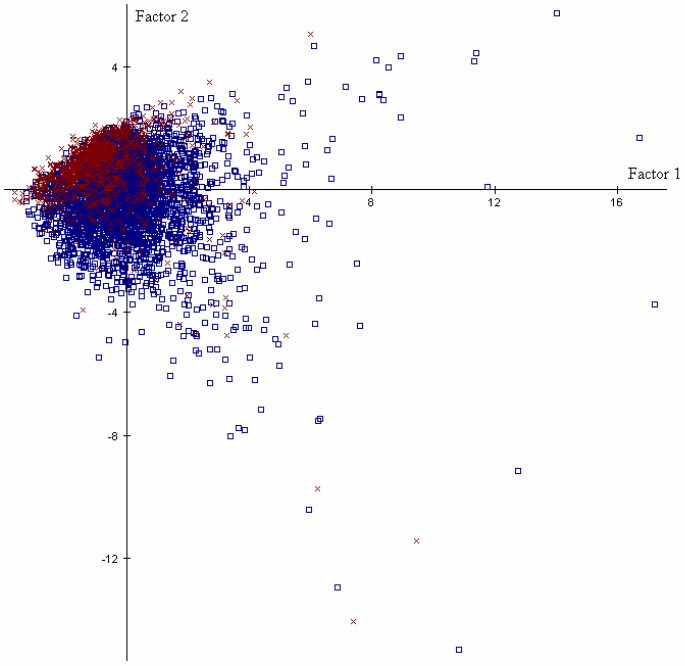
Tabla 4.5 Correlaciones de las variables activas con los factores

Variable	Eje 1	Eje 2	Eje 3	Eje 4	Eje 5	Eje 6	Eje 7	Eje 8	Eje 9
C1	0.15	-0.47	-0.54	-0.22	0.33	0.04	-0.13	-0.19	0.22
C2	0.15	0.37	0.24	-0.15	0.8	0.19	-0.06	0.15	-0.17
C3	0.22	-0.49	-0.54	-0.22	0.18	0.11	0.18	0.14	0.14
C4	0.17	-0.12	-0.18	0.3	-0.2	0.55	-0.64	0.27	-0.09
C5	0.2	0.11	-0.02	0.2	-0.08	0.72	0.55	-0.26	-0.03
C6	0.03	-0.03	-0.23	0.7	0.17	-0.22	0.32	0.49	0.07
C7	0.27	-0.2	-0.1	0.6	0.24	-0.22	-0.16	-0.54	-0.25
C8	0.77	0.24	-0.16	-0.01	-0.33	-0.17	0.02	-0.03	0.16
C9	0.53	-0.59	0.27	-0.09	-0.09	-0.05	0.1	0.11	-0.3
C10	0.2	-0.28	0.62	0.23	0.14	0.09	-0.11	-0.07	0.62
C11	0.83	0.45	-0.02	-0.08	0.14	-0.03	-0.02	0.06	0.02
C12	0.83	0.33	-0.06	-0.06	-0.07	-0.08	-0.06	-0.02	0.03
C13	0.42	-0.66	0.38	-0.11	-0.02	-0.01	0.1	0.13	-0.18



____Variables comunes (activas) _ _ _Variables específicas (ilustrativas)

Figura 4.15 Correlaciones de las variables activas e ilustrativas con los factores (Base 0)



□ Individuos activos (donantes) x Individuos ilustrativos (receptores)

Figura 4.16 Espacio común de los individuos en el primer plano factorial

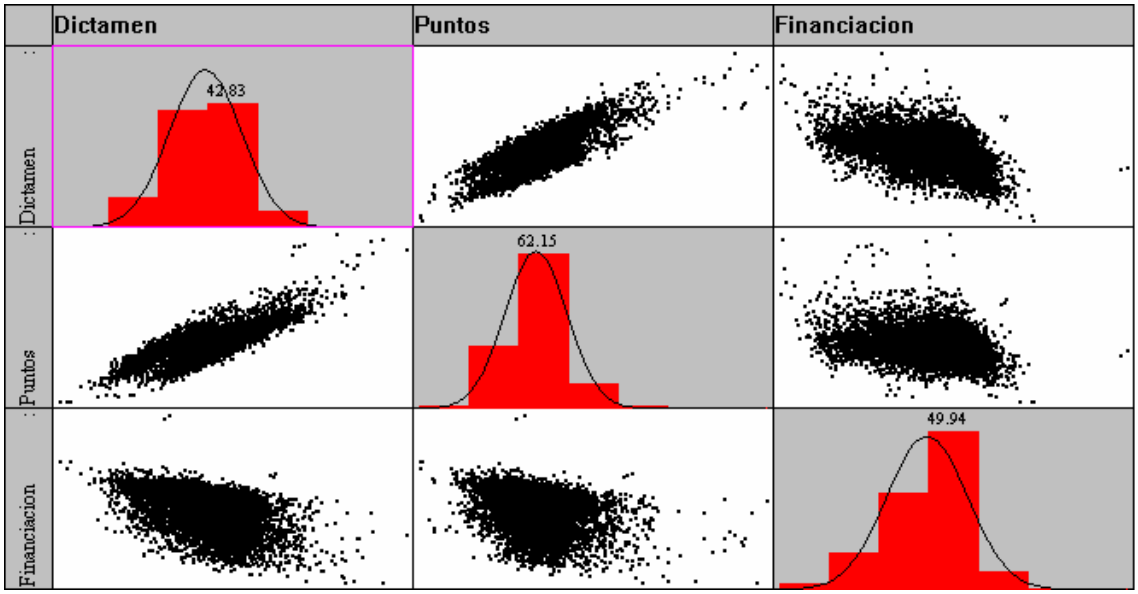


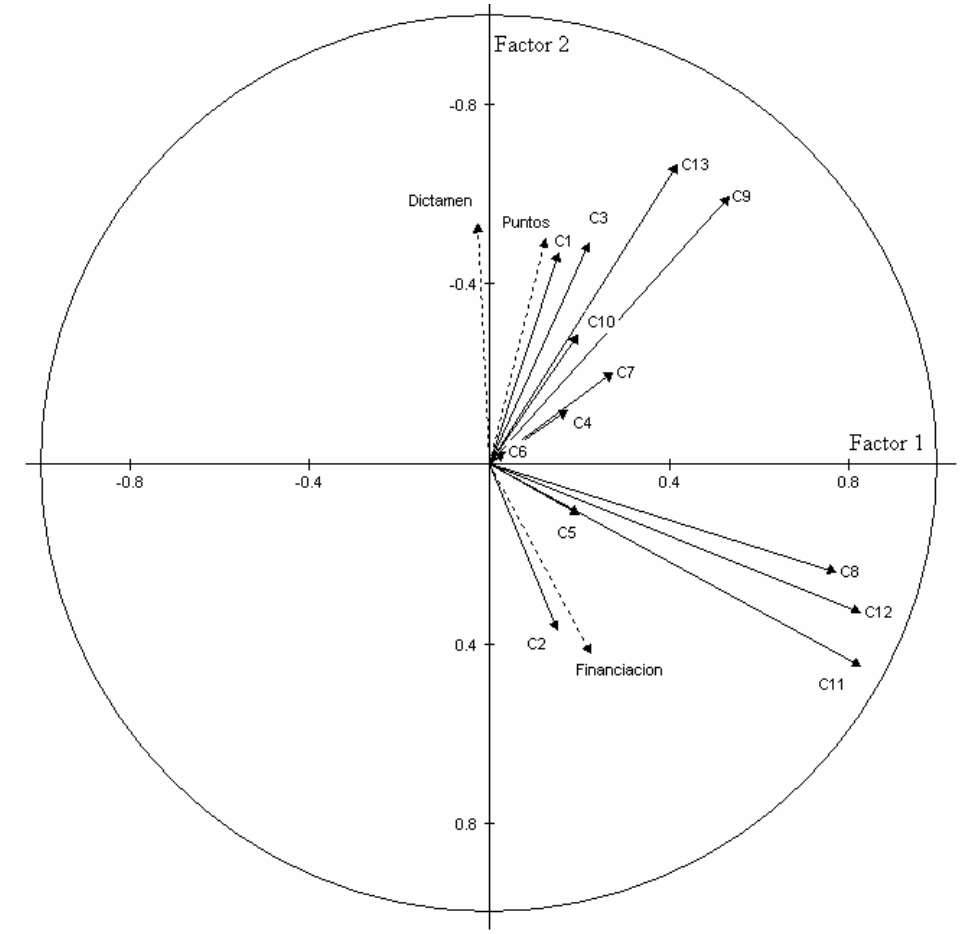
Figura 4.17 Características de las variables específicas (Base 0)

Base 1 :

VARIABLES	CORRELATIONS VARIABLE-FACTEUR								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
C2 - Dictamen	-0.03	-0.53	-0.20	-0.06	0.08	-0.36	-0.46	-0.27	-0.11
C3 - Puntos	0.13	-0.50	-0.11	0.18	0.22	-0.11	-0.32	-0.62	-0.21
C4 - Financiacion	0.22	0.42	0.14	-0.03	0.60	0.13	0.11	0.23	-0.07

MODALITES	VALEURS-TEST								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
4 . Dictamen_d									
AA_1 - C5=1	0.2	28.4	11.3	3.6	-4.7	19.4	27.6	14.1	3.9
AA_2 - C5=2	-0.2	-28.4	-11.3	-3.6	4.7	-19.4	-27.6	-14.1	-3.9
5 . Puntos_d									
AB_1 - C6=1	0.2	21.6	3.8	-5.8	-10.4	5.6	19.8	32.5	9.3
AB_2 - C6=2	-8.3	9.1	5.9	-0.4	-2.8	-0.2	1.1	6.1	-0.7
AB_3 - C6=3	-3.7	-1.3	1.9	-2.3	1.4	-0.7	-3.3	-6.0	-1.0
AB_4 - C6=4	-0.3	-5.9	-3.7	0.8	2.1	-1.0	-6.4	-9.5	-0.6
AB_5 - C6=5	4.0	-14.4	-4.8	1.0	6.4	-1.1	-10.0	-17.6	-2.6
AB_6 - C6=6	14.6	-25.6	-8.8	11.5	10.1	-5.7	-10.6	-24.2	-9.3
6 . Financiacion_d									
AC_1 - C7=1	-0.2	-4.8	-2.5	2.2	-6.2	-2.0	-2.2	-2.6	-0.7
AC_2 - C7=2	-9.1	-18.6	-8.5	3.0	-31.9	-6.3	-2.7	-11.6	5.8
AC_3 - C7=3	-8.3	-14.9	-4.7	0.5	-21.0	-4.2	-2.1	-5.2	3.5
AC_4 - C7=4	0.1	2.2	3.5	-2.1	7.8	-0.2	-4.2	0.1	-5.4
AC_5 - C7=5	7.8	16.6	5.7	-1.4	25.6	6.2	4.2	7.1	-1.5
AC_6 - C7=6	9.8	12.5	1.0	2.1	10.7	3.7	7.5	9.7	0.8

Al igual que en los ensayos anteriores (Simulación I), se puede ver en el reporte anterior que la capacidad predictiva es intermedia.



____Variables comunes (activas) _ _ _Variables específicas (ilustrativas)

Figura 4.18 Correlaciones de las variables activas e ilustrativas con los factores (Base 1)

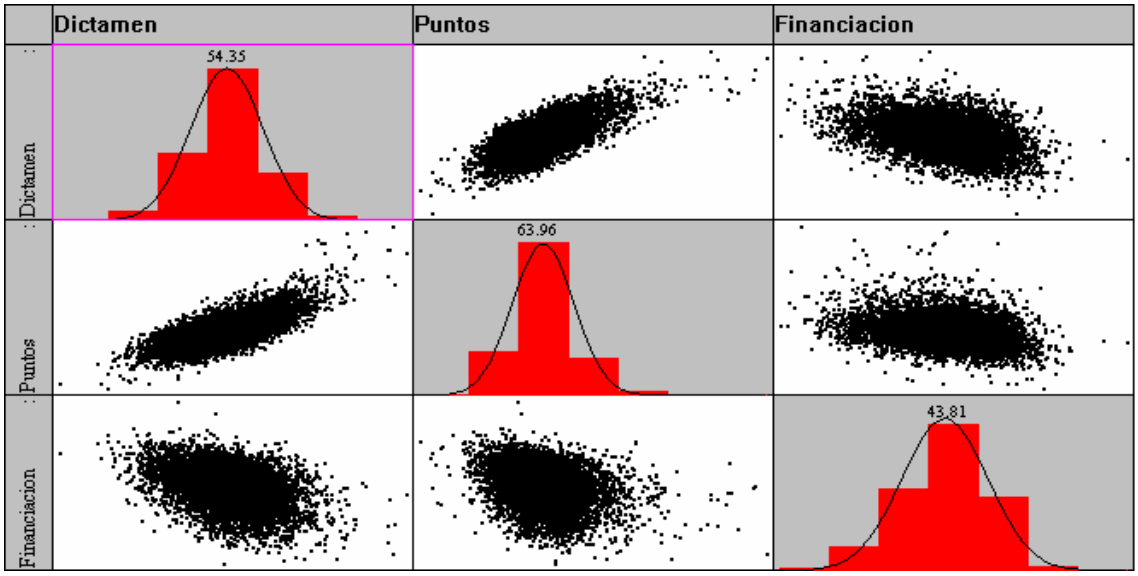


Figura 4.19 Características de las variables específicas (Base 1)

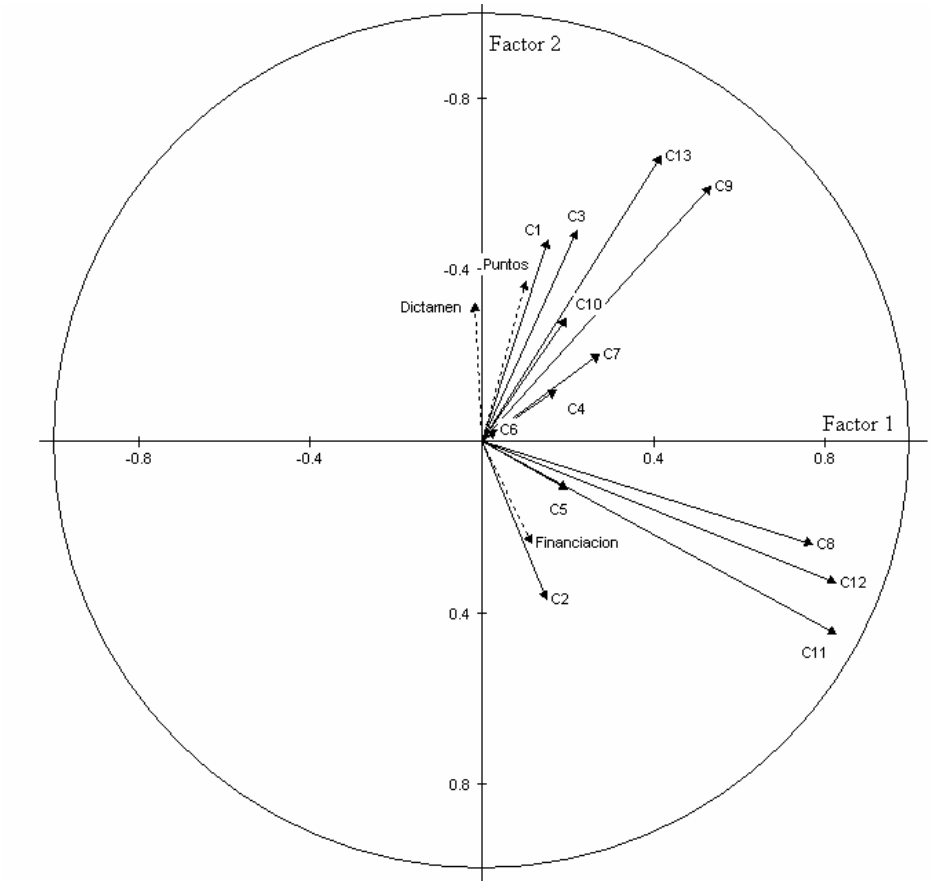
Base 2 :

VARIABLES	CORRELATIONS VARIABLE-FACTEUR								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
C2 - Dictamen	-0.02	-0.32	-0.12	-0.02	0.05	-0.22	-0.28	-0.15	-0.05
C3 - Puntos	0.10	-0.37	-0.08	0.13	0.17	-0.09	-0.25	-0.45	-0.15
C4 - Financiacion	0.12	0.24	0.07	0.00	0.32	0.08	0.05	0.11	-0.04

MODALITES	VALEURS-TEST								
IDEN - LIBELLE	1	2	3	4	5	6	7	8	9
4 . Dictamen_d									
AA_1 - C5=1	1.3	17.5	7.5	1.1	-2.4	12.7	16.2	8.0	3.0
AA_2 - C5=2	-1.3	-17.5	-7.5	-1.1	2.4	-12.7	-16.2	-8.0	-3.0
5 . Puntos_d									
AB_1 - C6=1	-1.8	17.5	4.0	-3.8	-7.9	4.0	15.5	23.1	5.9
AB_2 - C6=2	-3.3	3.9	2.6	-2.2	-2.1	1.2	-0.4	4.1	0.6
AB_3 - C6=3	-2.6	0.4	-0.5	-0.5	0.2	-1.2	-1.6	-1.4	0.2
AB_4 - C6=4	-2.8	-1.4	0.3	-0.1	1.1	-0.1	-3.6	-4.3	-0.9
AB_5 - C6=5	0.1	-8.5	-1.6	1.0	3.9	-0.9	-5.8	-10.1	-0.1
AB_6 - C6=6	10.9	-19.3	-6.7	7.3	8.3	-4.8	-10.6	-21.7	-8.1
6 . Financiacion_d									
AC_1 - C7=1	-3.6	-10.1	-4.3	-2.0	-12.0	-2.5	-1.2	-4.2	2.3
AC_2 - C7=2	-6.2	-10.1	-2.8	1.4	-15.5	-4.7	-1.2	-4.2	0.9
AC_3 - C7=3	-0.9	-2.4	-1.1	0.0	-3.6	-2.1	-1.0	-1.3	0.0
AC_4 - C7=4	-0.6	1.1	1.0	-0.3	3.1	1.5	-2.0	-0.6	0.0
AC_5 - C7=5	2.3	3.9	3.1	1.3	7.1	2.0	0.5	1.5	0.3
AC_6 - C7=6	6.5	11.2	1.8	-1.2	13.3	4.0	3.9	6.1	-2.1

Se puede observar que la Base 2 indica un bajo poder predictivo, como sucedió en los ensayos anteriores (Simulación I) de las variables específicas, augurando una fusión de datos problemática.

Los resultados relacionados con la capacidad predictiva de las variables comunes sobre las variables específicas, resultaron similares a los obtenidos en la Simulación I, aunque tal vez en este caso sean un poco inferiores.



____Variables comunes (activas) _ _ _Variables específicas (ilustrativas)

Figura 4.20 Correlaciones de las variables activas e ilustrativas con los factores (Base 2)

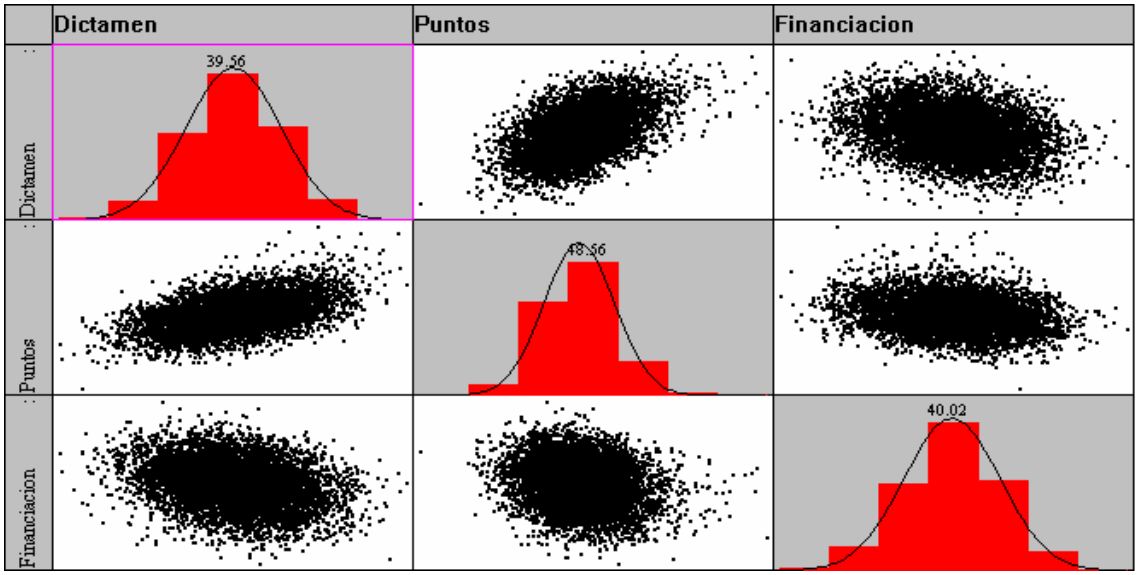


Figura 4.21 Características de las variables específicas (Base 2)

4.5 Ensayos (Simulación II)

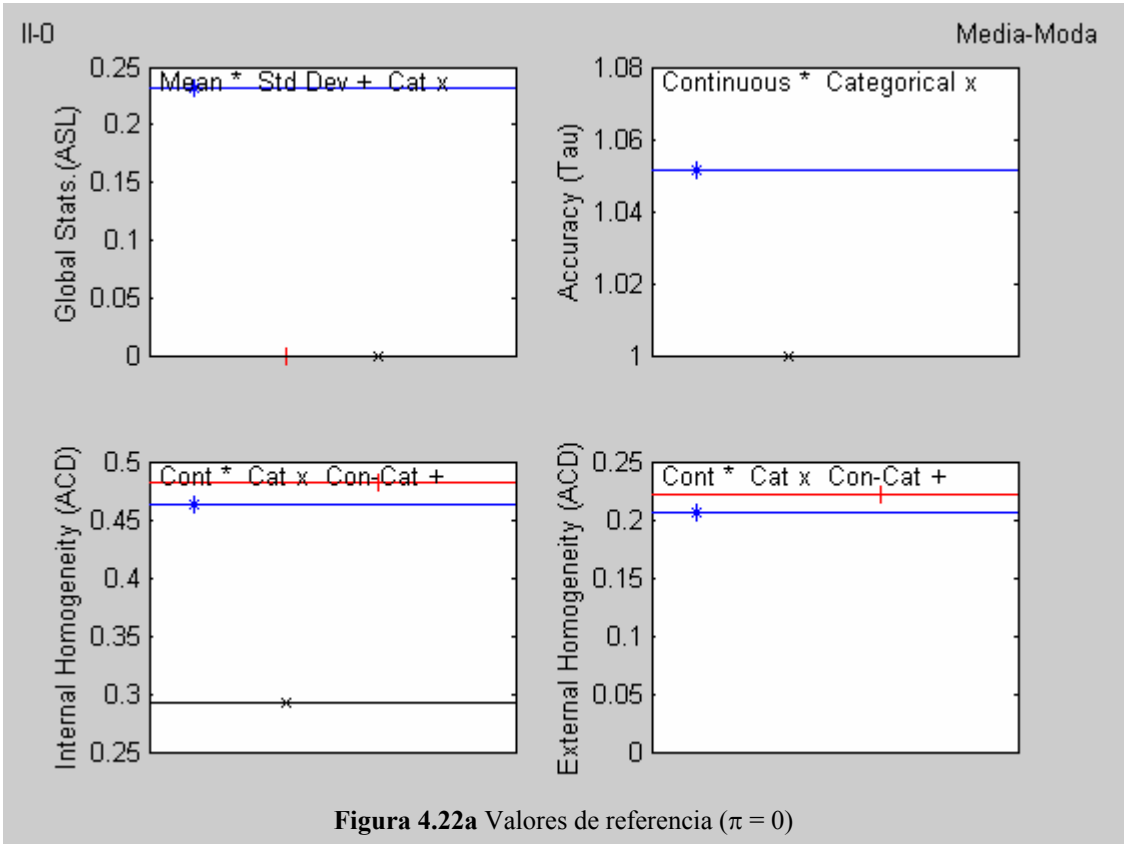


Figura 4.22a Valores de referencia ($\pi = 0$)

Omitir la gráficas de homogeneidad.

