Análisis de medidas repetidas mediante el uso de la curva media como síntesis.

Jorge Rodas
Departamento de Lenguajes y Sistemas Informáticos.
Universidad Politécnica de Cataluña.
email: jr@lsi.upc.es

J.Emilio Rojo
Serv. Psiquiatría. Hospital de Bellvitge.
Universidad de Barcelona.
email: jrojo@csub.scs.es

Karina Gibert
Departamento de Estadística e Investigación Operativa.
Universidad Politécnica de Cataluña.
email: karina@eio.upc.es

Ulises Cortés
Departamento de Lenguajes y Sistemas Informáticos.
Universidad Politécnica de Cataluña.
email: ia@lsi.upc.es

Febrero, 2003

Resumen

El presente documento reporta la comparación de los resultados obtenidos al aplicar la metodología KDSM[1] y el método de síntesis usando la media como función para el análisis de medidas seriadas muy cortas y repetidas con factor de bloque presentes en un *dominio poco estructurado* (DPE) del ámbito psiquiátrico.

Palabras Clave: Descubrimiento de Conocimiento, Medidas Seriadas y Clustering.

1 Introducción

El punto de partida del presente análisis tiene relación con el tipo de datos que presenta el estudio que realiza el Dr. J. Emilio Rojo del Servicio de Psiquiatría de la Ciudad Sanitaria y Universitaria de Bellvitge en Barcelona, que trata sobre el tiempo que tarda un paciente en reaccionar (Tiempo de Reacción) ante un estímulo, el cual puede ser: visual auditivo o ambos, justo después de la aplicación de electroshocks (Terapia Electroconvulsiva). Estos datos consisten en medidas seriadas muy cortas y repetidas del tiempo de reacción, datos relativos al electroshock y su aplicación e información sobre cada paciente tratado con electroshocks. Los detalles sobre dicha terapia y sus efectos se pueden consultar en [RGR01a]

Las medidas seriadas consisten en observaciones llevadas a cabo sobre una misma característica en diversos tiempos [Lin99]. Lo que las distingue de otras observaciones dentro de los modelos estadísticos tradicionales de datos son:

- El mismo atributo se mide sobre la misma unidad de observación más de una vez: ello implica que los atributos no son independientes como sucede en un análisis de regresión común y
- se involucra más de una unidad de observación: los atributos no conforman una serie de tiempo simple.

Las bases de datos que contienen medidas repetidas, suelen tener varias repeticiones—de las medidas seriadas del atributo de interés—para cada individuo incluido en el estudio. Sin embargo, no es correcto analizarlas todas juntas.

Una forma satisfactoria de analizar las medidas repetidas de cierto atributo para un conjunto de individuos es la propuesta de Matthews [Mat93]:

a) reducir las medidas repetidas a un conjunto pequeño de medidas independientes u obtener alguna función de síntesis apropiada para las observaciones en un objeto (promedio, área bajo la curva, etc) y

b) analizar esos sumarios, que ya son independientes entre sí, utilizando métodos estándar univariados.

[1] Del inglés Knowledge Discovery in Serial Measurement.

Por otra parte, en ocasiones la estructura de datos incluye un factor de bloque debido a los individuos en estudio o el número de medidas seriadas es tan pequeño que no permitirá un análisis clásico de series de tiempo. En [RGRC01] se propone toda una metodología para realizar de forma conveniente un análisis donde los datos tengan este tipo de estructura. Se analiza también que la reducción de todas las series de medidas de un mismo individuo a una curva síntesis (curva media) representa una pérdida de información demasiado valiosa, que puede presentar una idea errónea del real comportamiento del fenómeno objeto de estudio. En [RGRC01] se justifica que cuando las series de medidas son muy cortas resumirlas en un conjunto independiente de descriptores (como los coeficientes de la serie modelo) no es posible sencillamente porque con pocos datos no se pueden estimar. Por lo que decidimos realizar el presente análisis para ilustrar lo que sucede al utilizar la propuesta de Matthews [Mat93] de trabajar con síntesis de medidas.

2 Objetivo y Método de análisis

El objetivo de este análisis es valorar si la propuesta de Matthews [Mat93] es de utilidad para la aplicación introducida en [RGRC01]. Para ello, se realizarán las siguientes tareas:

1. Obtener la curva media de cada bloque definido por individuo.
2. Clasificación Basada en Reglas de las curvas medias obtenidas en el paso anterior, utilizando la base de conocimiento obtenida en [RGRC01].
3. Interpretación de los resultados por parte del experto y comparación de los mismos con respecto de los obtenidos en [RGRC01].