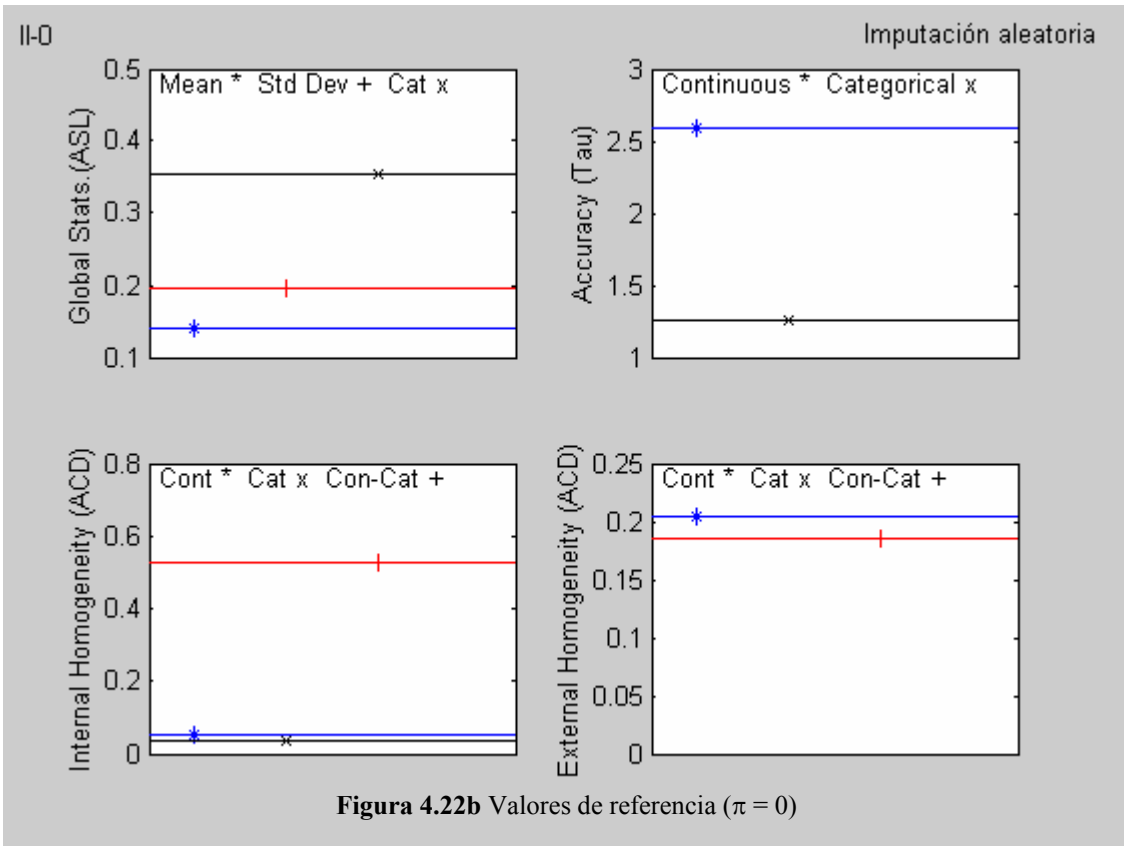


Figura 4.22b Valores de referencia ($\pi = 0$)

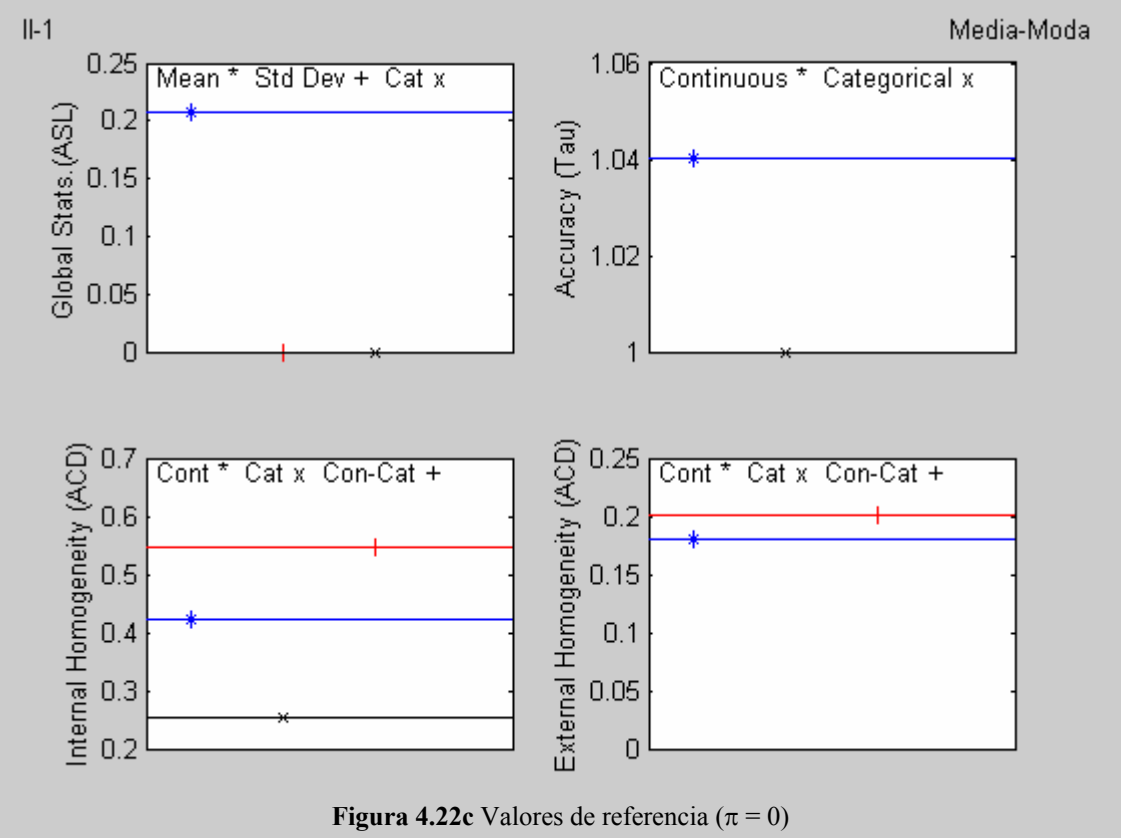


Figura 4.22c Valores de referencia ($\pi = 0$)

Omitir la gráficas de homogeneidad.

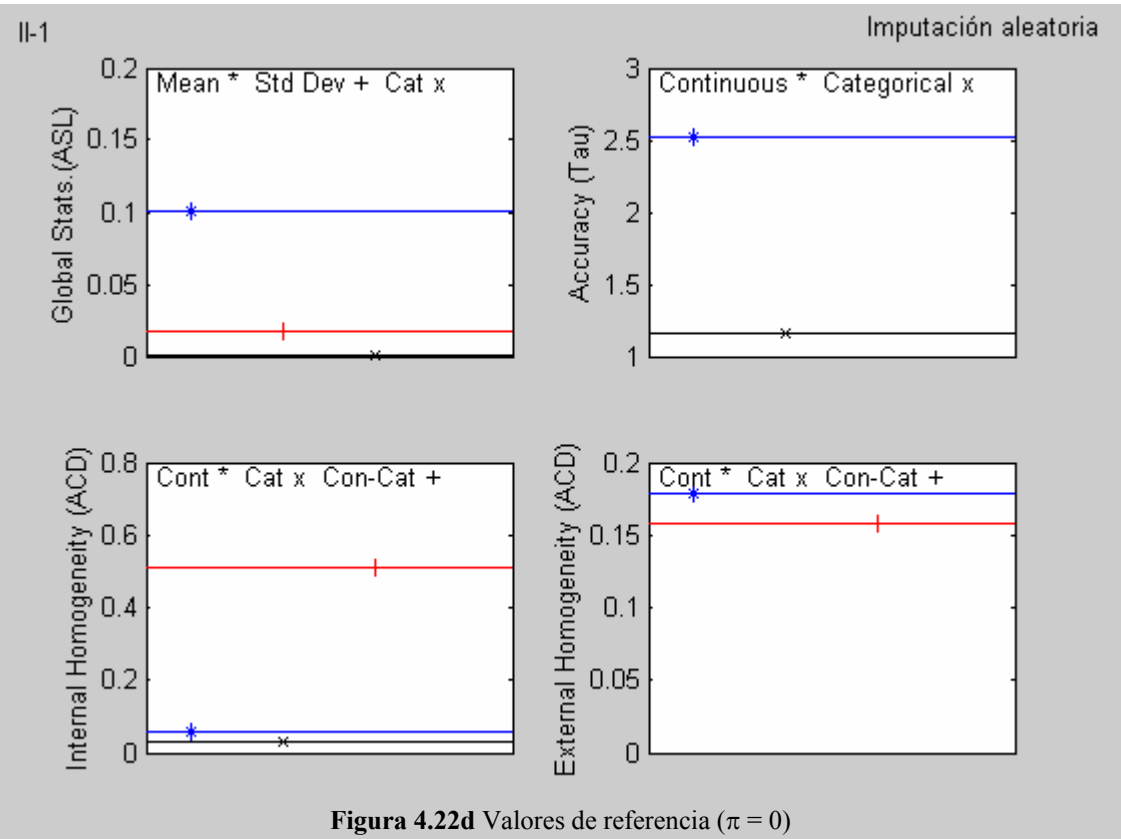


Figura 4.22d Valores de referencia ($\pi = 0$)

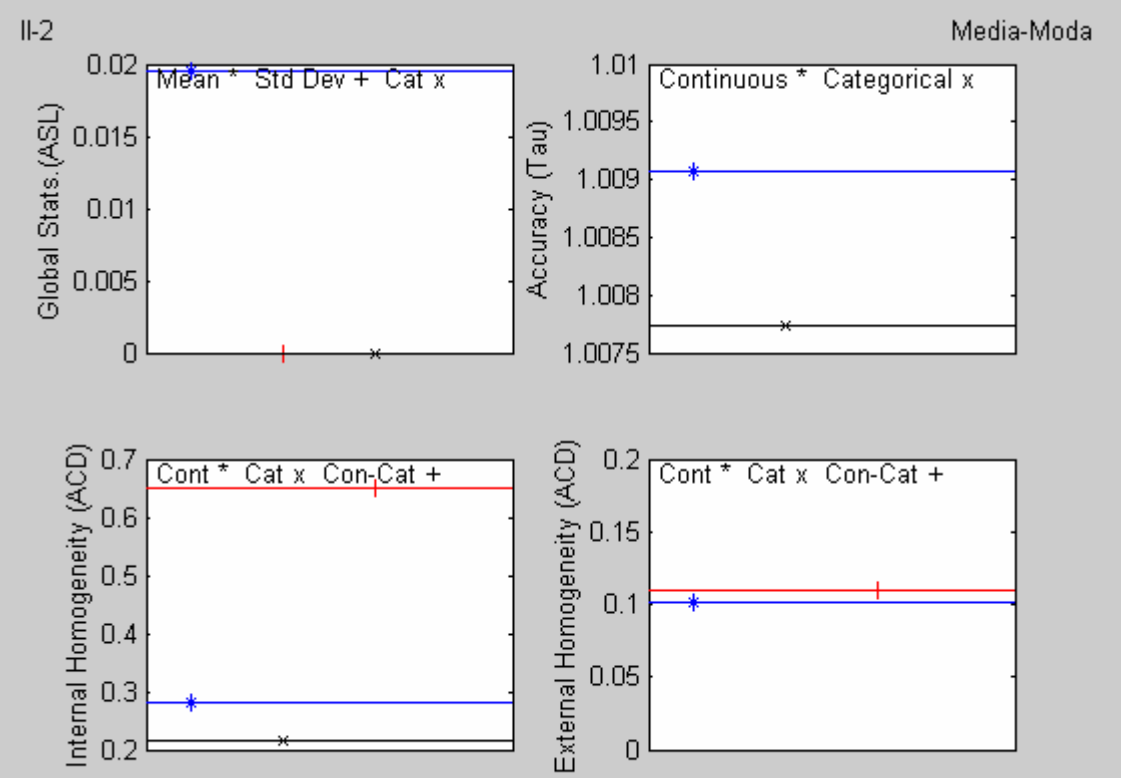


Figura 4.22e Valores de referencia ($\pi = 0$)

Omitir la gráficas de homogeneidad.

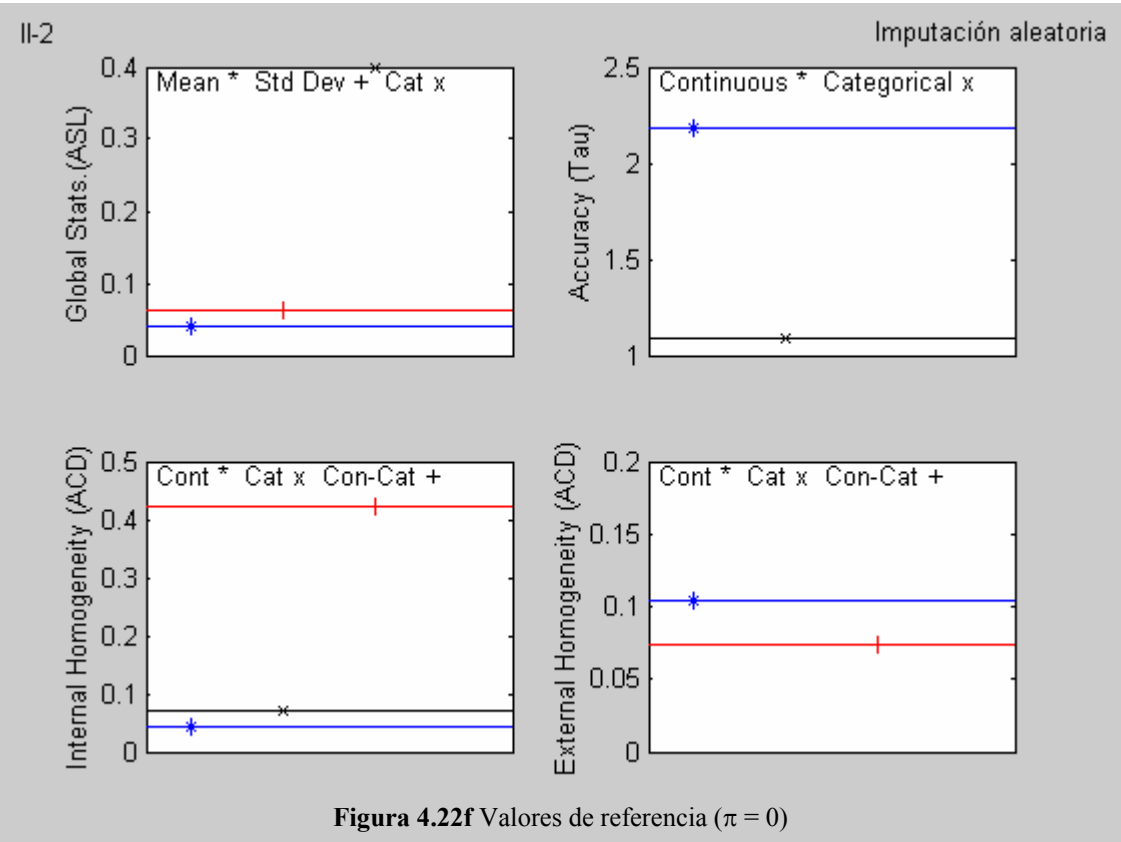


Figura 4.22f Valores de referencia ($\pi = 0$)

4.5.1 Ensayos T1DM

Ensayo 2.1

Procedimiento de imputación:	T1DM (Take One Deterministic Multivariate)
Parámetro:	P (probabilidad de repetición de los donantes)
Valor del Parámetro:	P = 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0
Archivos:	Base0, Base 1, Base 2

Los valores de referencia (imputación empleando la media –moda-) se pueden observar en las figuras 4.22a, 4.22c y 4.22e.

4.5.1.1Resultados

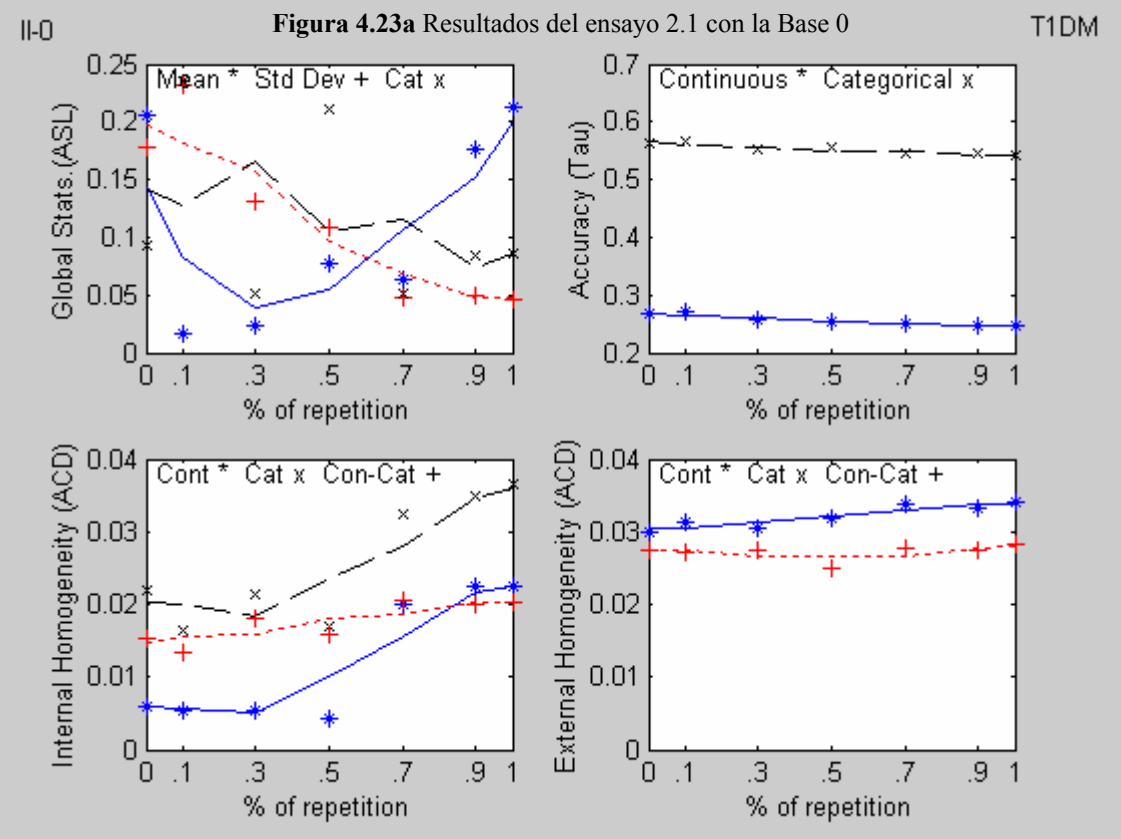


Figura 4.23a Resultados del ensayo 2.1 con la Base 0

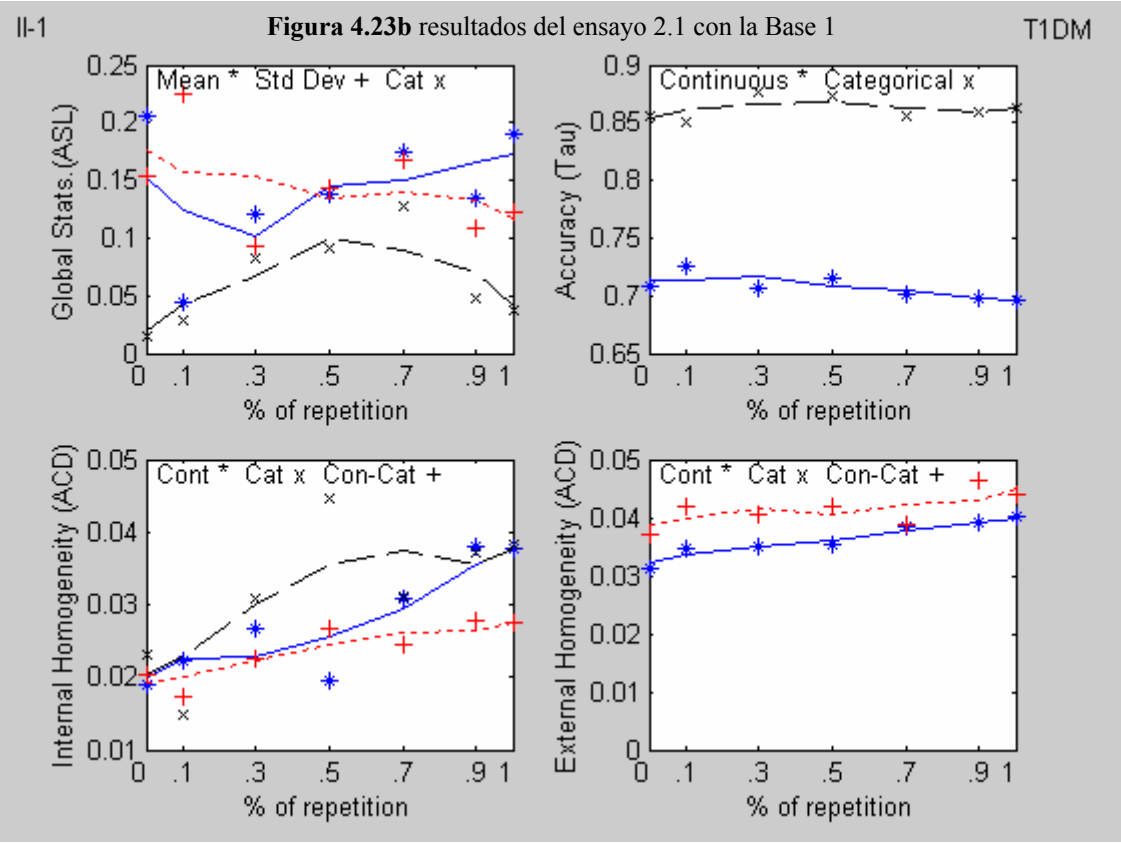


Figura 4.23b resultados del ensayo 2.1 con la Base 1

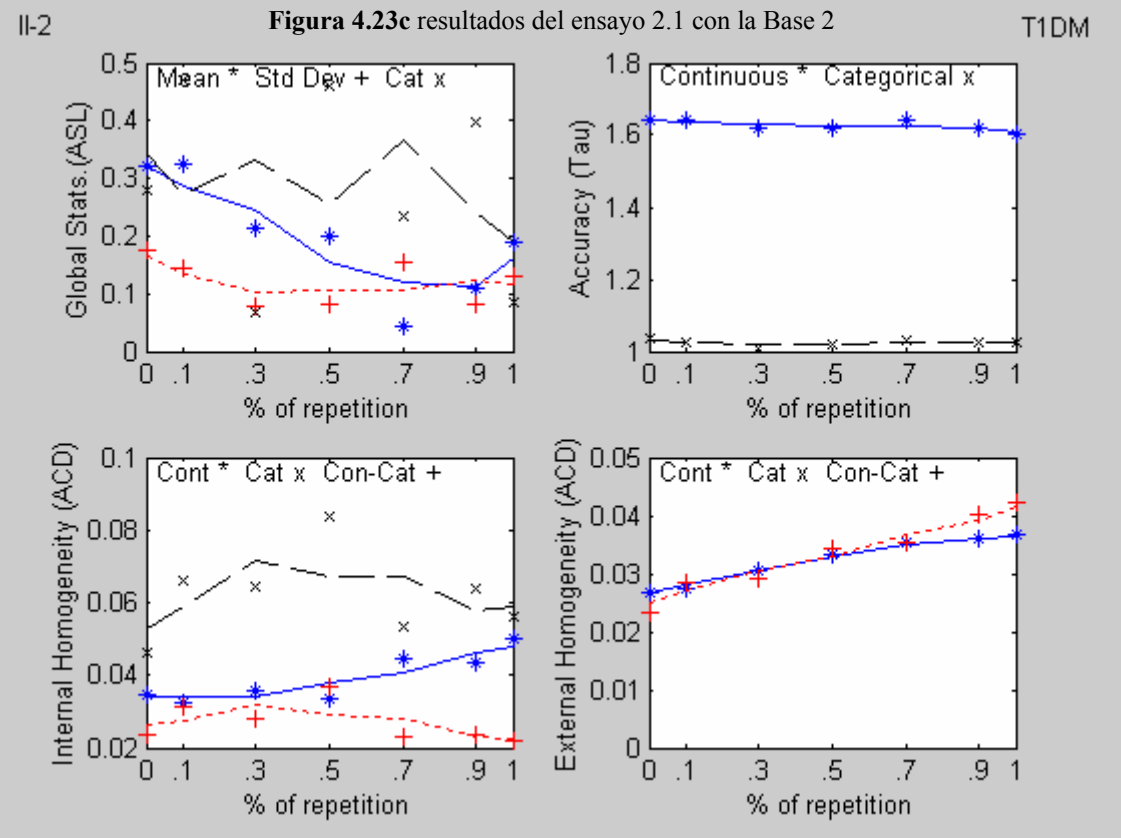


Figura 4.23c resultados del ensayo 2.1 con la Base 2

4.5.1.2 Análisis de los resultados

- ASL.- La reproducción de la media para los casos de la varianza residual normal y reducida, presenta una tendencia creciente para $P > 0$. En el caso del ruido aleatorio amplificado, el comportamiento que se observa es decreciente; se reproduce mejor la desviación estándar y solo en el caso de la varianza residual reducida se observa una tendencia decreciente. En el caso de la varianza residual normal, la reproducción categórica presenta resultados buenos en los valores intermedios de P.

- Tau.- Para los casos de la varianza residual normal y reducida, los valores obtenidos del índice (continuo y categórico) son buenos y no parecen depender del parámetro P. En el caso del ruido aleatorio amplificado, la imputación por la media, comparada con este método resulta mejor opción.

- ACD (Interna).- Los valores obtenidos del índice son buenos y en el caso continuo, se observa una tendencia creciente con el parámetro P. El caso categórico y continuo-categórico parecen no verse influenciados por el valor de P en el caso del ruido aleatorio amplificado.

- ACD (Externa).- Los valores obtenidos del índice son buenos y en los casos de la varianza residual normal y reducida, parece no tener efecto el parámetro P. En el caso del ruido aleatorio amplificado, se observa una ligera tendencia creciente con el parámetro P.

4.5.2 Ensayos T1SM

Ensayo 2.2

Procedimiento de imputación: T1SM (Take One Stochastic Multivariate)

Parámetro: Ninguno

7 repeticiones

Archivos: Base0, Base 1, Base 2

Los valores de referencia (imputación aleatoria) se pueden observar en las figuras 4.22b, 4.22d y 4.22f.

4.5.2.1 Resultados

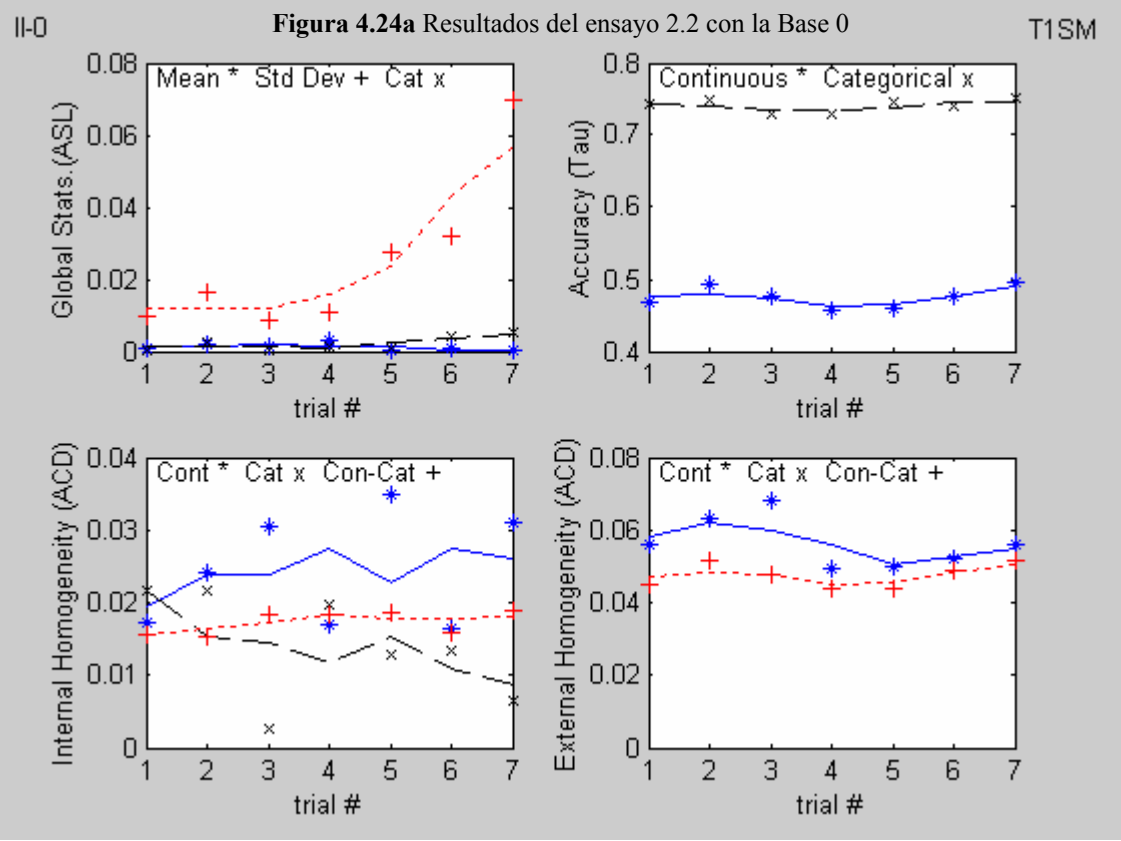


Figura 4.24a Resultados del ensayo 2.2 con la Base 0

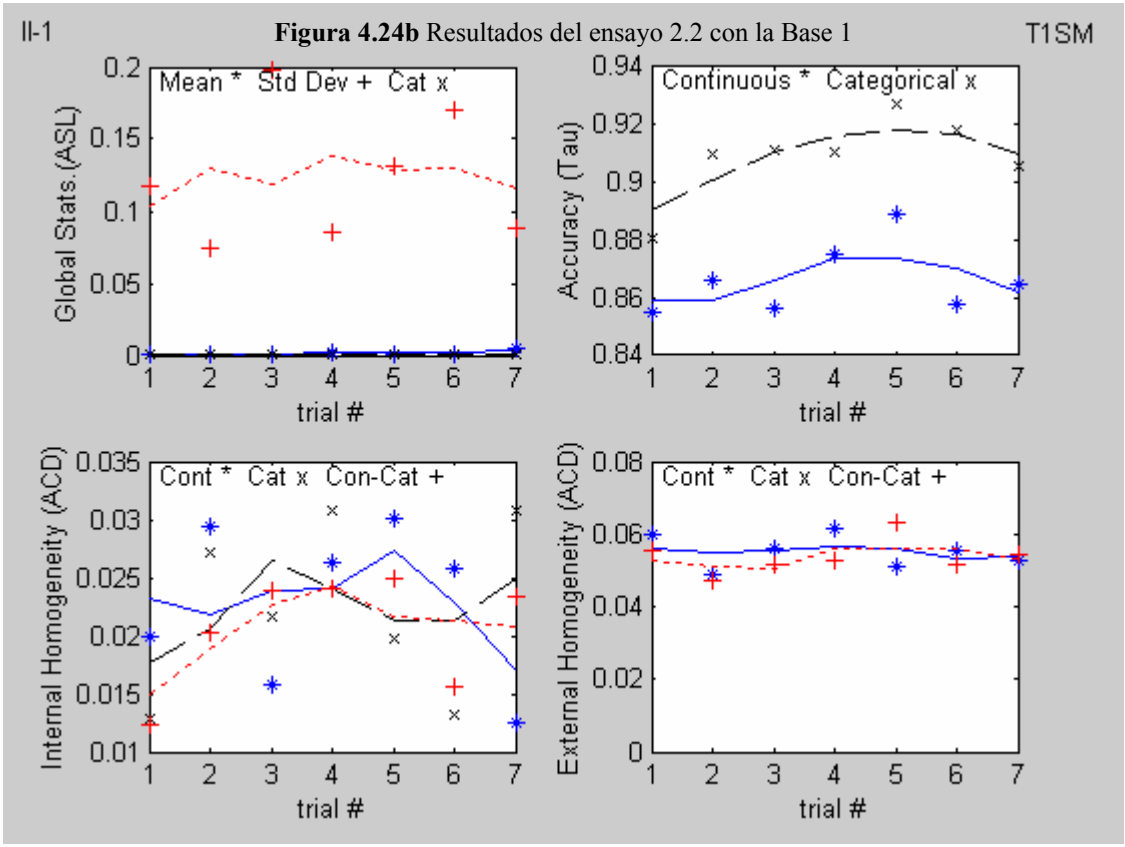


Figura 4.24b Resultados del ensayo 2.2 con la Base 1

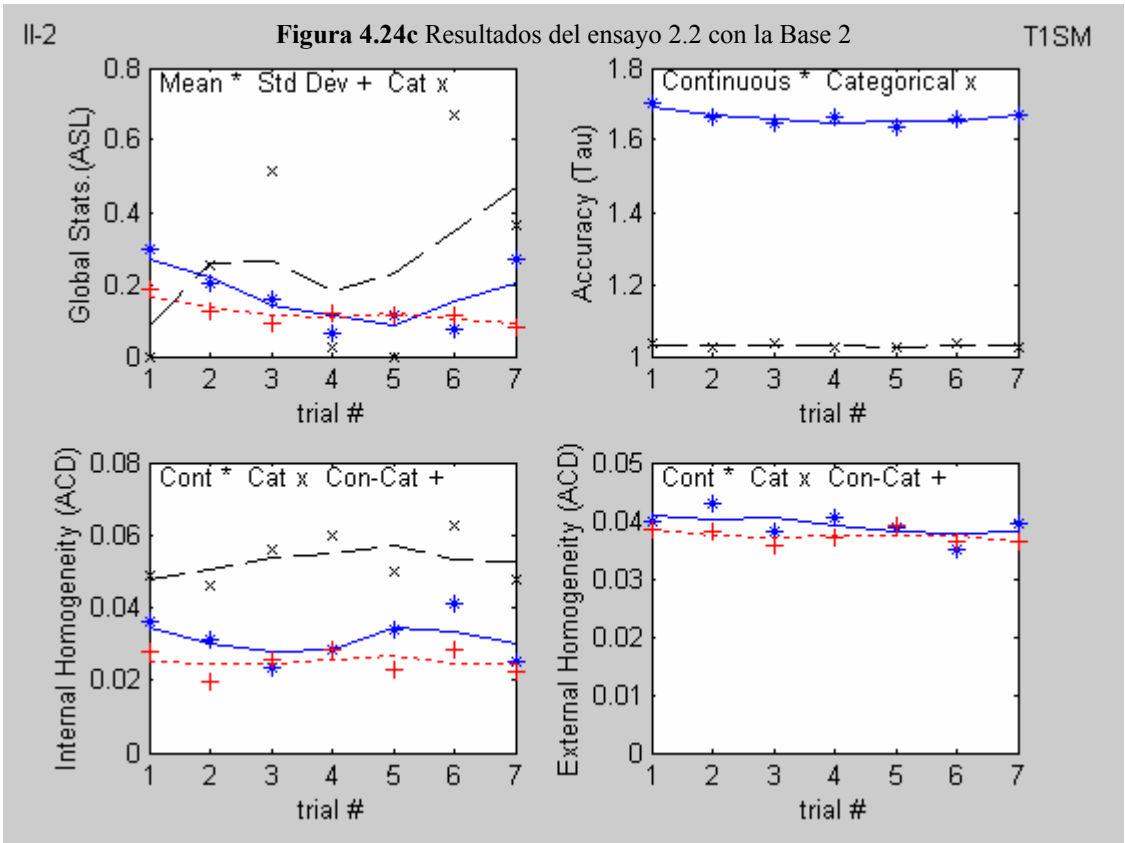


Figura 4.24c Resultados del ensayo 2.2 con la Base 2

4.5.2.2 Análisis de los resultados

- ASL.- Para los casos de la varianza residual normal y reducida, no se reproduce bien la media, no así en el caso del ruido aleatorio amplificado. La desviación estándar no se reproduce bien en el caso de la varianza residual reducida. La reproducción categórica solo presenta buenos resultados en el caso de la varianza residual amplificada.

- Tau.- Para los casos de la varianza residual normal y reducida, los valores obtenidos del índice (continuo y categórico) son buenos en promedio. En el caso del ruido aleatorio amplificado, la imputación por la media, comparada con este método resulta mejor opción.

- ACD (Interna).- No parece tener influencia el incremento de fluctuación aleatoria en los resultados, solo en el caso categórico, con el ruido aleatorio amplificado.

- ACD (Externa).- No parece tener influencia el incremento de fluctuación aleatoria en los resultados. Nuevamente se observa que los resultados obtenidos con el método T1DM son ligeramente mejores.

4.5.3 Ensayos TKDM

Ensayo 2.3.

Procedimiento de imputación: TKDM (Take K Deterministic Multivariate)

Parámetro: K (número de vecinos)

Valor del Parámetro: K = 1, 5, 10, 15, 20, 25, 30, 35, 40

Archivos: Base1, Base 2, Base 3

Los valores de referencia (imputación empleando la media –moda-) se pueden observar en las figuras 4.22a, 4.22c y 4.22e.