3 Cálculo de los valores medios

3.1 Datos

El estudio presentado en [RGRC01], da seguimiento a un conjunto 108 ES pertenecientes a 13 pacientes que presentan desórdenes depresivos o esquizofrenia y que se han tratado con TEC durante un cierto tiempo en el servicio de psiquiatría de la Ciudad Sanitaria y Universitaria de Bellvitge (CSUB).

Tras cada sesión se miden los *tiempos de reacción*[2] basados en estímulos visuales y auditivos: *Tiempo de reacción visual simple* (prueba S5), *tiempo de reacción auditivo simple* (prueba S6), *tiempos de categorización y de reacción visual: Visual Complejo* (prueba S7) y *tiempos de categorización y de reacción visoauditivo: Visoauditivo Complejo* (prueba S8).

Las pruebas se realizan en varias ocasiones a cada paciente: el día previo a la TEC (basales) y después de la aplicación de cada electroshock a las 2, 4, 6, 12 y 24 horas. Así se obtienen por un lado las medidas seriadas muy cortas (6 tomas de medidas) y son repetidas pues se vuelven a tomar tras cada ES aplicado (más detalles en [RGR01b]).

[2]En este caso, el tiempo de reacción (TR) es el tiempo entre la emisión de un estímulo y el momento en que el paciente reacciona ante dicho estímulo.

	t_1	t_2	...	t_R	
E_{10}	Y_{10}^1	Y_{10}^2	...	Y_{10}^r	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	bloque 1
E_{1n_1}	$Y_{1n_1}^1$	$Y_{1n_1}^2$	...	$Y_{1n_1}^r$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	bloque 2
E_{2n_2}	$Y_{2n_2}^1$	$Y_{2n_2}^2$	...	$Y_{2n_2}^r$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
E_{n0}	Y_{n0}^1	Y_{n0}^2	...	Y_{n0}^r	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	bloque n
E_{nn_i}	$Y_{nn_i}^1$	$Y_{nn_i}^2$	...	$Y_{nn_i}^r$	

Tabla 1: *Bloques de series formados por los individuos del estudio.*

3.2 Cálculo de los valores medios

Para realizar el cálculo de los valores medios el procedimiento consiste en sustituir las n_i filas de cada bloque (Tabla 1) Y_{ij}^t donde $i = \{1 \dots n\}$ es el individuo, $j = \{0 \dots n_i\}$ indica la j-ésima ocurrencia de E en el individuo i y $t \in \{1 \dots r\}$ (donde r es muy pequeño) $j = \{1 \dots n_i\}$ por una sola curva media $\bar{Y_i}^t = \frac{\sum_j Y_{ij}^t}{n_i}$.

Nota: La totalidad de los datos con que se realizó el cálculo de los valores medios se encuentra en [RGR01b].

4 Clasificación

Para poder comparar los resultados que se obtengan en este análisis con la propuesta de Matthews [Mat93] y los obtenidos previamente en [RGRC01] se utilizó la metodología de Clasificación Basada en Reglas (ClBR) [Gib94] ya que consiste en una estrategia mixta que combina la gestión de una Base de Conocimiento (IA) con un análisis cluster (Estad.). Así, el experto puede aportar una base de conocimiento (BC) con el conocimiento adicional, expresado en reglas de lógica de primer orden sobre su campo de estudio. Procesando las reglas, se detecta una primera estructura en los pacientes (eventualmente no completa) que se combina con el resultado de un proceso de cluster jerárquico hasta obtener un único dendrograma con todos los elementos (ver [RGRC01]).

Así que, utilizamos la misma base de conocimiento para que los resultados sean comparables. El estudio [RGRC01], quiere observar la estructura de los ES a partir de las curvas obtenidas en las primeras 24 horas siguientes a cada sesiónde ES. No obstante, para aplicar la ClBR es conveniente mantener la independencia entre las filas de datos. La matriz de ES, como se ha visto antes (Tabla 1), contiene bloques con todas las curvas relativas a un mismo paciente, las que por supuesto no son independientes entre sí, por estar influenciadas por características propias de cada paciente que se mantienen constantes en cada *bloque-paciente*.