4.5.3.1 Resultados:

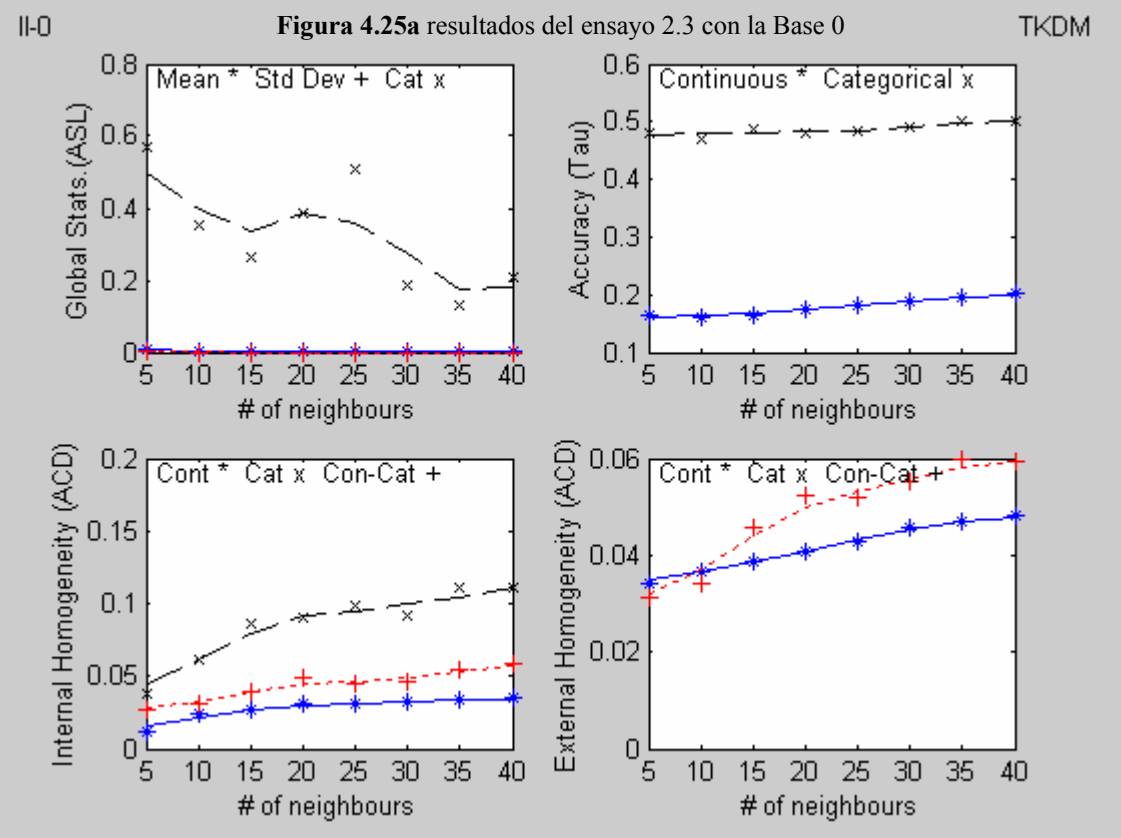


Figura 4.25a resultados del ensayo 2.3 con la Base 0

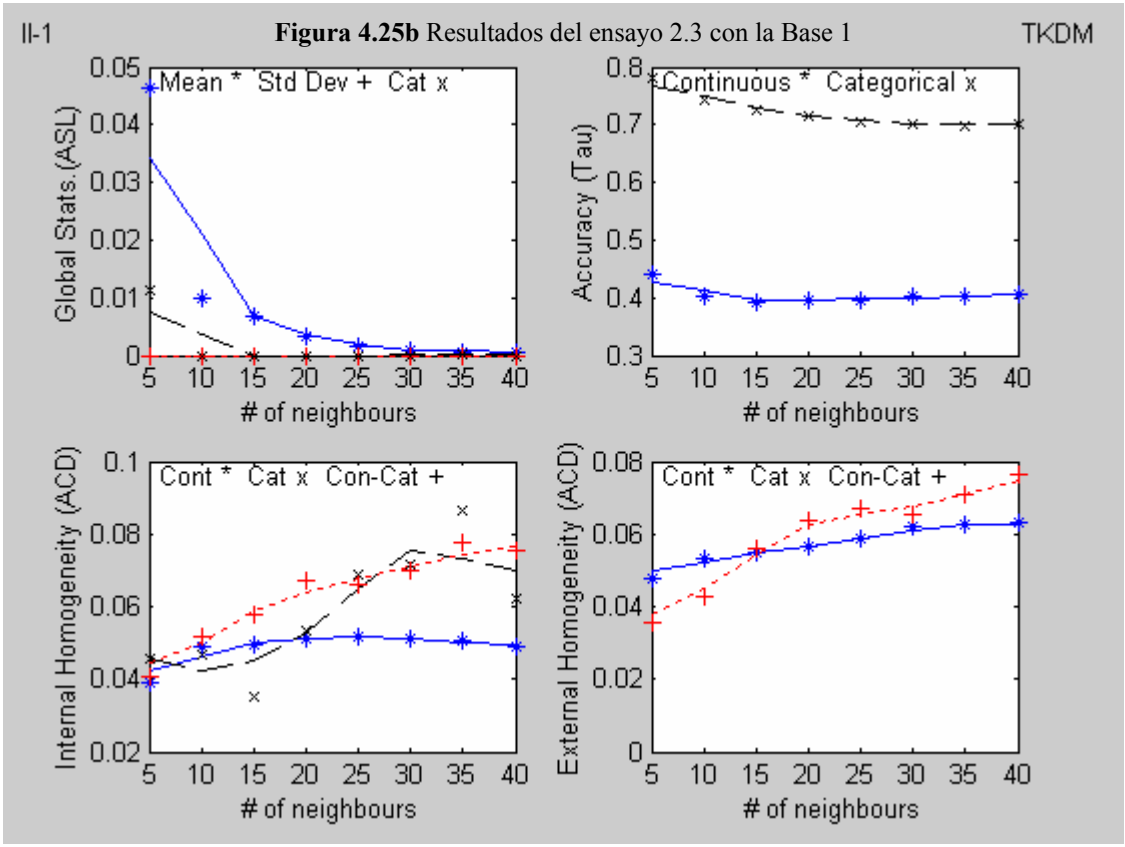


Figura 4.25b Resultados del ensayo 2.3 con la Base 1

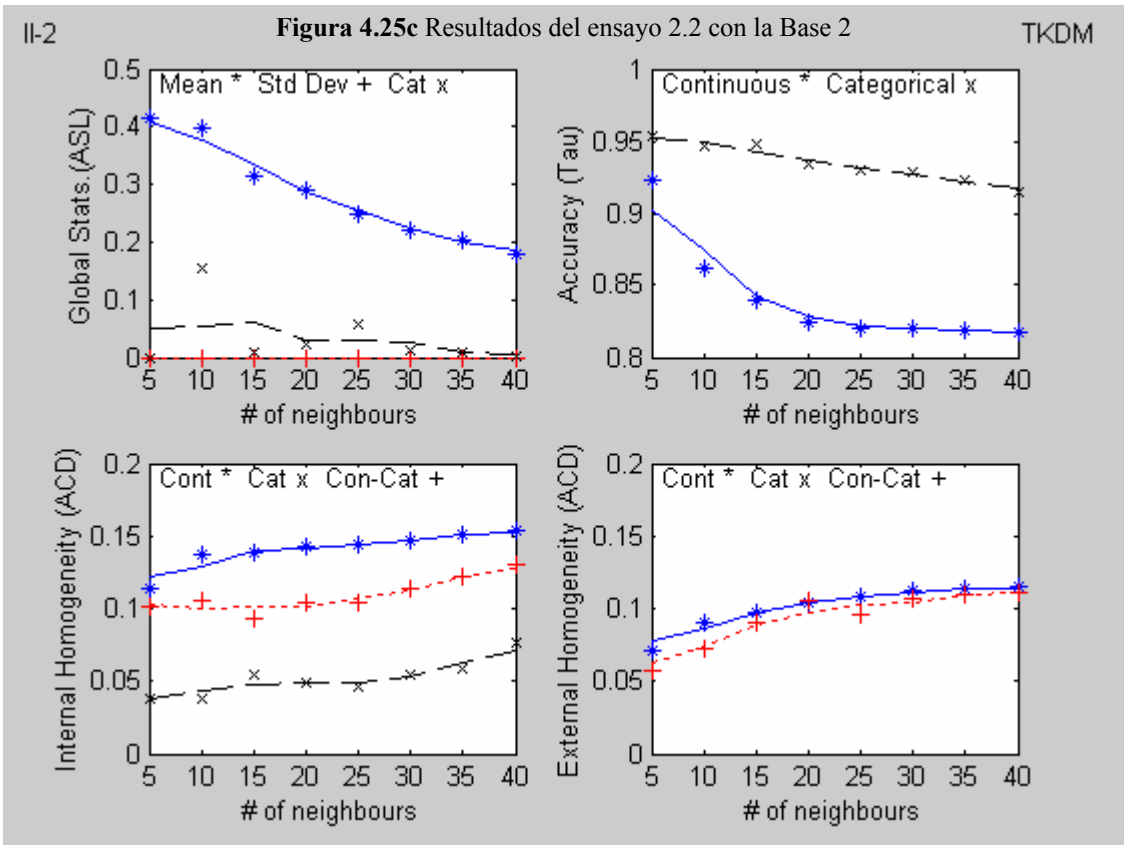


Figura 4.25c Resultados del ensayo 2.2 con la Base 2

4.5.3.2 Análisis de los resultados

- ASL (caso continuo).- Solo se obtienen buenos resultados para la reproducción de la media en el caso del ruido aleatorio amplificado con una tendencia a empeorar con el incremento de K. En el caso categórico se obtienen buenos resultados para la varianza residual reducida. La desviación estándar no se reproduce bien en ningún caso.

- Tau (caso continuo).- Para los casos de la varianza residual normal y reducida, los valores obtenidos del índice (continuo y categórico) son buenos y no parecen depender del parámetro K. En el caso del ruido aleatorio amplificado, la imputación mejora incrementando el valor de K.

- ACD (Interna).- Los valores obtenidos del índice son buenos en todos los casos de la varianza residual; se observa una ligera tendencia creciente con el parámetro K. Los valores aumentan en general con la varianza residual.

- ACD (Externa).- Los valores obtenidos del índice son buenos en los casos de la varianza residual normal y reducida, no tanto así en el caso del ruido aleatorio amplificado. Se observa una ligera tendencia creciente con el parámetro K. Los valores aumentan en general con la varianza residual.

4.5.4 Ensayos TKSM

Ensayo 2.4.

Procedimiento de imputación: TKSM (Take K Stochastic Multivariate)

Parámetro: K (número de vecinos)

Valor del Parámetro: K = 5, 10, 15, 20, 25, 30, 35, 40

Archivos: Base0, Base 1, Base 2

Los valores de referencia (imputación aleatoria) se pueden observar en las figuras 4.22b, 4.22d y 4.22f.

4.5.4.1 Resultados:

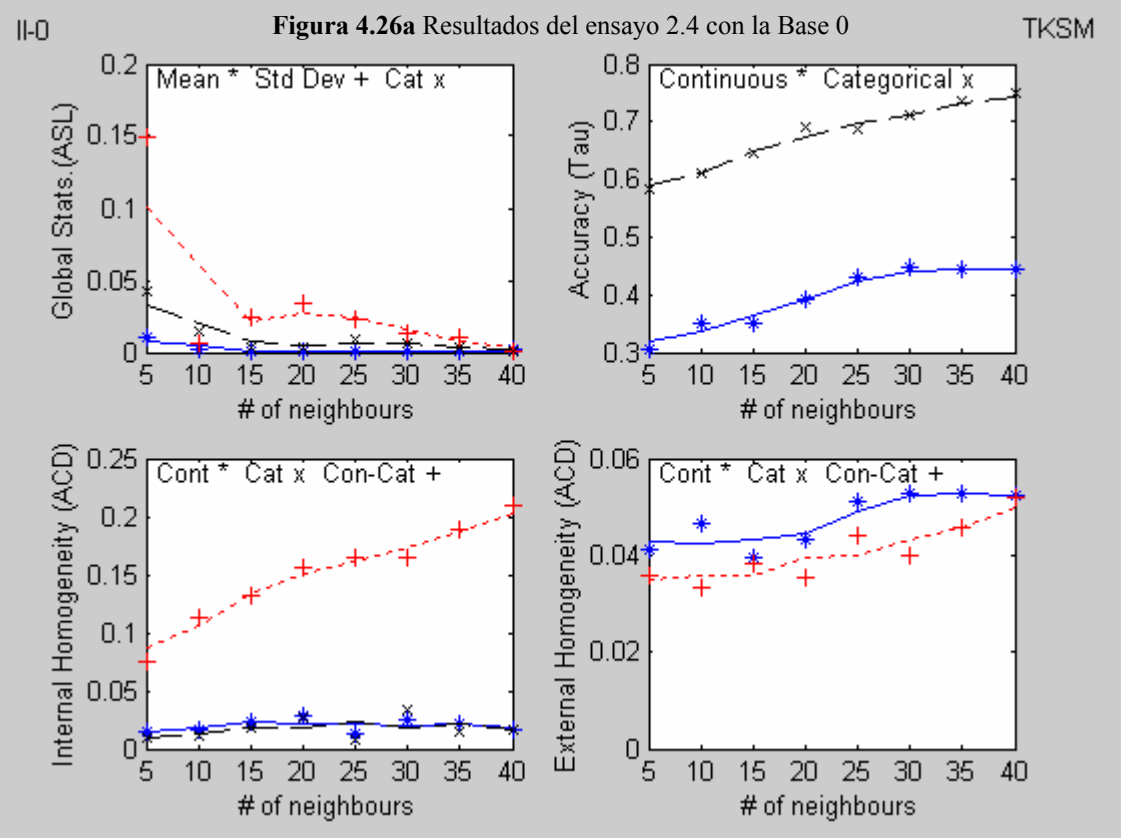
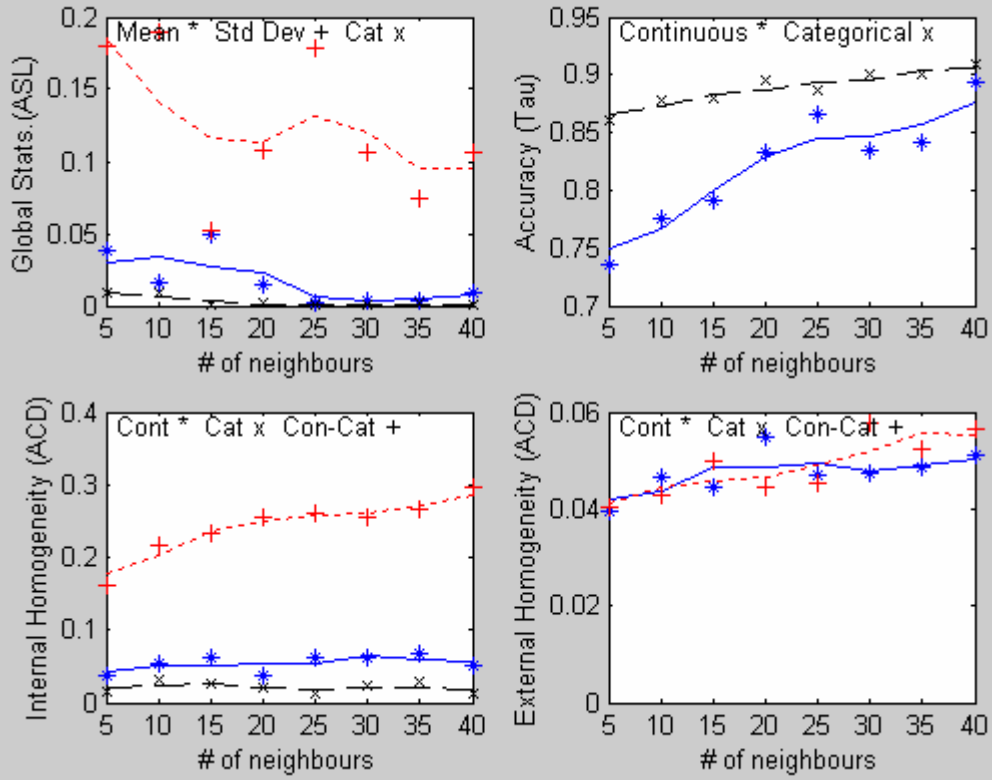
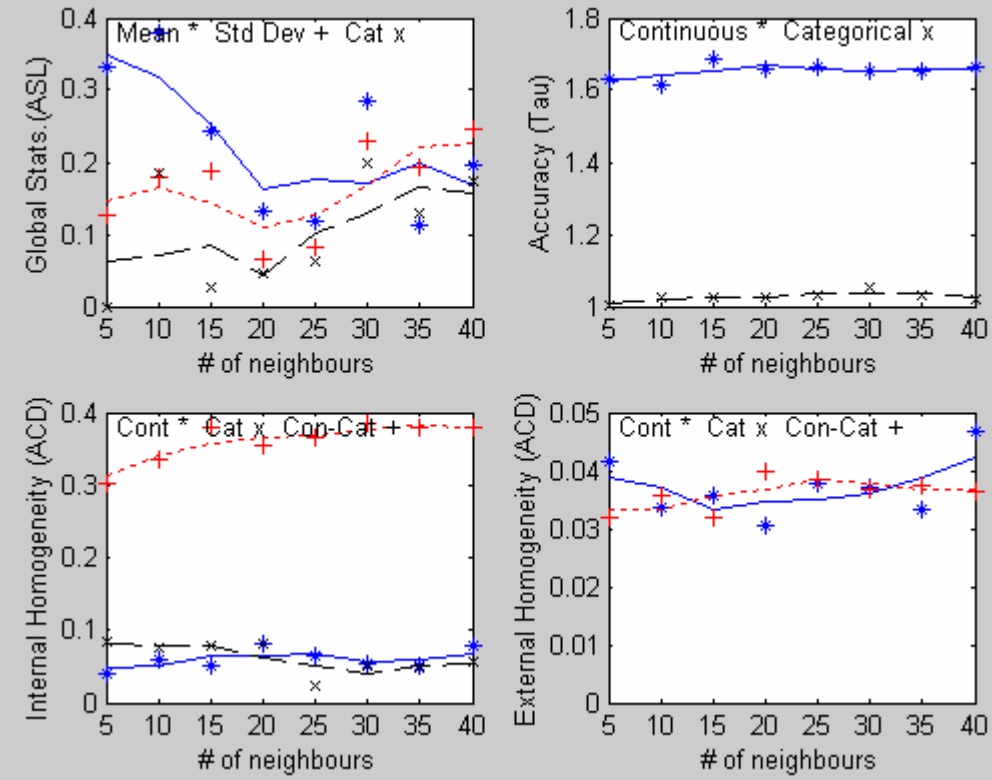


Figura 4.26a Resultados del ensayo 2.4 con la Base 0

II-1 **Figura 4.26b** Resultados del ensayo 2.4 con la Base 1 TKSM



II-2 **Figura 4.26c** Resultados del ensayo 2.4 con la Base 2 TKSM



4.5.4.2 Análisis de los resultados

- ASL.- Se reproducen bien la media y el caso categórico solo en el caso de ruido aleatorio amplificado. La desviación estándar solo se reproduce mal en el caso de la varianza residual reducida.

- Tau.- Para los casos de la varianza residual normal y reducida, los valores obtenidos del índice (continuo y categórico) son buenos y presentan una tendencia creciente con el parámetro K. En el caso del ruido aleatorio amplificado, los valores parecen no depender del valor de K. Los valores para el caso continuo son superiores a 1.

- ACD (Interna).- En todos los casos, los valores obtenidos son buenos aunque tienden a crecer en función de la varianza residual.

- ACD (Externa).- En todos los casos, los valores obtenidos son buenos. Ligeramente mejores para el caso de ruido aleatorio amplificado.

Parte II
4.6 Ensayos (parámetro π)

En esta parte de la simulación, el objetivo es evaluar el efecto en los ensayos del parámetro π (cf. 3.4). Los ensayos se hicieron con el método TKSM. El número de vecinos seleccionados para este propósito fueron: 3, 10, 20 y 40 con los siguientes valores de π: 0.0, 0.1, 0.3, 0.5, 0.7, 0.9 y 1.0. Los resultados se presentan en las figuras 4.27a, 4.27b, 4.27c y 4.27d. Los archivos empleados en los distintos ensayos corresponden a la base de datos generada de forma aleatoria.

4.6.1 Ensayos (Simulación I)

- Ensayo 1. Procedimiento de imputación: TKSM (3 vecinos)
- Ensayo 2. Procedimiento de imputación: TKSM (10 vecinos)
- Ensayo 3. Procedimiento de imputación: TKSM (20 vecinos)
- Ensayo 4. Procedimiento de imputación: TKSM (40 vecinos)

Parámetro: π (peso de la muestra local)
Valor del Parámetro: π = 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0
Archivo: Base 0

Tabla 4.6 Valores de π según (3.15) para la Base 1

K	Variable 1	Variable 2	Variable 3	Promedio
3	0.8686	0.8948	0.8319	0.8651
10	0.9424	0.9443	0.9225	0.9364
20	0.9643	0.9606	0.9519	0.9590
40	0.9783	0.9727	0.9704	0.9738

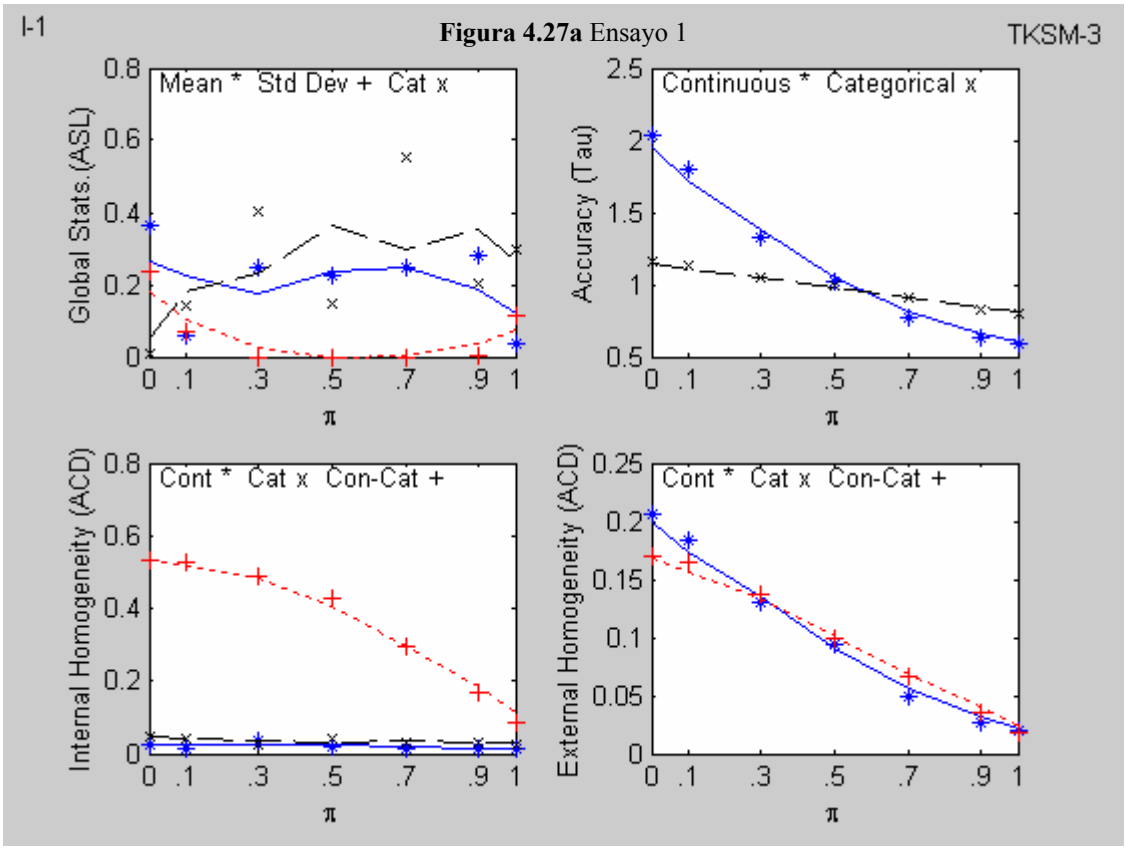


Figura 4.27a Ensayo 1

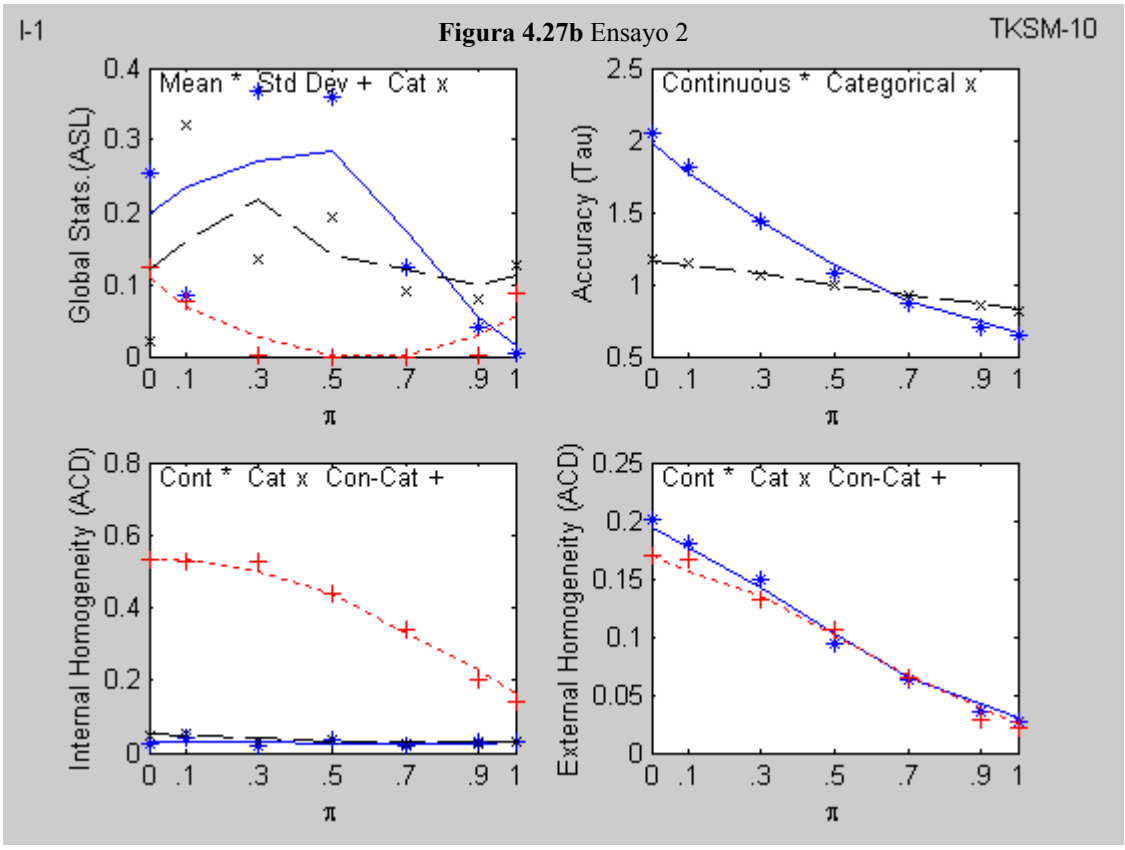


Figura 4.27b Ensayo 2

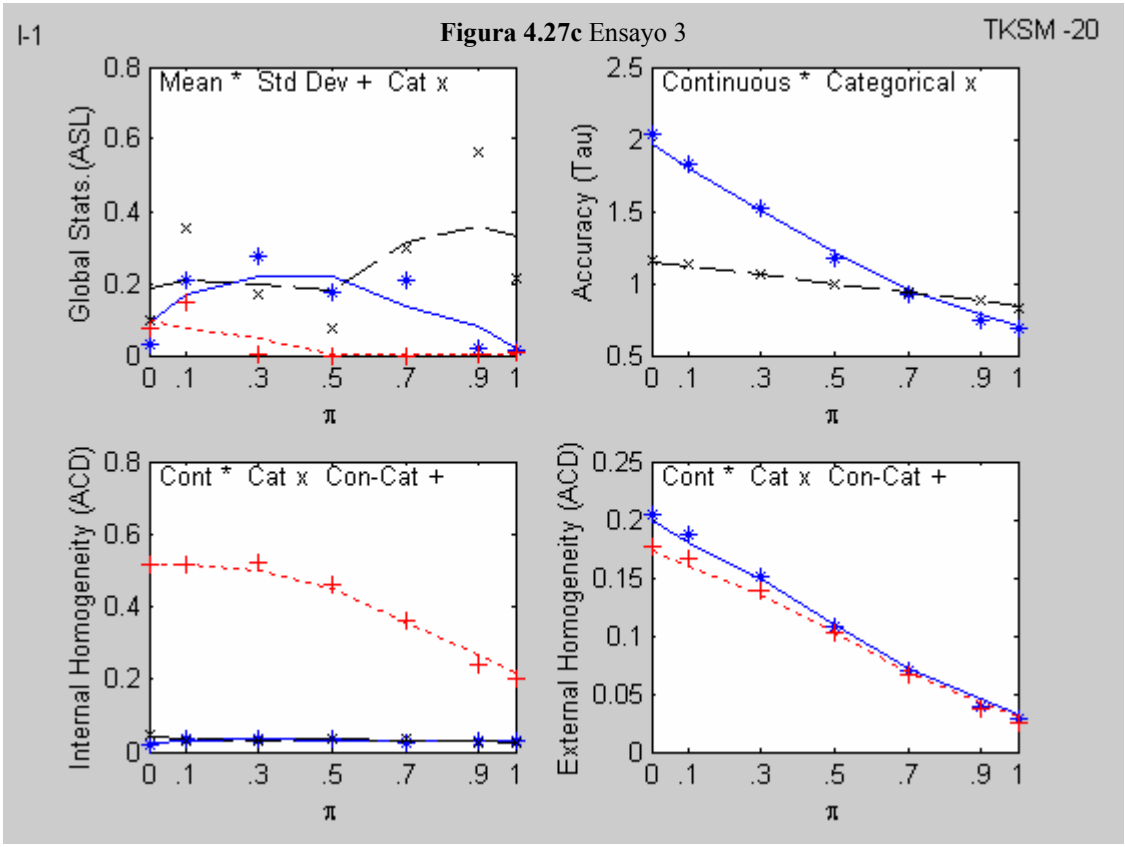


Figura 4.27c Ensayo 3

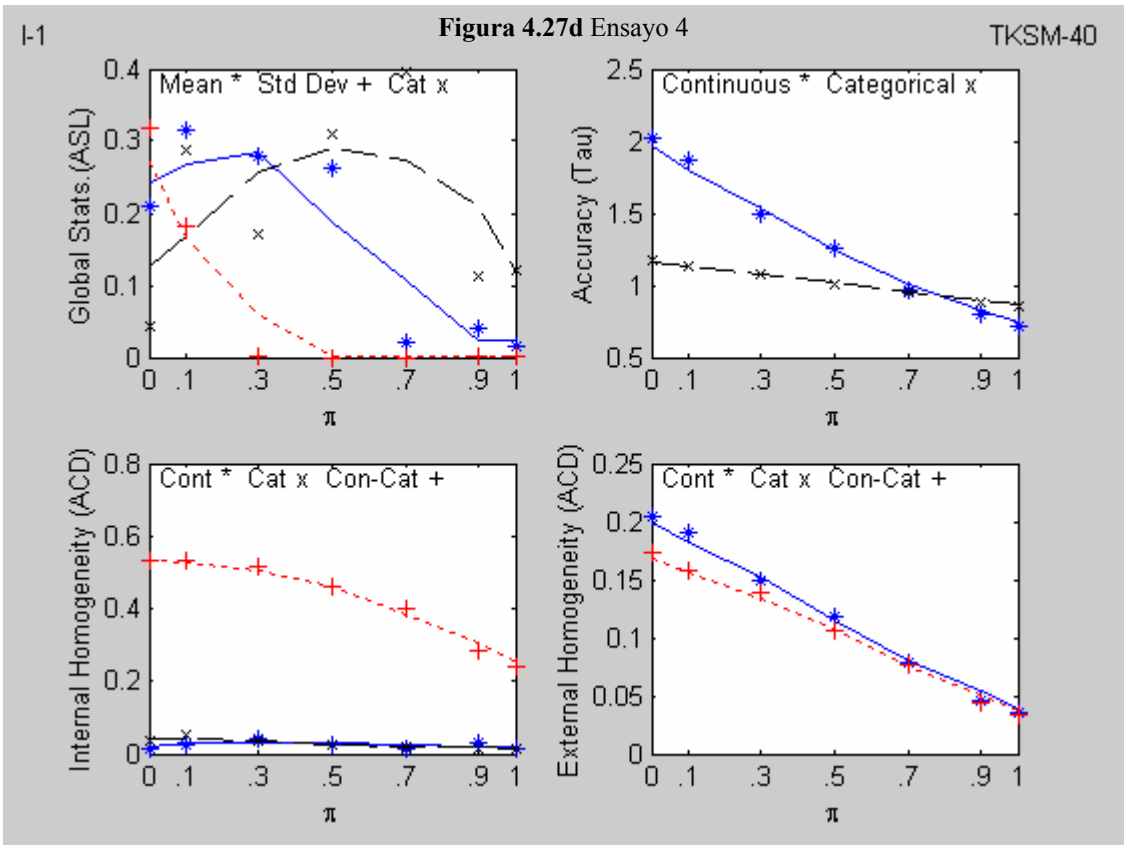


Figura 4.27d Ensayo 4

4.6.2 Análisis de los resultados

- ASL.- Se reproducen bien la media y el caso categórico para valores intermedios de p en los ensayos desarrollados. La desviación estándar solo se reproduce bien para valores pequeños de π.

- Tau.- Se observa en las gráficas que el valor del índice mejora en función de los valores crecientes de π. En π = 0, que representa una extracción aleatoria global, el valor es 2, como se podría esperar.
  Los valores teóricos calculados se pueden ver en la tabla 4.6.

- ACD (Interna).- Las gráficas no muestran evidencia de que el parámetro π tenga algún efecto sobre el índice. Los valores obtenidos son buenos para los casos continuo y categórico, no así para la relación continua-categórica.

- ACD (Externa).- En los 4 ensayos realizados, se aprecia una tendencia decreciente.

4.6.3 Ensayos (Simulación II)

Los ensayos se hicieron con el método TKSM. El número de vecinos seleccionados para evaluar el efecto del parámetro π fueron: 3, 10, 20 y 40 con los siguientes valores de π: 0.0, 0.1, 0.3, 0.5, 0.7, 0.9 y 1.0 Los resultados se presentan en las figuras 4.28a, 4.28b, 4.28c y 4.28d. Los archivos empleados en los distintos ensayos corresponden a la base de datos generada de manera no aleatoria.

- Ensayo 5. Procedimiento de imputación: TKSM (3 vecinos)
- Ensayo 6. Procedimiento de imputación: TKSM (10 vecinos)
- Ensayo 7. Procedimiento de imputación: TKSM (20 vecinos)
- Ensayo 8. Procedimiento de imputación: TKSM (40 vecinos)

Parámetro: π (peso de la muestra local)
Valor del Parámetro: π = 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0
Archivos: Base 1

Tabla 4.7 Valores de π según (3.15) para la Base 1

K	Variable 1	Variable 2	Variable 3	Promedio
3	0.8869	0.9199	0.8611	0.8893
4	0.9506	0.9542	0.9338	0.9462
5	0.9707	0.9688	0.9594	0.9663
10	0.9822	0.9784	0.9747	0.9785

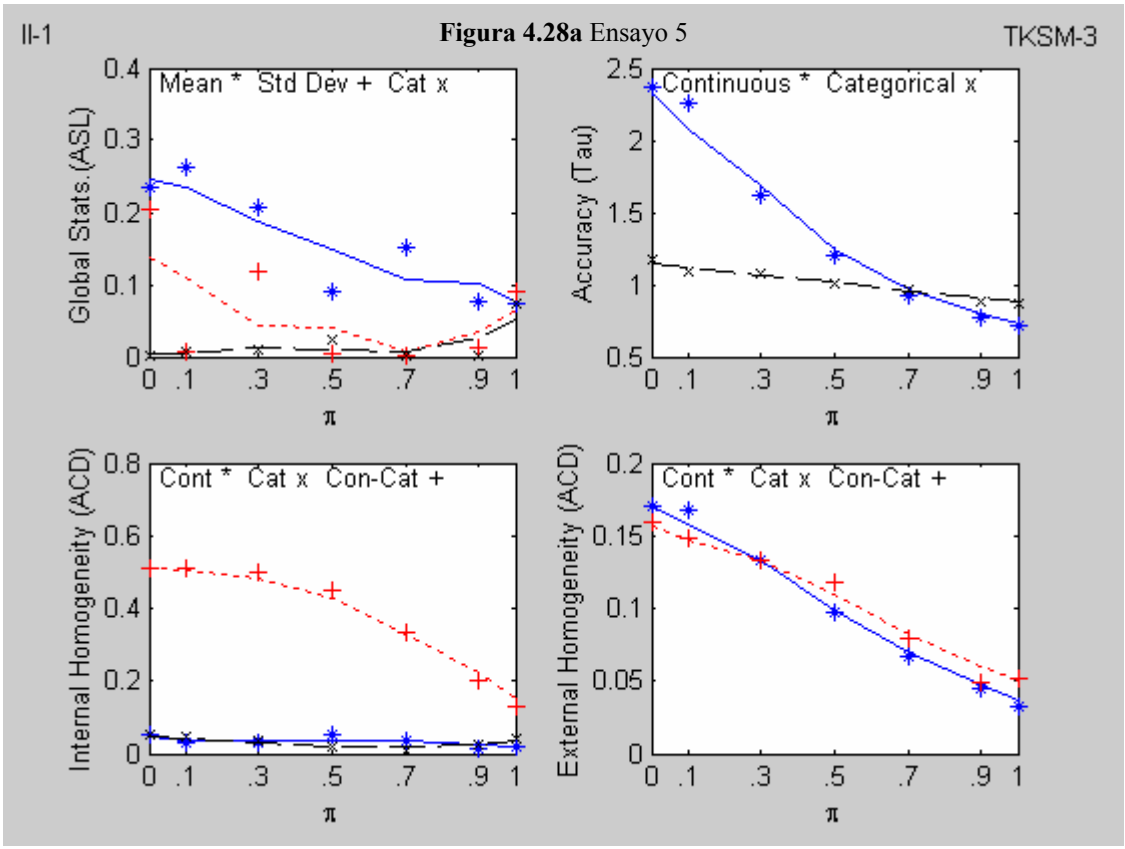


Figura 4.28a Ensayo 5

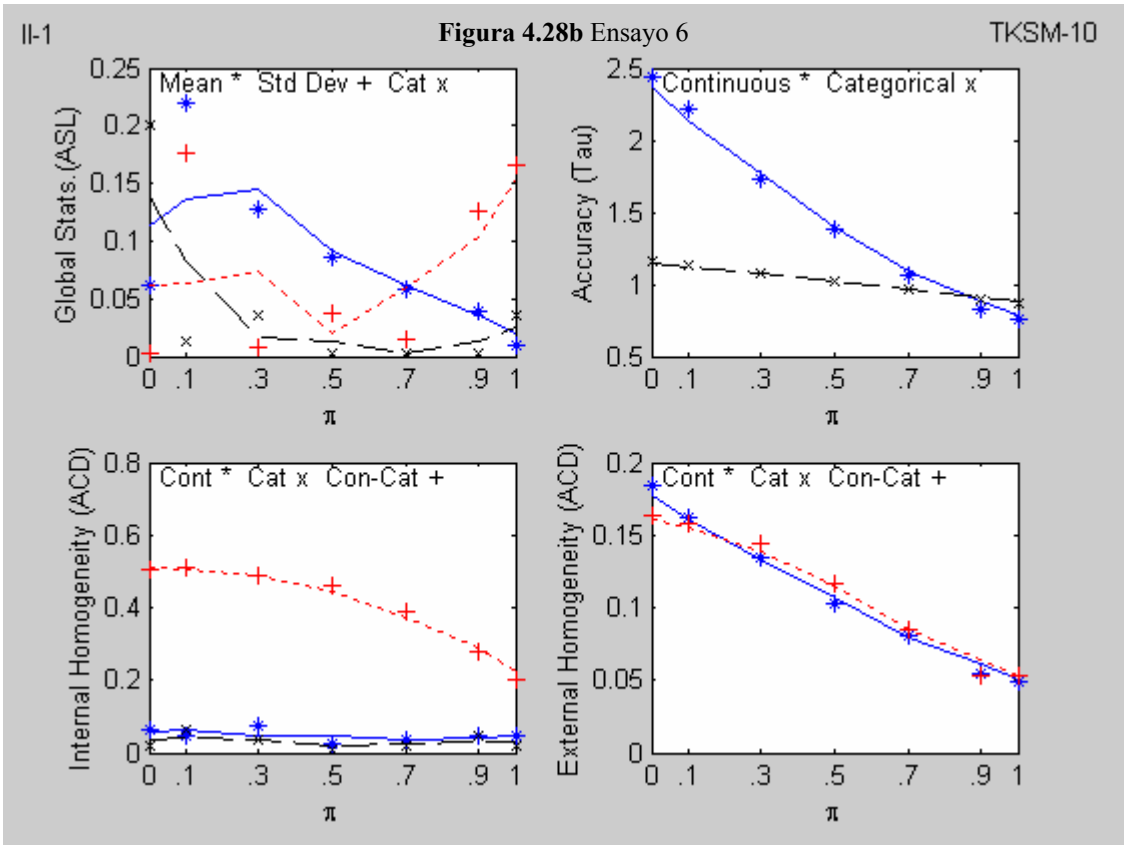


Figura 4.28b Ensayo 6

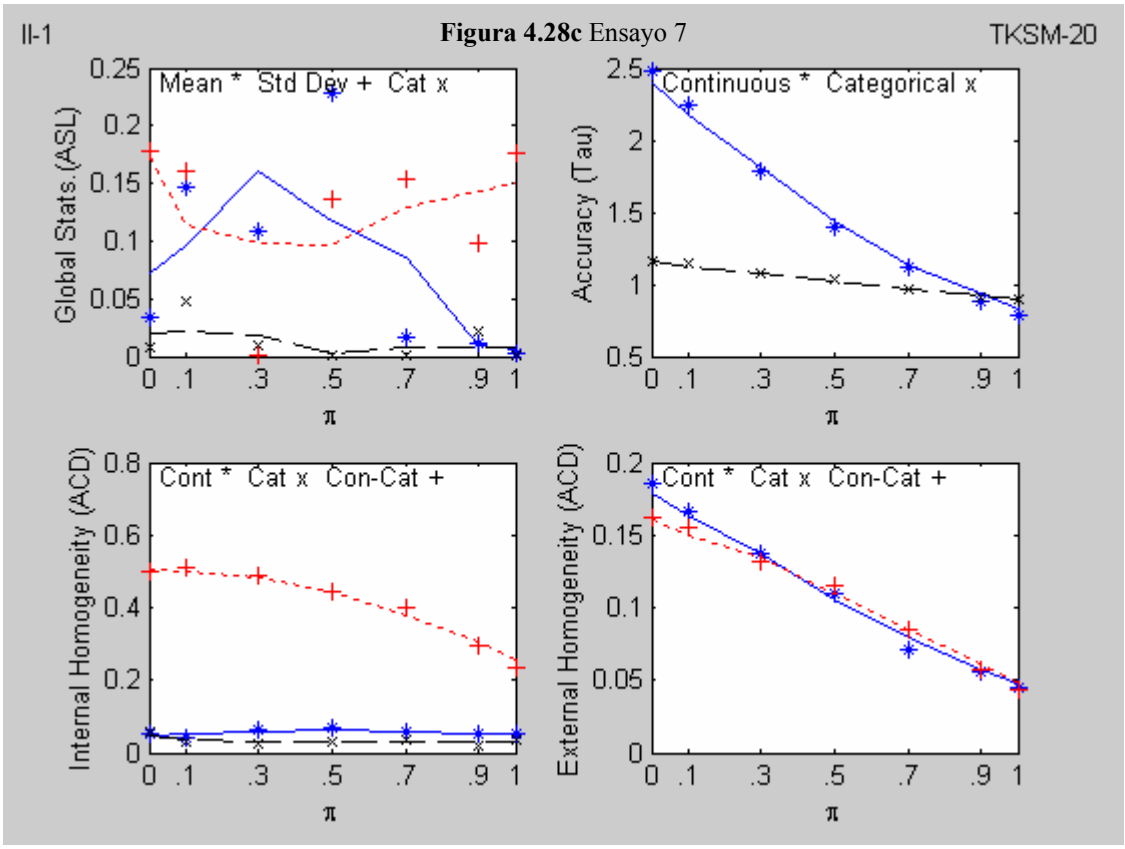


Figura 4.28c Ensayo 7

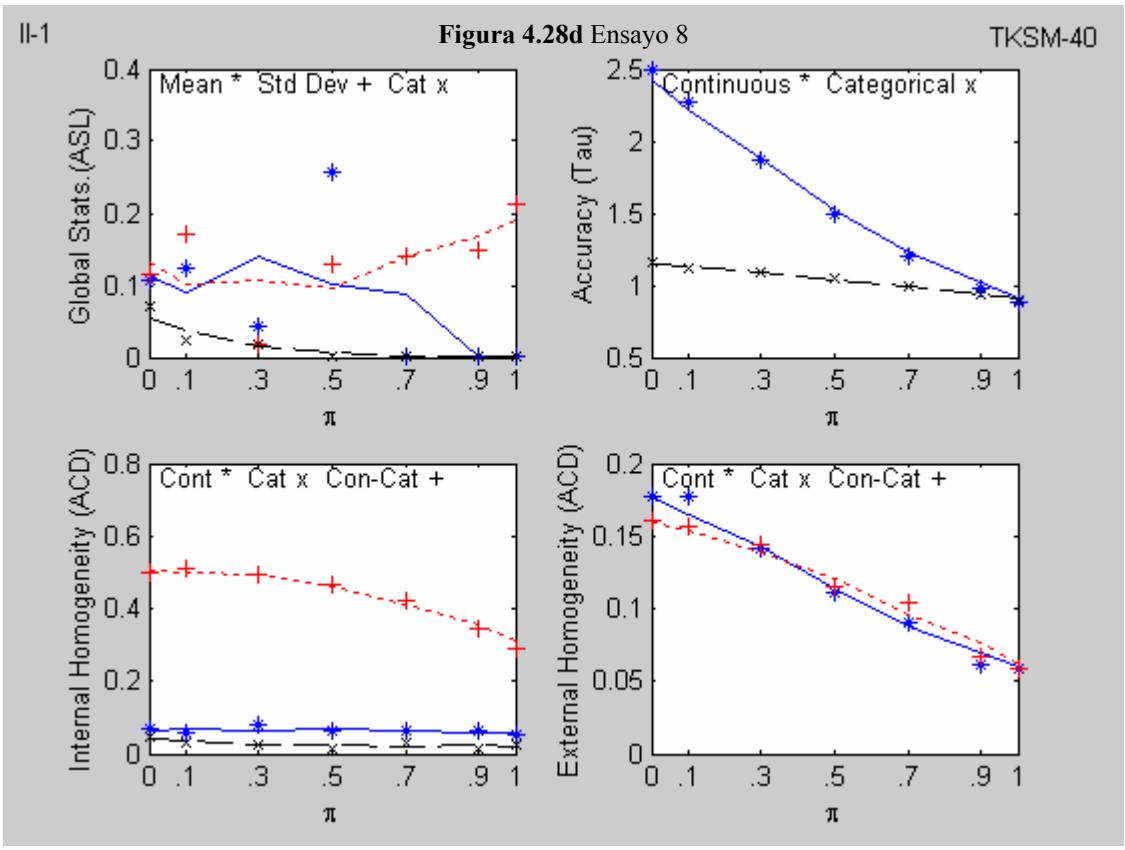


Figura 4.28d Ensayo 8

4.6.4 Análisis de los resultados

- ASL.- Se reproducen bien la media y el caso categórico para valores intermedios de p en los ensayos desarrollados. La desviación estándar solo se reproduce bien para valores pequeños de π.

- Tau.- Se observa en las gráficas que el valor del índice mejora en función de los valores crecientes de π. En π = 0, que representa una extracción aleatoria global, el valor es 2, como se podría esperar.
  Los valores teóricos calculados se pueden ver en la tabla 4.7.

- ACD (Interna).- Las gráficas no muestran evidencia de que el parámetro π tenga algún efecto sobre el índice. Los valores obtenidos son buenos para los casos continuo y categórico, no así para la relación continua-categórica.

- ACD (Externa).- En los 4 ensayos realizados, se aprecia una tendencia decreciente.

4.7 Imputación EM

Se presenta a continuación, la imputación hecha con el método EM. La imputación se ha hecho solo para las variables continuas. Los ensayos se desarrollaron con los archivos (Base 0, Base 1, Base 2) correspondientes a la selección aleatoria (Simulación I) y los archivos (Base 0, Base 1, Base 2) correspondientes a la selección no aleatoria de los individuos (Simulación II).

El objetivo de estos ensayos, es el de poner dentro de un marco de referencia los resultados obtenidos con los métodos propuestos en esta investigación.

La figura 4.29a y 4.29b presentan los resultados de la imputación EM determinística y EM con un componente aleatorio para los distintos archivos en el caso de la selección aleatoria de los individuos (Simulación I).

La figura 4.30a y 4.30b presentan los resultados de los ensayos con los archivos correspondientes a la selección no aleatoria de los individuos (Simulación II). Los resultados mostrados corresponden a la imputación EM determinística y EM con un componente aleatorio.

4.7.1 Resultados

- EM-D Imputación EM determinística.

- EM-R Imputación EM con un componente aleatorio.