4.1 ClBR de Medidas Seriadas Medias

De acuerdo con la propuesta de Matthews redujimos cada bloque de medidas seriadas a una sola (media) que representa a todo el bloque. Con este procedimiento de síntesis en lugar de clasificar 108 medidas seriadas (una por cada ES aplicado a cada paciente) se clasificaron 11 series medias (1 por paciente). Seguidamente realizamos la ClBR utilizando como Base de Conocimiento adicional la misma que en el estudio [RGRC01], la cual se compone de 2 reglas simples:

1. Si EDAD $\leq 40 \rightarrow$ pacientes jóvenes y,
2. Si EDAD $> 50 \rightarrow$ pacientes mayores.

Se obteniene el dendrograma de la Figura 1.

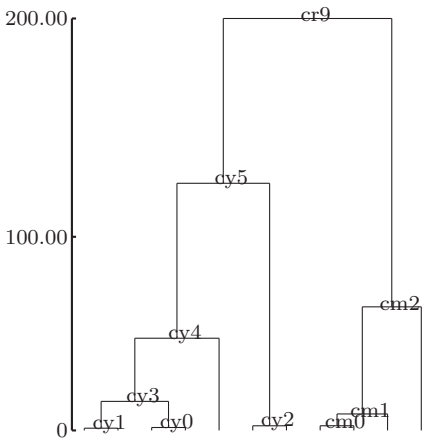


Figura 1: Árbol general de clasificación.

El análisis de la estructura del dendrograma Figura 1, sugiere un corte en 3 clases: dos clases (cy4 y cy2) del grupo de los pacientes jóvenes y la otra (m2) del grupo de los pacientes mayores.

Las 3 clases están conformadas de la siguiente forma:

- *cy4*= pac01, pac05, pac08, pac10 y pac13,
- *cy2*= pac03 y pac06, y
- *cm2*= pac02, pac07, pac11 y pac12.

4.2 ClBR de Medidas Seriadas

Del estudio [RGRC01] reproducimos la estructura del dendrograma Figura 2 obtenida para poder visualizar de forma clara la diferencia entre utilizar la propuesta de Matthews y no hacerlo.

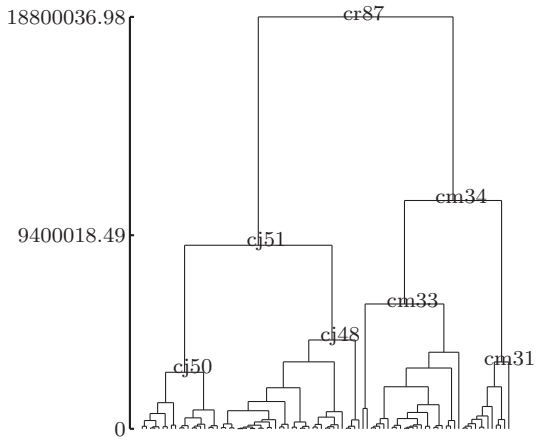


Figura 2: Árbol general de clasificación.

El análisis de la estructura del dendrograma Figura 2, sugirió 4 clases: dos clases (cj50 y cj48) del grupo de los pacientes jóvenes y otras dos (cm33 y cm31) del grupo de los pacientes mayores.