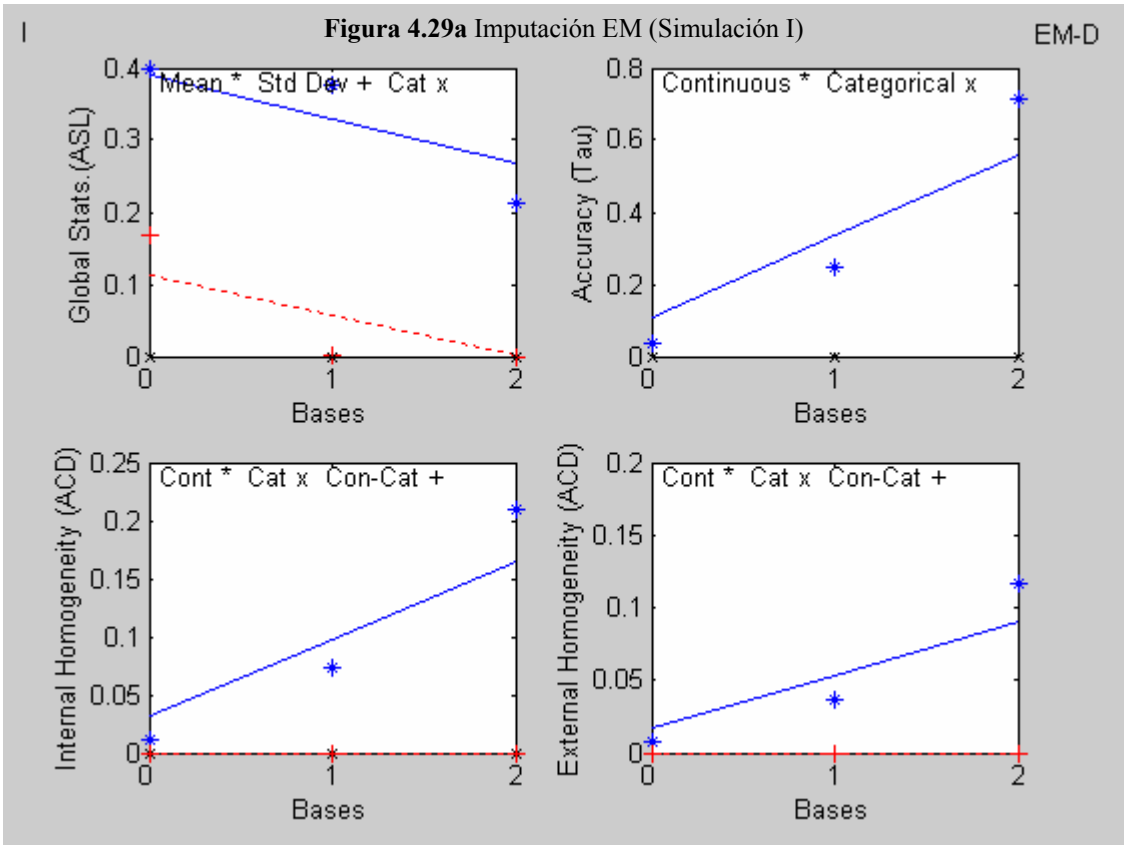


Figura 4.29a Imputación EM (Simulación I)

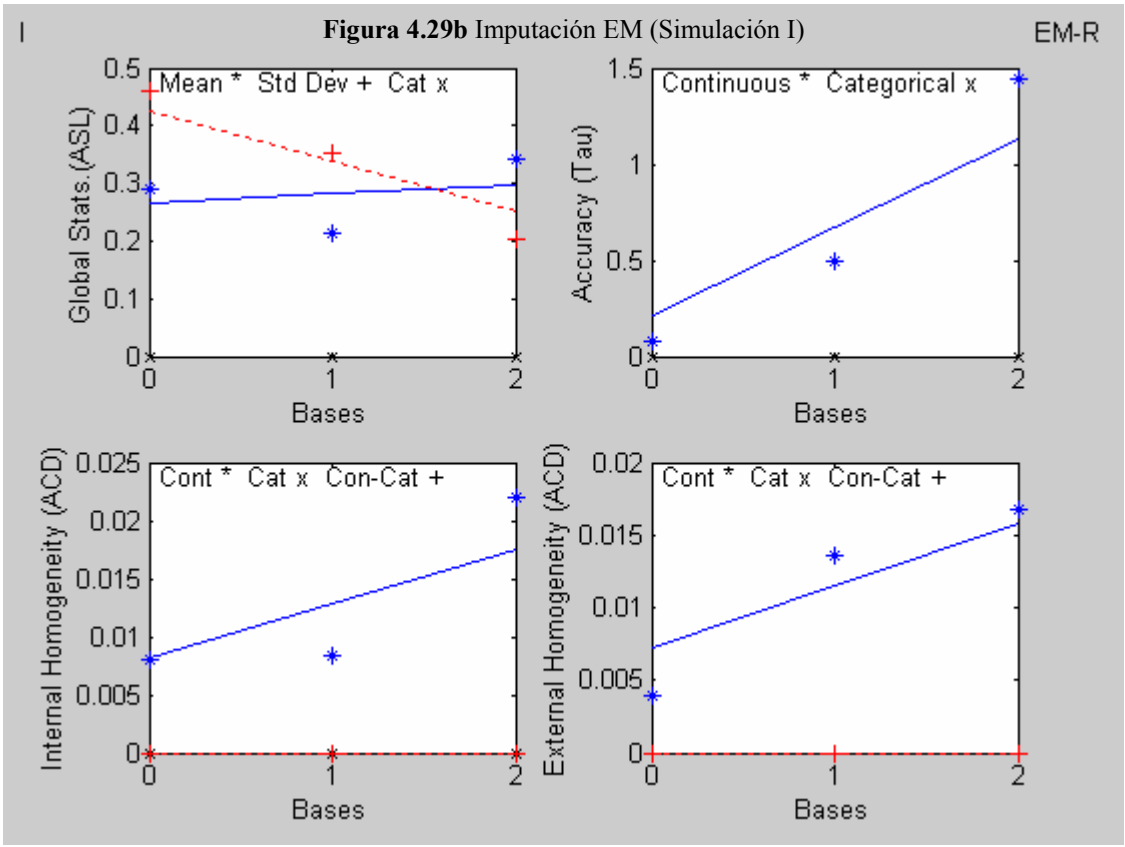
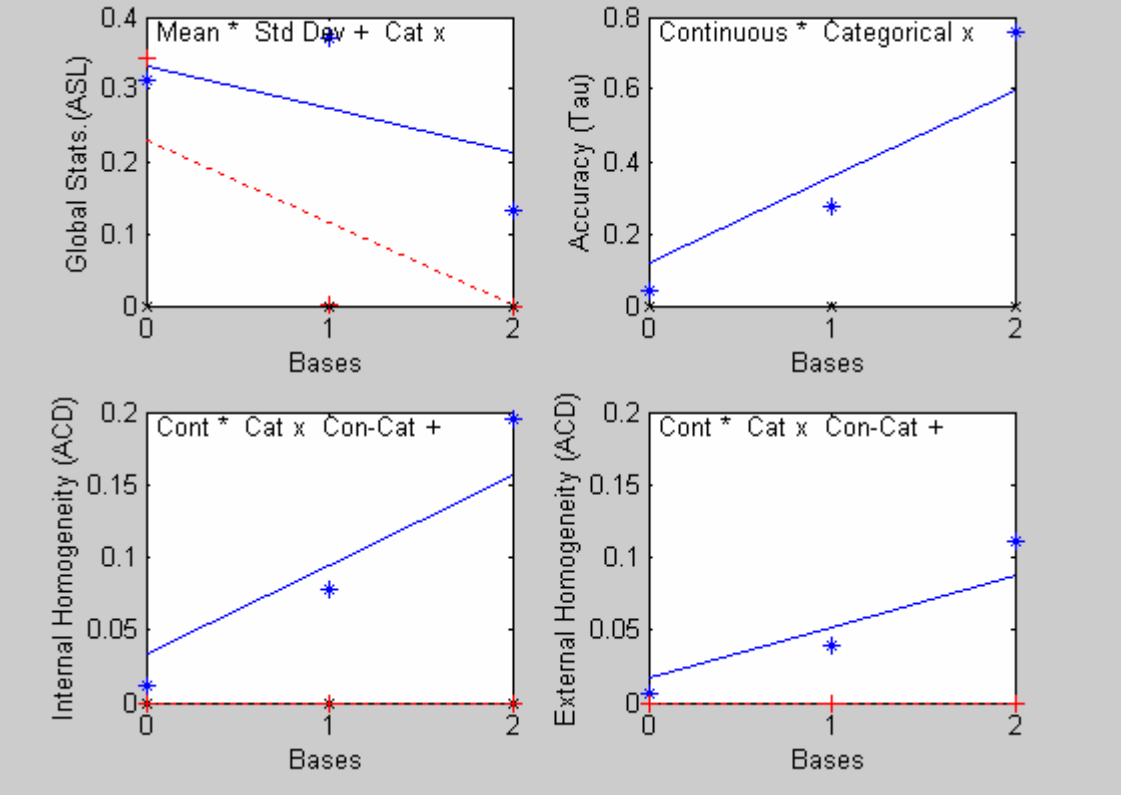
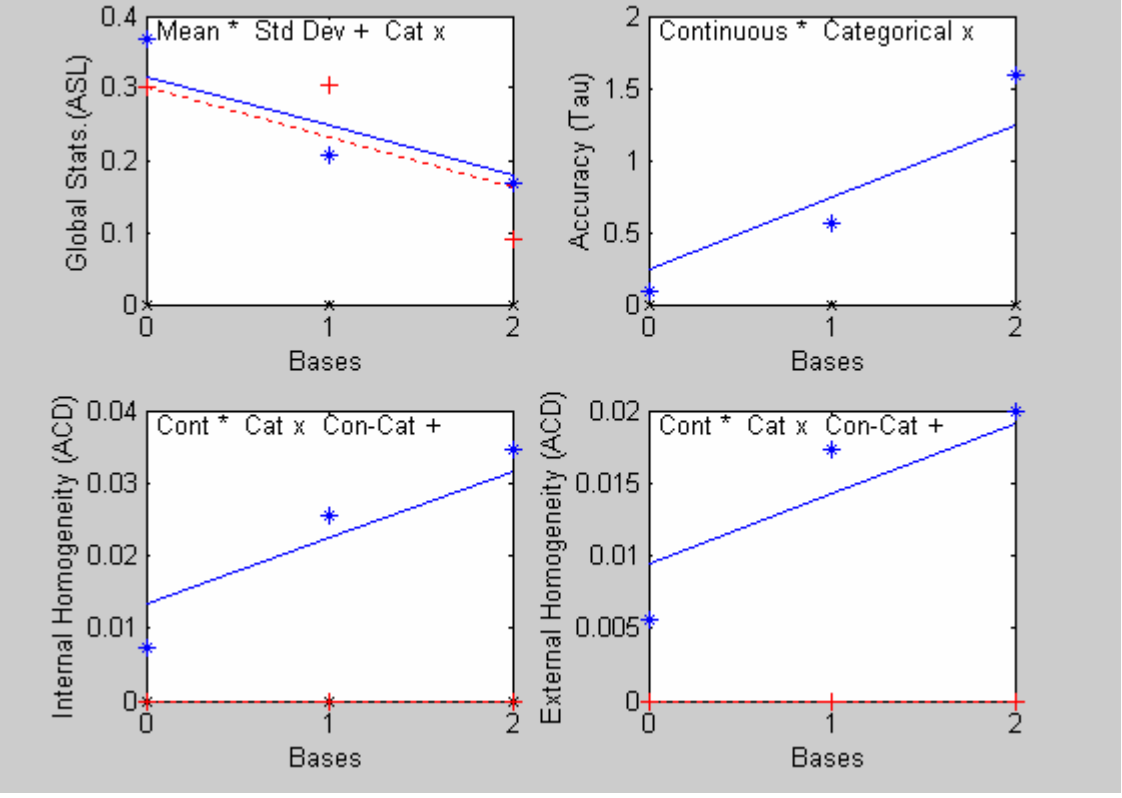


Figura 4.29b Imputación EM (Simulación I)

II **Figura 4.30a** Imputación EM (Simulación II) EM-D



II **Figura 4.30b** Imputación EM (Simulación II) EM-R



4.7.2 Análisis de los resultados

- ASL.- Se observa en las gráficas de los ensayos en la simulación I como en la II, que la media se reproduce bien. En el caso determinístico de la simulación I y el aleatorio de la simulación II se observa que el índice disminuye en función de la varianza residual. La desviación estándar se reproduce bien para el caso de la varianza residual disminuida en el caso determinístico. En el caso aleatorio, se reproduce bien la desviación estándar.

- Tau.- Los valores del índice son mayores que 1 para el caso aleatorio con el ruido aleatorio aumentado. Se observa en las gráficas una tendencia creciente con la varianza residual.

- ACD (Interna).- Los valores del índice son malos para el caso del ruido aleatorio amplificado en el caso determinístico.

- ACD (Externa).- Los valores del índice son malos para el caso del ruido aleatorio amplificado en el caso determinístico. Mejoran con respecto al índice interno.

4.8 Imputación PLS

Se presenta a continuación, la imputación hecha con el método PLS. La imputación se ha hecho solo para las variables continuas. Los ensayos se desarrollaron con los archivos (Base 0, Base 1, Base 2) correspondientes a la selección aleatoria (Simulación I) y los archivos (Base 0, Base 1, Base 2) correspondientes a la selección no aleatoria de los individuos (Simulación II).

La figura 4.31a y 4.31b presentan los resultados de la imputación PLS determinística y PLS con un componente aleatorio para los distintos archivos en el caso de la selección aleatoria de los individuos (Simulación I).

La figura 4.32a y 4.32b presenta los resultados de los ensayos con los archivos correspondientes a la selección no aleatoria de los individuos (Simulación II). Los resultados mostrados corresponden a la imputación PLS determinística y PLS con un componente aleatorio.

Al igual que con la imputación EM, se trata de poner dentro de un marco de referencia los resultados obtenidos con los métodos propuestos en esta investigación.

4.8.1 Resultados

- PLS-D Imputación PLS determinística.
- PLS -R Imputación PLS con un componente aleatorio.

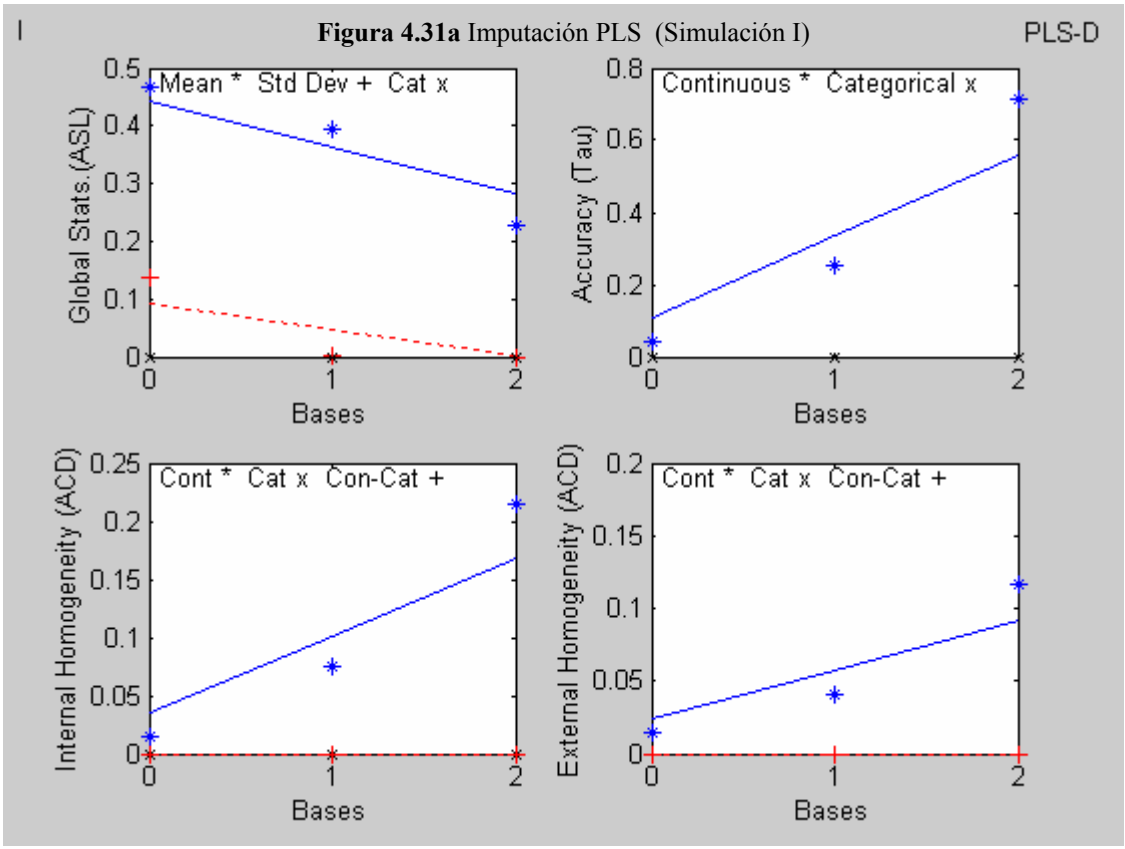


Figura 4.31a Imputación PLS (Simulación I)

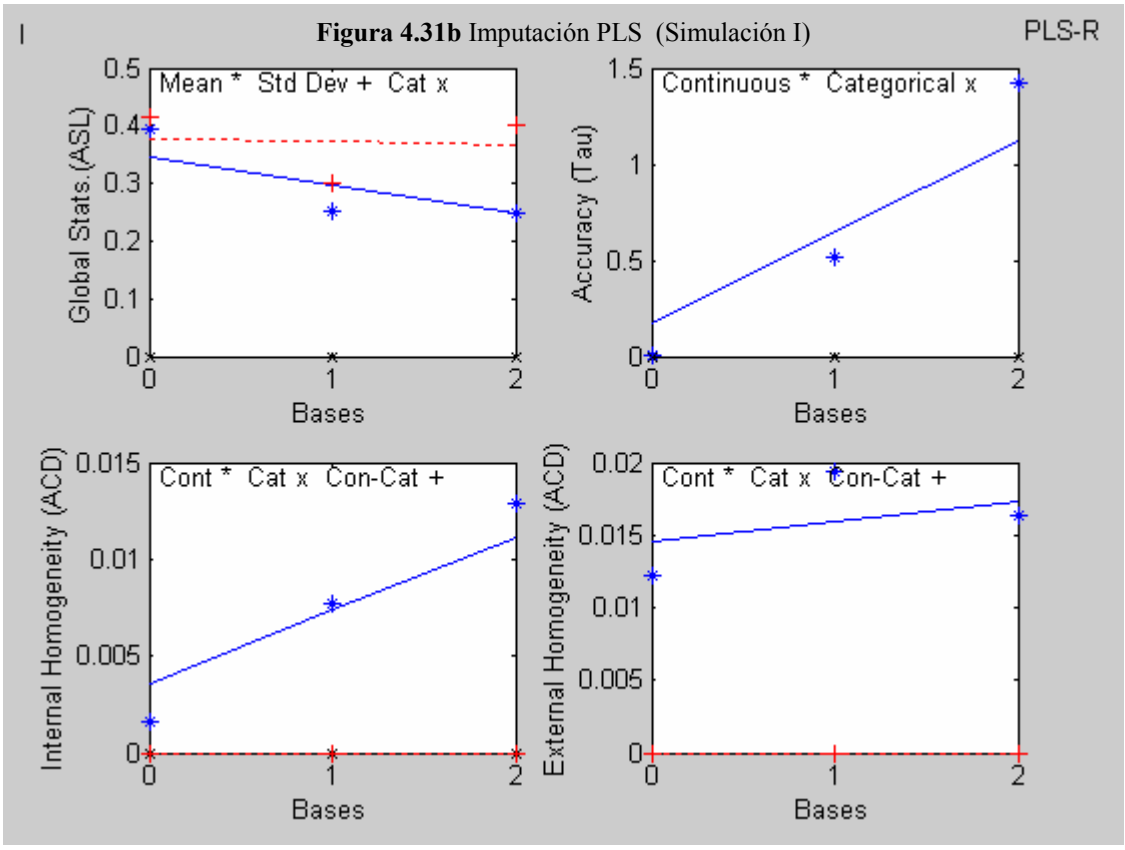


Figura 4.31b Imputación PLS (Simulación I)

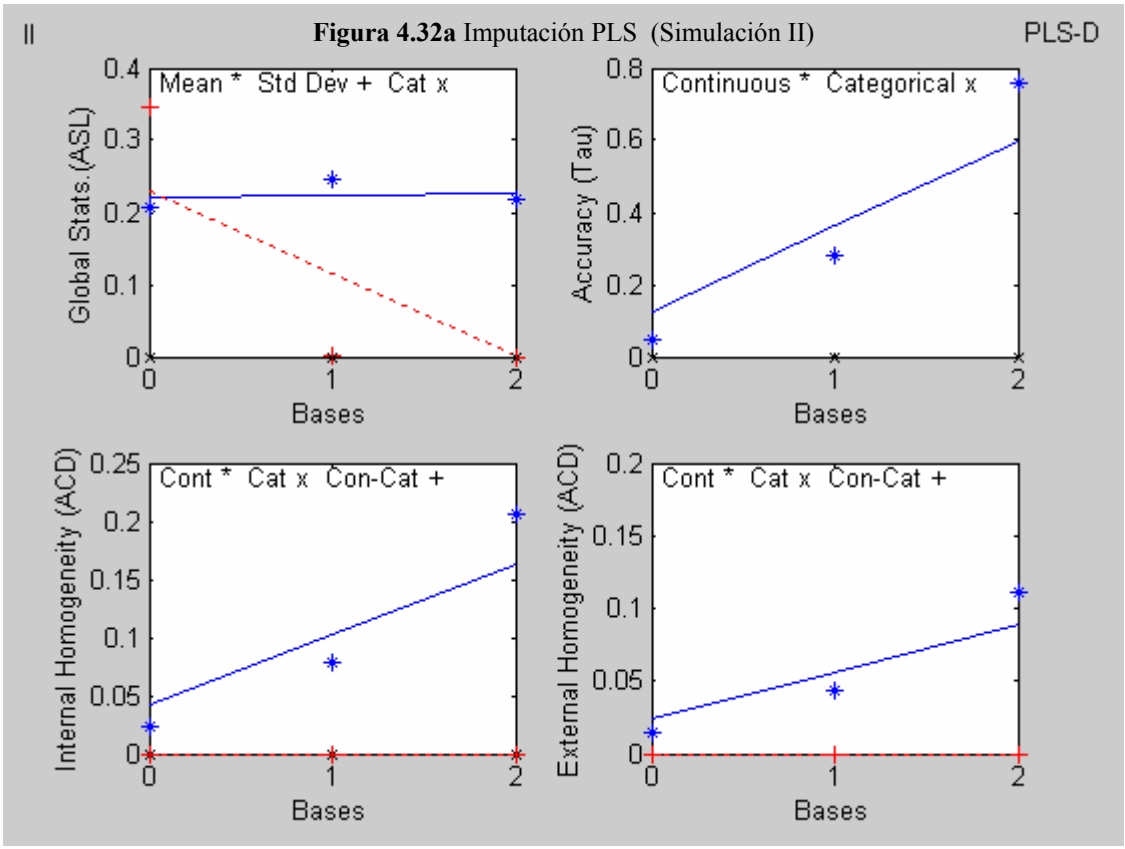


Figura 4.32a Imputación PLS (Simulación II)

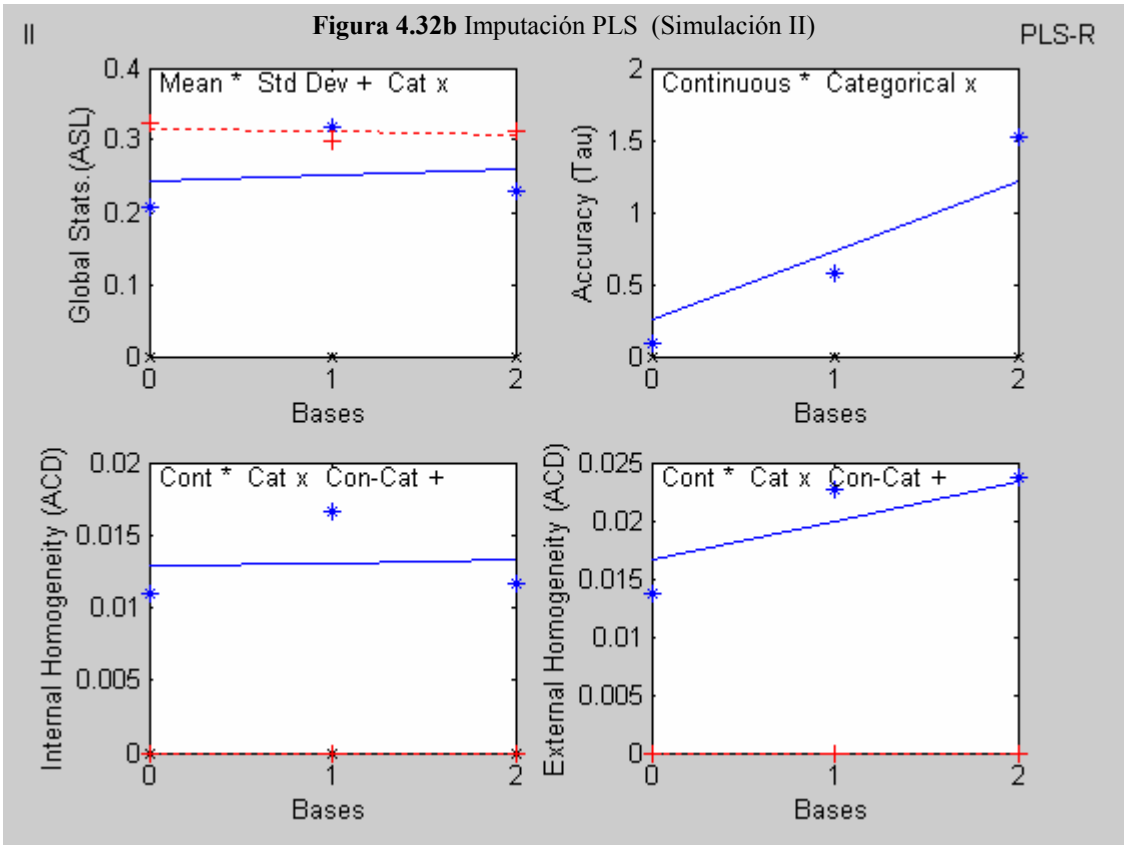


Figura 4.32b Imputación PLS (Simulación II)

4.8.2 Análisis de los resultados

- ASL.- Se observa en las gráficas de los ensayos en la simulación I como en la II, que la media se reproduce bien. En el caso determinístico de la simulación I y el aleatorio de la simulación II se observa que el índice disminuye en función de la varianza residual. La desviación estándar se reproduce bien para el caso de la varianza residual disminuida en el caso determinístico. En el caso aleatorio, se reproduce bien la desviación estándar.

- Tau.- Los valores del índice son mayores que 1 para el caso aleatorio con el ruido aleatorio aumentado. Se observa en las gráficas una tendencia creciente con la varianza residual.

- ACD (Interna).- Los valores del índice son malos para el caso del ruido aleatorio amplificado en el caso determinístico.

- ACD (Externa).- Los valores del índice son malos para el caso del ruido aleatorio amplificado en el caso determinístico. Mejoran con respecto al índice interno.

Los valores observados son similares a los obtenidos con el método EM. En el caso de la homogeneidad interna, se aprecian resultados ligeramente mejores con este método, en comparación con el método EM. En el caso de la homogeneidad externa, se observan resultados ligeramente mejores con el método EM.

4.9 Tabla comparativa

Tabla 4.8 Resumen de resultados obtenidos con las bases generadas de forma aleatoria

		ASL			Precisión			Homogeneidad Interna			Homogeneidad Externa		
		Base0	Base1	Base2	Base0	Base1	Base2	Base0	Base1	Base2	Base0	Base1	Base2
	EM-D	0.3978	0.3775	0.2133	0.0375	0.2191	0.7149	0.0111	0.0724	0.2098	0.0075	0.0366	0.1171
	PLS-D	0.4684	0.3956	0.2268	0.0404	0.2518	0.7143	0.0141	0.0755	0.2142	0.0149	0.0409	0.1172
P = 0	T1DM	0.3245	0.2402	0.1707	0.2120	0.6102	1.4714	0.0121	0.0186	0.0136	0.0155	0.0193	0.0183
K = 5	TKDM	0.0301	0.0508	0.3455	0.1199	0.4002	1.3282	0.0086	0.0328	0.0367	0.0157	0.0248	0.0361
	EM-S	0.2911	0.2118	0.3415	0.0760	0.4972	1.4415	0.0081	0.0083	0.0221	0.0040	0.0136	0.0168
	PLS-S	0.3939	0.2524	0.2469	0.0032	0.5106	1.4252	0.0016	0.0077	0.0129	0.0122	0.0194	0.0163
	T1SM	0.0026	0.0391	0.2540	0.3739	0.7293	1.5568	0.0159	0.0257	0.0155	0.0348	0.0357	0.0312
K = 5	TKSM	0.0436	0.0210	0.2299	0.2191	0.6125	1.4653	0.0124	0.0109	0.0186	0.0190	0.0213	0.0222

Tabla 4.9 Resumen de resultados obtenidos con las bases generadas de forma no aleatoria

		ASL			Precisión			Homogeneidad Interna			Homogeneidad Externa		
		Base0	Base1	Base2	Base0	Base1	Base2	Base0	Base1	Base2	Base0	Base1	Base2
	EM-D	0.3136	0.3720	0.1328	0.0434	0.2771	0.7606	0.0120	0.0780	0.0949	0.0069	0.0389	0.1114
	PLS-D	0.2063	0.2463	0.2185	0.0487	0.2808	0.7590	0.0227	0.0790	0.2061	0.0151	0.0429	0.1115
P = 0	T1DM	0.2062	0.2062	0.3203	0.2665	0.7074	1.6431	0.0060	0.0189	0.0344	0.0300	0.0310	0.0265
K = 5	TKDM	0.0111	0.0464	0.4136	0.1636	0.4388	0.9230	0.0121	0.0391	0.1136	0.0341	0.0478	0.0713
	EM-S	0.3691	0.2065	0.1676	0.0887	0.5569	1.5847	0.0074	0.0254	0.0345	0.0056	0.0174	0.0199
	PLS-S	0.2063	0.3172	0.2295	0.0971	0.5751	1.5261	0.0110	0.0167	0.0116	0.0136	0.0226	0.0236
	T1SM	0.0032	0.0023	0.0666	0.4560	0.8748	1.6612	0.0170	0.0262	0.0284	0.0494	0.0615	0.0404
K = 5	TKSM	0.0107	0.0378	0.3327	0.3054	0.7359	1.6296	0.0142	0.0372	0.0409	0.0412	0.0396	0.0414

Las tablas 4.8 y 4.9, muestran de manera comparativa los resultados obtenidos en los distintos ensayos desarrollados.

La tabla 4.8 corresponde a los resultados obtenidos con los archivos correspondientes a la selección aleatoria de los individuos (Simulación I).

La tabla 4.9 corresponde a los resultados obtenidos con los archivos correspondientes a la selección aleatoria de los individuos (Simulación II).

En los dos casos se han presentado los resultados en dos partes: ensayos determinísticos y ensayos estocásticos. Los parámetros seleccionados corresponden a los observados en las distintas gráficas mostradas anteriormente.

Tomando como referencia la base 1, se puede observar para el indicador de Precisión, que los valores son buenos en los casos de imputación determinística, tanto en la tabla 4.8 como 4.9.

Se observa que los valores son menores en los casos de los métodos EM y PLS.
En el caso de imputación con componente aleatorio, se ve que en todos los métodos para variabilidad grande (Base 2), los resultados se pueden calificar de malos.

Para el indicador de Homogeneidad interna, en términos generales, los resultados son mejores para los métodos propuestos en el caso determinístico, no así en el caso aleatorio. Los métodos determinísticos para EM y PLS presentan valores muy grandes de Homogeneidad Interna cuando la variabilidad es grande.

En el caso de la Homogeneidad Externa, los resultado para todos lo métodos son equiparables. Los valores obtenidos con los métodos propuestos son mejores que los obtenidos con PLS y EM en el caso determinístico.
Los valores del índice son bastante buenos.

En el caso del índice ASL, los resultados observados con los métodos propuestos son buenos para ruido aleatorio grande.

Parte III

4.10 Prueba del sistema

En esta parte del trabajo se hará una aplicación del sistema a un conjunto de datos proporcionados por el Instituto de Estadística de Cataluña (Idescat) relacionados con el estudio sobre la utilización de las Tecnologías de la Información y las comunicaciones (TIC) en los hogares de Cataluña correspondientes al mes de noviembre de 2002. Los datos representan una parte de la muestra y están relacionados con los sujetos que se conectan a Internet. En esta muestra se hace un seguimiento especial para conocer el uso de Internet en sus distintas categorías:

T_emailing (correo electrónico)
T_chating (charlar)
T_leisure (ocio)
T_buying (comprar)
shopping (despensa)
paying (pagar)
T_banking (operaciones bancarias)
studying (estudio)
T_government (trámites oficiales)
gtransactions (transacciones oficiales)
health (salud)
htransactions (transacciones de salud)
teleworking (trabajo a distancia)

Una tercera parte de esta información da lugar a los individuos donantes y el resto, a los individuos receptores. Se dispone en total de 1116 individuos.
Las variables consideradas como comunes se escogerán de la siguiente lista:

age (edad)
sex (sexo)
studies (nivel de estudios)
occupation (oficio)
children (niños)
quickcon (conexión rápida)
flatrate (tarifa plana)
freqcomp (frecuencia frente al ordenador)
ofimatics (software de oficina)
advanced (uso avanzado)
specifics (uso específico)
cleisure (ocio por ordenador)
webhome (Internet en casa)
webwork (Internet en el trabajo)
webstudy (Internet para estudios)
freqweb (frecuencia de conexión a Internet)
timewebhome (tiempo de Internet en casa)
timewebwork (tiempo de Internet en el trabajo)
english, spanish, catalán

4.10.1 Descripción del problema

Se trata de validar la metodología propuesta y el sistema "GRAFT" implementado para reproducir los hábitos de consumo de Internet a partir de datos socioeconómicos.
En esta aplicación, no se intenta efectuar una imputación con un índice alto de precisión (accuracy) para ser utilizados a nivel individual, sino que se pretende obtener un conjunto de datos imputados que simulen la realidad. Se trata de conservar índices adecuados de homogeneidad (interna y externa) que permitan desarrollar modelos del uso de Internet.
Este objetivo permite seleccionar de acuerdo a lo visto en las secciones anteriores el método de imputación TKSM como el indicado para esta situación.

Los datos contenidos en las variables específicas, representan proporciones, y por lo tanto sus valores estarán contenidos en el intervalo [0 , 1]. Esta característica de las variables permite hacer uso de las transformaciones, que para esta aplicación se considera como adecuada la transformación logit(x).

Debido a que no todos los usuarios de ordenador utilizan Internet para los consumos, se hará uso de las variables mixtas.

Descripción de las variables específicas

La figura 4.33 presenta un mosaico con algunas características de las variables específicas (histograma de frecuencias con la curva normal sobrepuesta, correlaciones). Se puede apreciar en algunas de ellas, el beneficio de la transformación efectuada (logit(x)).

Verificación de que la muestra de donantes y receptores son equivalentes.

La figura 4.34 muestra en el plano de los primeros dos ejes factoriales la representación de los individuos donantes y los individuos receptores. Si bien es posible contrastar esta equivalencia mediante pruebas estadísticas específicas, nos limitaremos a dar una apreciación visual de la equivalencia entre ambas muestras. Se puede establecer a partir de la figura que las dos muestras provienen de la misma población.

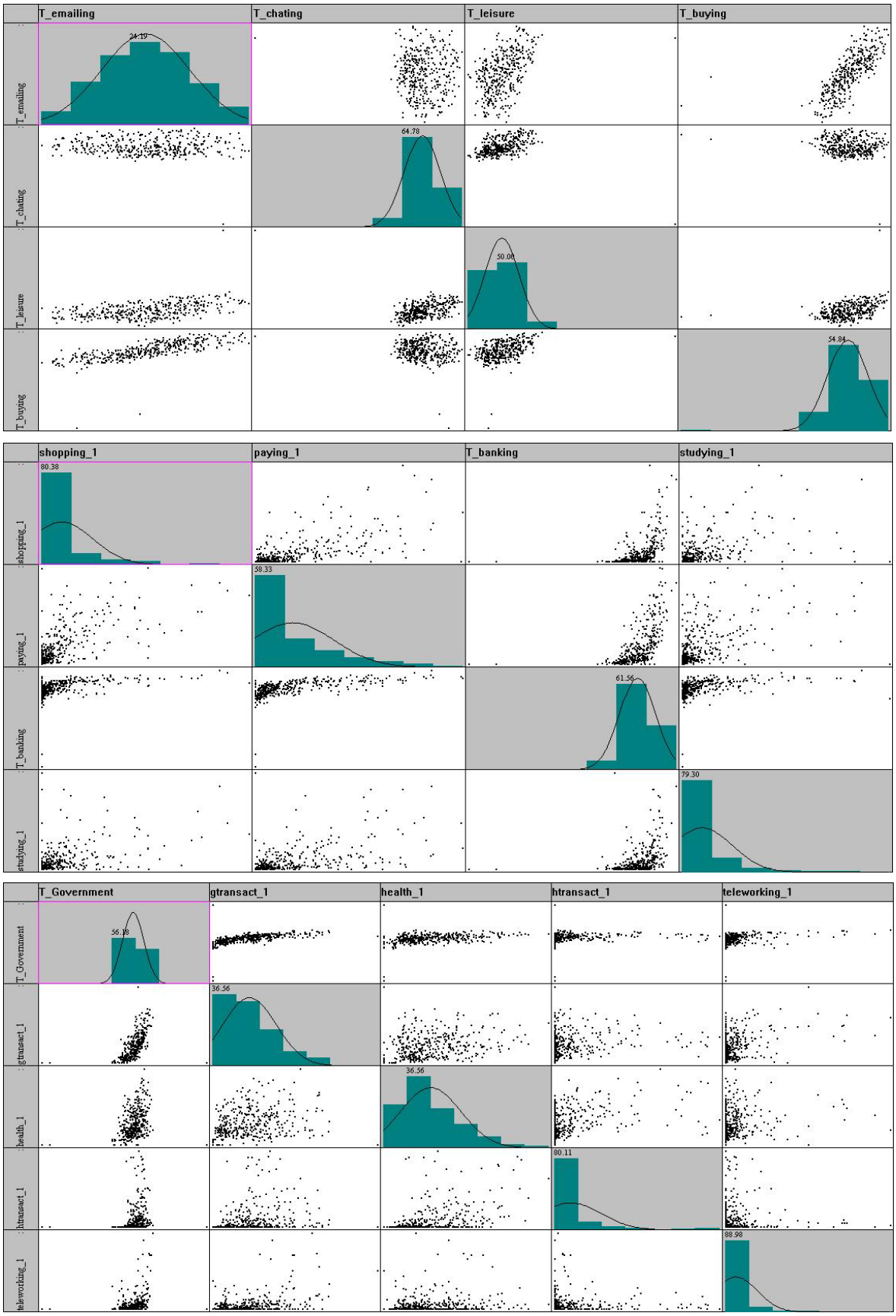


Figura 4.33 Histogramas y correlaciones de las variables específicas

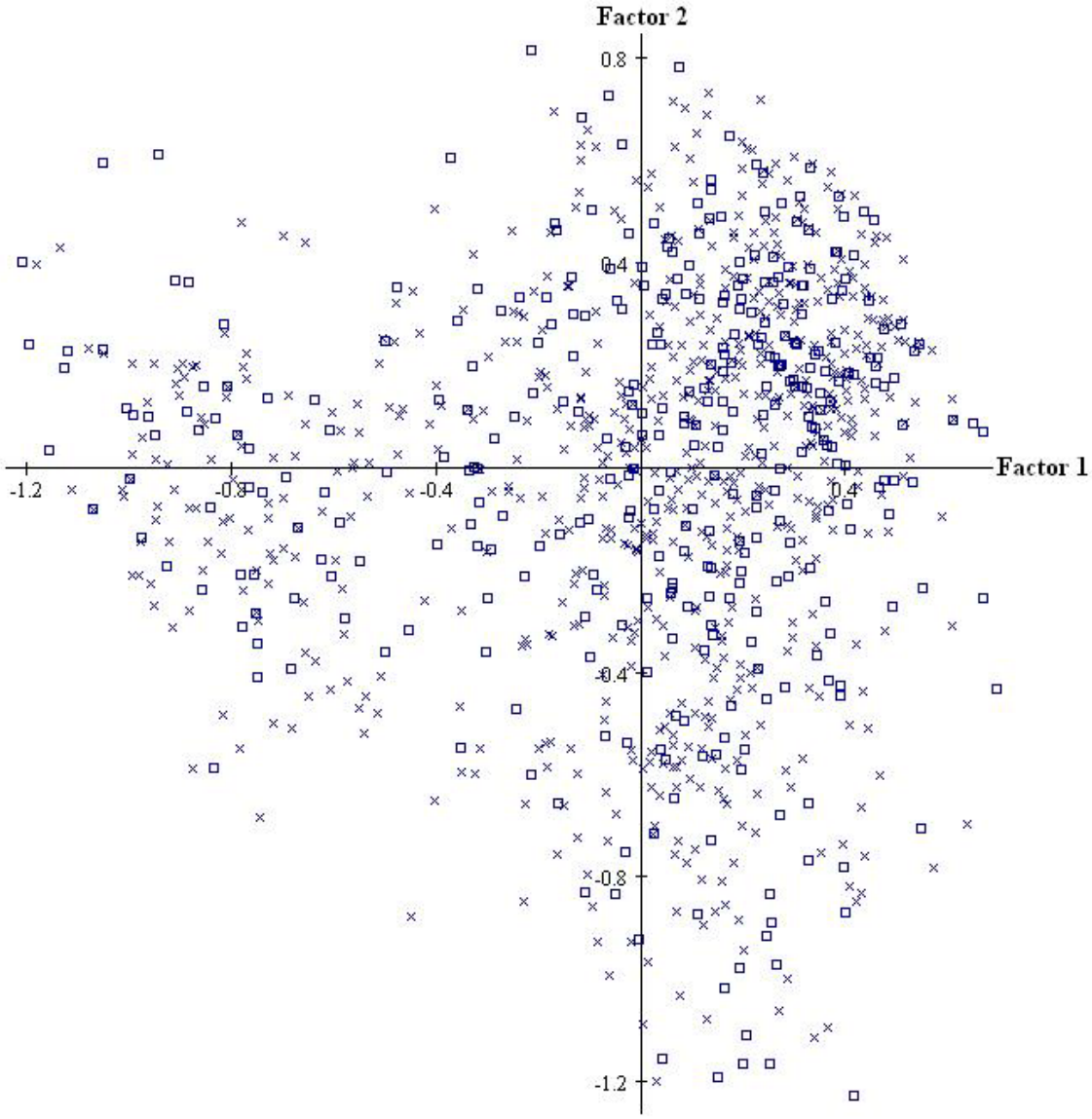


Figura 4.34 Representación de los individuos Donantes □ y Receptores x

Selección de las variables comunes

La tabla 4.10 presenta las distintas variables categóricas de las cuales se escogerán aquellas que se emplearán como variables comunes para la aplicación del método. Se utilizó la técnica de agrupación de variables. Se pretende agrupar variables semejantes entre ellas, para poner de manifiesto aquellas que aportan información parecida, dado que inicialmente desconocemos el número de grupos.

El resultado que se consideró adecuado fue de 15 grupos, lo cual da lugar a seleccionar 15 variables comunes categóricas. Se reduce por lo tanto la dimensión del problema.

Tabla 4.10 Conjunto de variables categóricas

Variables Comunes Categóricas:	*Grupo	Variables agrupadas
age	1	age
sex	2	sex
studies	3	studies
occupation	4	occupation
children	5	children
quickcon	6	quickcon flatrate webhome timewebh
flatrate	7	freqcomp freqweb
freqcomp	8	ofimatic
ofimatics	9	advanced
advanced	10	specific webwork timewebw
specifics	11	cleisure
cleisure	12	webstudy
webhome	13	catalan
webwork	14	spanish
webstudy	15	english
freqweb		
timewebhome		
timewebwork		
catalan		
spanish		
english		

* Clustering of Variables

Se puede ver en la tabla que existen tres grupos con más de una variable. Intuitivamente se puede apreciar la conexión entre estas variables y se ha decidido entonces seleccionar a una de cada grupo como representante del mismo.