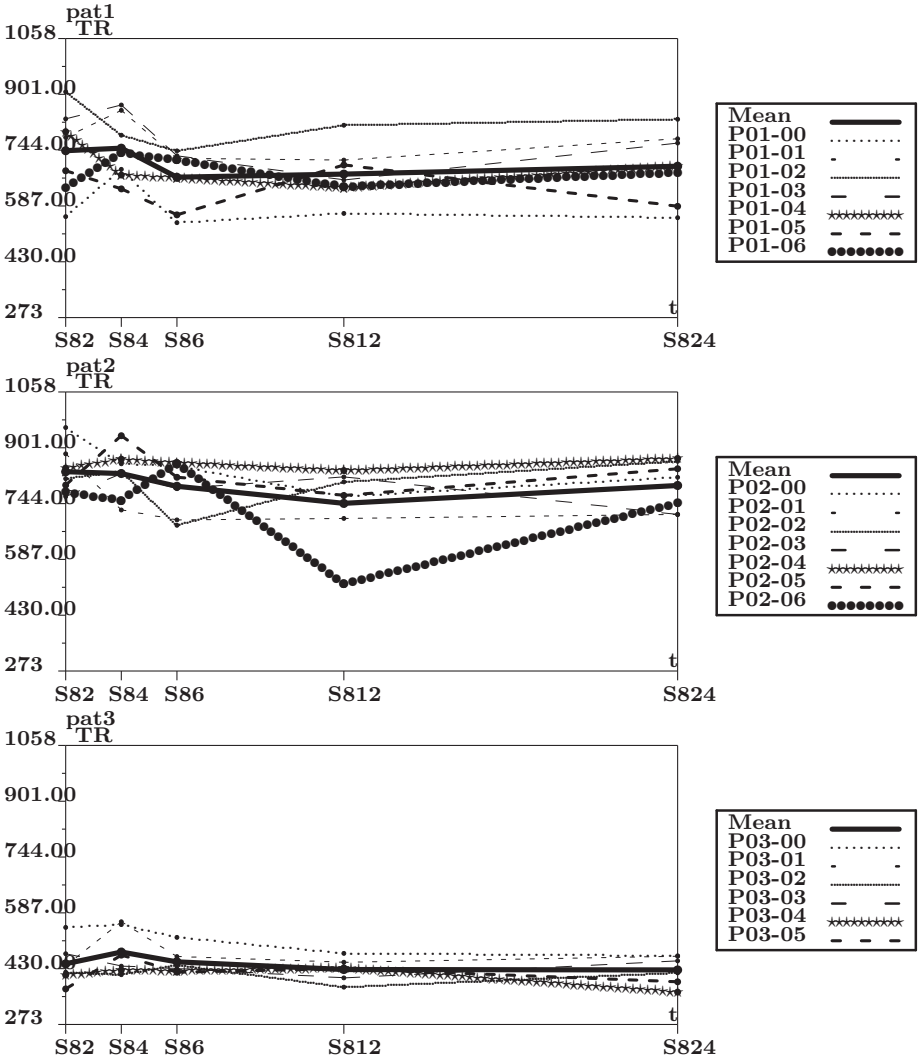
Las 4 clases están conformadas de la siguiente forma:

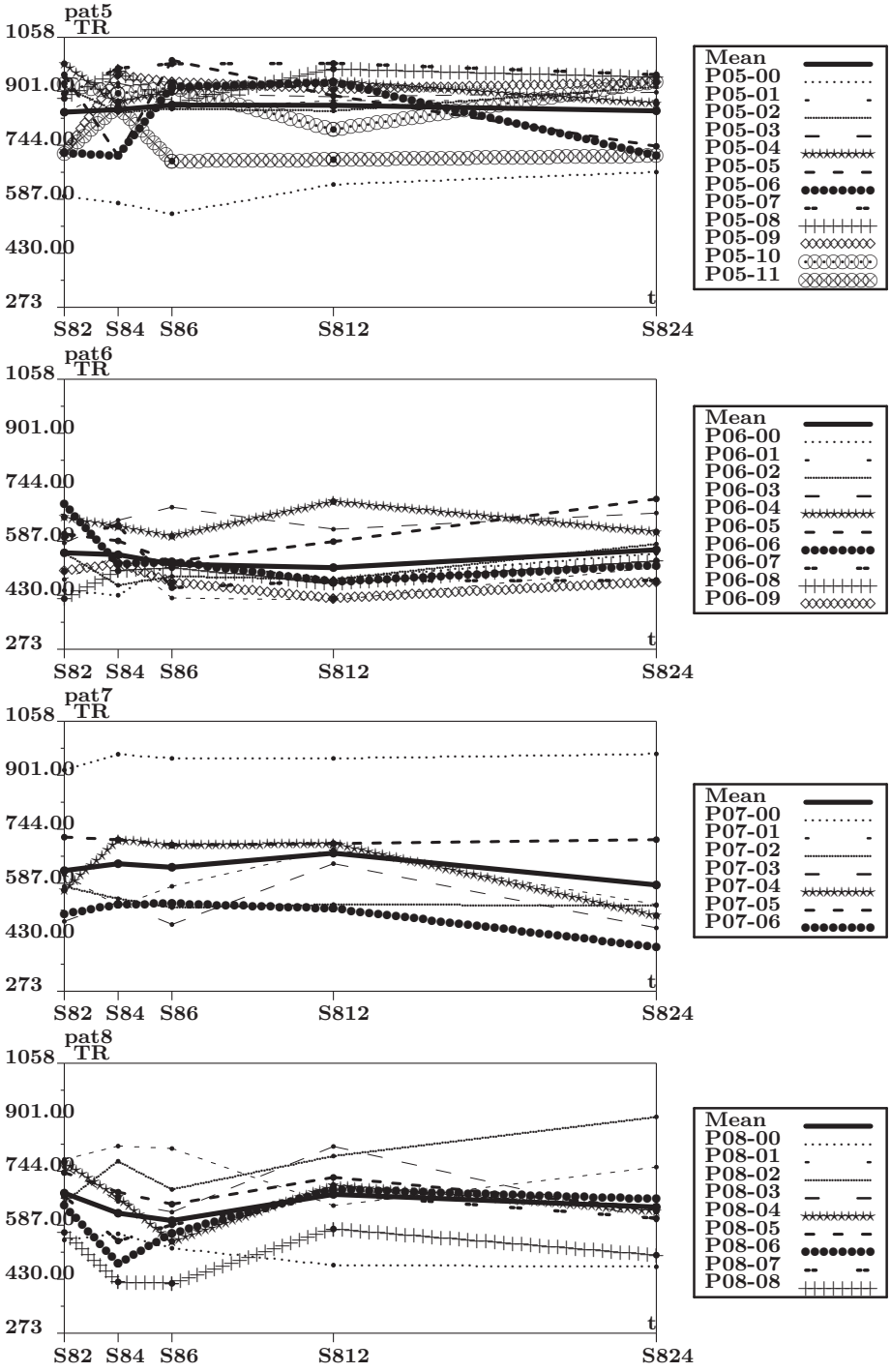
- *cj50*=$\{E_{133}, E_{511}, E_{135}, E_{32}, E_{102}, E_{84}, E_{510}, E_{131}, E_{11}, E_{67}, E_{13}, E_{59}, E_{88}, E_{15}, E_{65}, E_{56}, E_{103}, E_{86}$ y $E_{82}\}$,
- *cj48*=$\{E_{105}, E_{68}, E_{101}, E_{51}, E_{134}, E_{33}, E_{69}, E_{35}, E_{34}, E_{106}, E_{104}, E_{58}, E_{52}, E_{64}, E_{53}, E_{16}, E_{55}, E_{132}, E_{12}, E_{62}, E_{61}, E_{81}, E_{87}, E_{85}, E_{31}, E_{107}, E_{14}, E_{54}, E_{66}, E_{83}, E_{136}, E_{108}, E_{63}$ y $E_{57}\}$,
- *cm33*=$\{E_{115}, E_{22}, E_{21}, E_{113}, E_{75}, E_{128}, E_{112}, E_{23}, E_{1211}, E_{114}, E_{72}, E_{25}, E_{1210}, E_{123}, E_{71}, E_{116}, E_{111}, E_{24}, E_{124}, E_{74}$ y $E_{126}\}$ y
- *cm31*=$\{E_{1212}, E_{117}, E_{122}, E_{129}, E_{73}, E_{26}, E_{76}, E_{127}, E_{125}$ y $E_{121}\}$.