Las variables seleccionadas fueron:

webhome (grupo 6)

freqcomp (grupo 7)

webwork (grupo 10)

Todas las variables son explicativas del consumo de Internet. La selección se ha efectuado por la semántica de la variable.

La figura 4.35 muestra las correlaciones de las variables específicas seleccionadas con los dos primeros ejes factoriales. Esta gráfica se complementa con la tabla 4.11 en la que se presentan las correlaciones de las variables con los dos primeros factores. Estas correlaciones las podemos comparar con el valor crítico de no significación dado por:

$$\frac{1.96}{\sqrt{n}} = 0.1016 .$$

Además se incluye en la tabla la suma de los cuadrados de las correlaciones en los 15 primeros factores[3], lo que representa la aportación de los factores a la variabilidad de cada variable.

La suma de estos cuadrados representa la variabilidad común debida a los factores y permite establecer la calidad de la representación de cada variable a partir de este conjunto de factores (comunalidad).

Tabla 4.11 Correlaciones de las variables especificas con los factores

	Factores		
Variables	1	2	Comunalidad
T_emailing	-0.05	0.74	0.71
T_chating	-0.68	-0.07	0.52
T_leisure	-0.36	0.24	0.58
T_buying	0.08	0.65	0.56
shopping	0.08	0.36	0.33
paying	0.10	0.60	0.53
T_banking	0.23	0.55	0.55
studying	0.11	0.42	0.47
T_government	0.13	0.55	0.44
gtransact	0.22	0.58	0.54
health	0.27	0.27	0.51
htransact	0.26	0.22	0.18
teleworking	0.07	0.18	0.27

[3] Más adelante se justificará la elección de 15 factores.

Se puede ver en la tabla que la comunalidad es baja para las variables:

shopping
htransact
teleworking

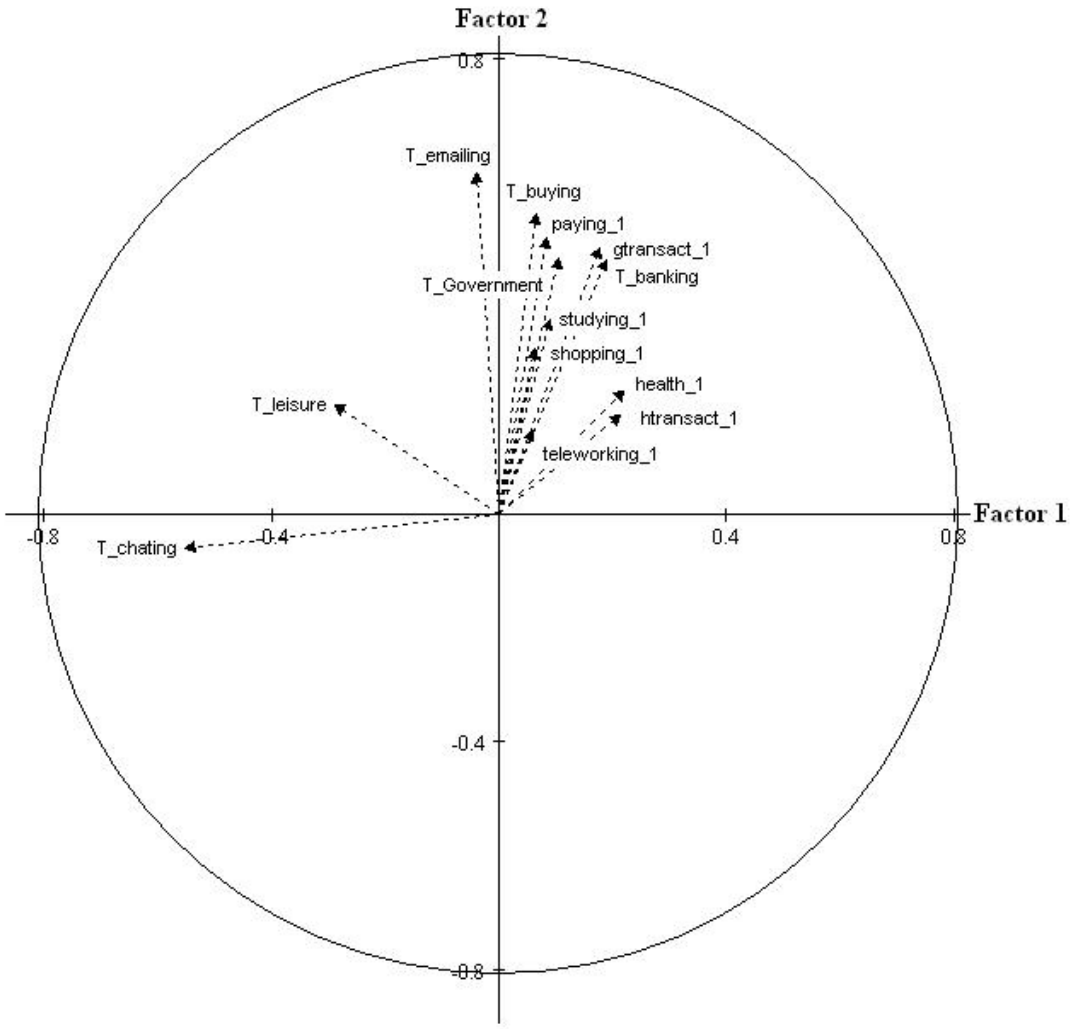


Figura 4.35 Diagrama de correlaciones en los dos primeros ejes factoriales

Determinación del número de ejes factoriales para la fusión

La figura 4.36 muestra el resultado de distintas corridas del sistema empleando el método de imputación TKSM en función del número de ejes factoriales. En este caso, para evaluar la calidad de la fusión efectuada y determinar los valores óptimos de los parámetros, ya

que estamos simulando un caso real, la comparación de los estadísticos ASL, Tau y ACD interno y externo, se ha efectuado entre los datos imputados de los donantes con sus valores reales observados, dado que los datos reales de los receptores no estarán disponibles en una fusión real.

En la figura se puede observar que el número adecuado de ejes factoriales es de 15. Los mejores resultados se obtienen con 15 ejes factoriales a excepción del índice de precisión.

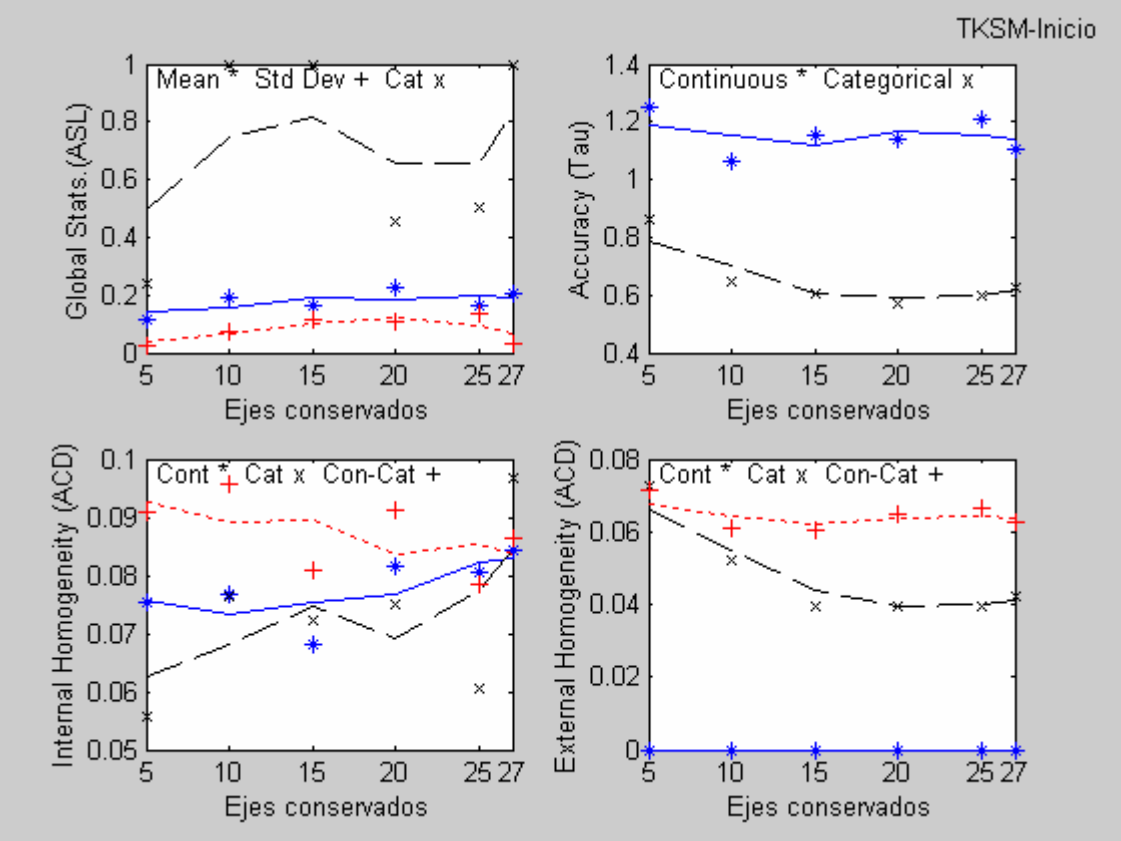


Figura 4.36 Aplicaciones con distintos ejes

4.10.2 Imputación TKSM. Determinación del número *k* de vecinos.

Procedimiento de imputación:	TKSM (Take K Stochastic Multivariate)
Parámetro:	K (número de vecinos)
Valor del Parámetro:	K = 5, 10, 15, 20, 25, 30, 35, 40
Archivo:	idescat

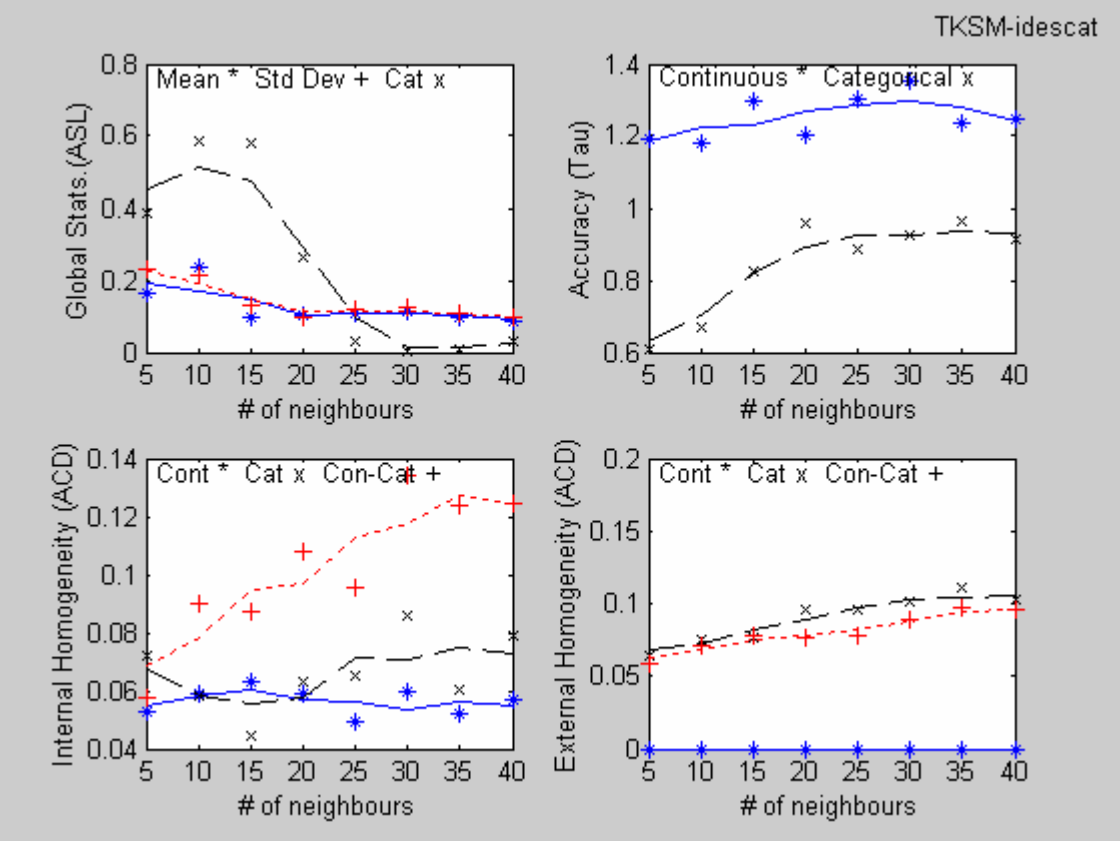


Figura 4.37 Imputación TKSM

Se puede ver en la figura 4.37 que el número de vecinos más cercanos adecuado para el experimento oscila entre 5 y 10. Los mejores resultados se obtienen para este número de vecinos. Los resultados obtenidos corresponden con lo esperado.

El índice ASL es superior a 0.05 en todos los casos (K = 5, 10, . . . , 40).

El índice de homogeneidad interna es pequeño 0.0531, lo mismo que en el caso de la homogeneidad externa 0.0583, lo cual constituyen valores aceptables para la fusión realizada.

Verificación de distribución equivalente entre los datos imputados y reales

Para verificar que efectivamente se ha obtenido un conjunto de valores imputados para las variables específicas que son "como los reales", se ha procedido a efectuar un análisis de componentes principales de la tabla de datos formada por los receptores y los valores de las variables especificas observadas. A continuación se proyectan los mismos individuos

receptores sobre el primer plano factorial, obtenido "según los valores imputados en las mismas variables". Para apreciar mejor la equivalencia entre ambas distribuciones, se ha procedido a comparar los valores reales observados en los receptores con sus valores imputados, si bien en general se haría entre los valores reales observados en los donantes y los valores imputados en los receptores.

Estos resultados se pueden observar en la figura 4.38, en donde se han representado en los dos primeros ejes factoriales los individuos donantes y los receptores con la diferencia ahora que los donantes son los individuos originales y los receptores, los individuos imputados.

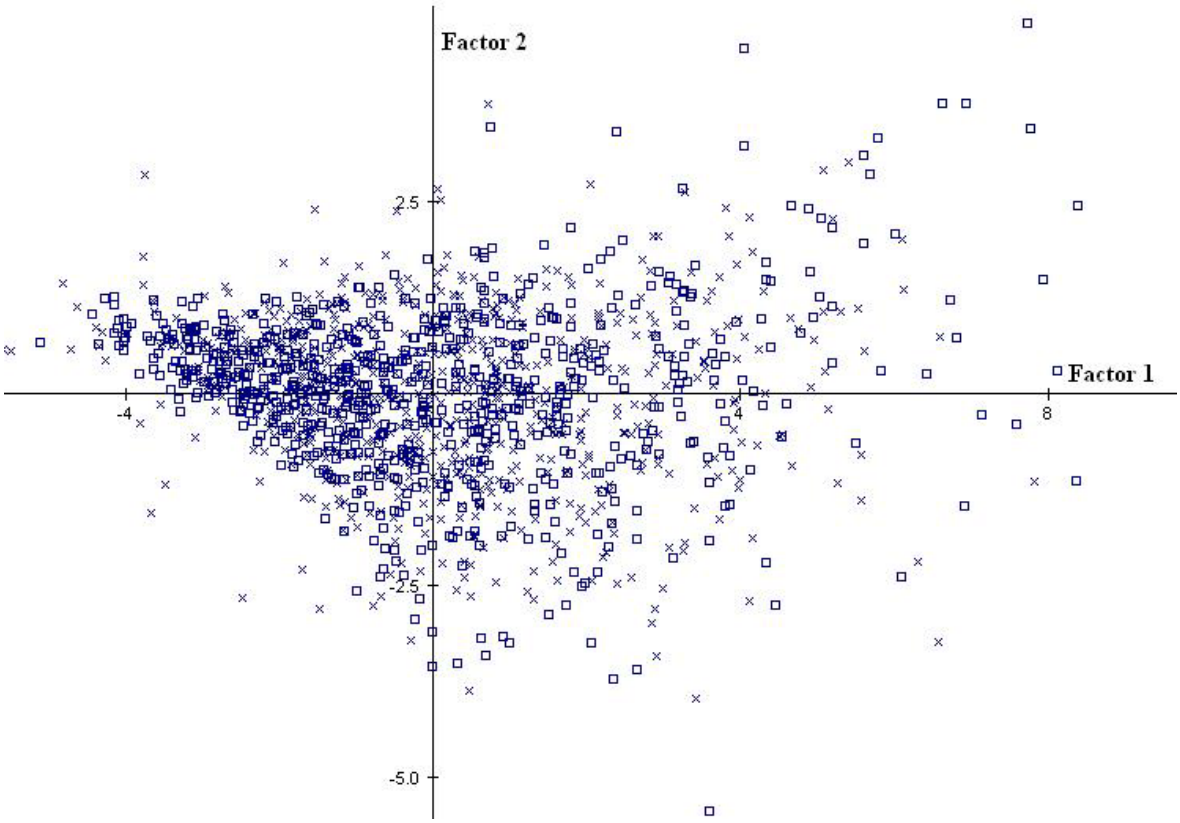


Figura 4.38 Individuos receptores: originales□ e imputados x

Para ver si la estructura de correlaciones entre los datos reales observados y los datos imputados es la misma, se representa el círculo de correlaciones de las variables activas (específicas observadas en la muestra de receptores), conjuntamente con las correlaciones de las variables específicas imputadas, con los mismos dos primeros ejes factoriales obtenidos en el análisis de componentes principales mencionado.

Puede observarse visualmente la equivalencia entre las dos estructuras de correlación (tabla 4.12, figura 4.39)

Tabla 4.12 Estructuras de correlación

Variable	Φ_1_recep_imp	Φ_2_recep_imp	Φ_1_recep_orig	Φ_2_recep_orig
emailing	0.814	-0.016	0.86	-0.062
chating	-0.305	0.537	-0.238	0.217
leisure	0.350	0.525	0.458	0.429
buying	0.866	-0.048	0.887	0.114
shopping	0.644	0.152	0.664	0.265
paying	0.818	0.148	0.836	0.241
banking	0.789	-0.184	0.795	-0.022
studying	0.665	-0.078	0.693	-0.045
government	0.835	-0.289	0.809	-0.181
gtransact	0.737	-0.229	0.722	-0.155
health	0.494	-0.614	0.435	-0.574
htransact	0.290	-0.706	0.38	-0.629
teleworking	0.469	0.230	0.577	0.391

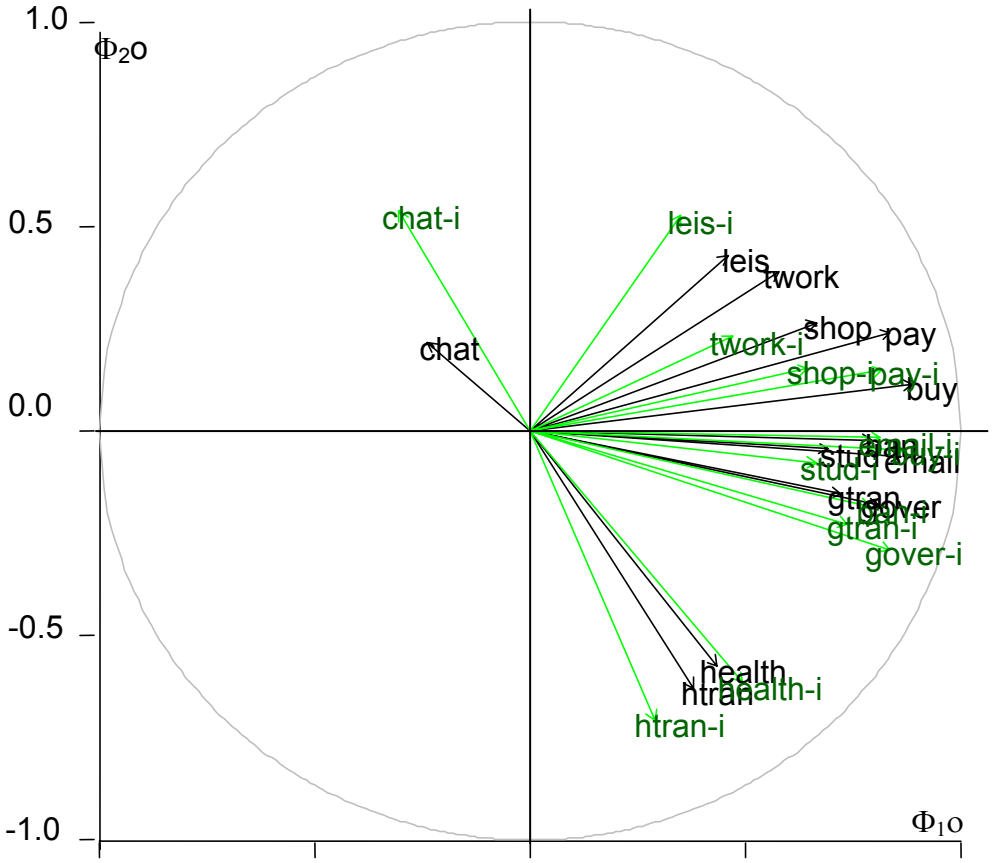


Figura 4. 39 Círculo de correlaciones

Resultados mediante imputación EM

Para evaluar la calidad de los resultados obtenidos con el método K-NN se ha repetido el experimento utilizando el método EM. Para poder comparar los resultados con la imputación previa realizada mediante el método TKSM, se ha agregado una componente aleatoria teniendo de esta manera lo que se llama aquí como: EM-R. Los resultados se presentan en la tabla 4.13.

En la tabla se presentan los resultados obtenidos mediante el método EM-R, comparando las imputaciones realizadas con los valores reales observados en los receptores. Se incluyen a su vez los resultados obtenidos mediante TKSM comparando los datos imputados con los valores observados de los receptores y con los valores observados de los donantes.

Tabla 4.13 Resultados del ensayo

	Validación con:		
Índice	TKSM Receptores	TKSM Donantes	EM-R Receptores
ASL	0.1902	0.1663	0.1880
Homogeneidad Interna (ACD)	0.0746	0.0531	0.0660
Homogeneidad Externa (ACD)	0.0595	0.0583	0.0508
Precisión(τ)	1.0274	1.1913	1.0728

Se puede apreciar en la tabla los buenos resultados obtenidos con el método propuesto. El índice de precisión muestra un valor que ya se esperaba. En el caso EM-R, los resultados son similares a los obtenidos con el sistema GRAFT.