5 Discusión

Se puede observar que en la Figura 3 que la media en cada paciente ejerce un efecto de suavizado sobre las curvas, sin tendencia clara durante las 24 horas. Lo anterior indica que los tiempos de reacción (TR) se mantienen básicamente constantes, lo cual es falso. Se compensan unos valores a otros pero el TR registrados por los pacientes *nunca* es estable. Precisamente ocurre porque en un mismo paciente se dan curvas con tendencias contrarias, pues en ciertas aplicaciones de ES los pacientes *reducen* su tiempo de reacción y en otras *aumentan*) y cuando se realiza la media la tendencia se anula.

Del gráfico de medias Figura 4, obtenido del análisis de la clasificación siguiendo la propuesta de Matthews, se puede ver que solo la clase y2 muestra una reducción en su tiempo de reacción en las 24 horas. Aunque no tiene ningún sentido en cuanto a valorar


pat1
TR
1058
901.00
744.00
587.00
430.00
273
t
S82 S84 S86 S812 S824
Mean
P01-00
P01-01
P01-02
P01-03
P01-04
P01-05
P01-06
pat2
TR
1058
901.00
744.00
587.00
430.00
273
t
S82 S84 S86 S812 S824
Mean
P02-00
P02-01
P02-02
P02-03
P02-04
P02-05
P02-06
pat3
TR
1058
901.00
744.00
587.00
430.00
273
t
S82 S84 S86 S812 S824
Mean
P03-00
P03-01
P03-02
P03-03
P03-04
P03-05


pat5
TR
1058
901.00
744.00
587.00
430.00
273
t
S82
S84
S86
S812
S824
Mean
P05-00
P05-01
P05-02
P05-03
P05-04
P05-05
P05-06
P05-07
P05-08
P05-09
P05-10
P05-11
pat6
TR
1058
901.00
744.00
587.00
430.00
273
t
S82
S84
S86
S812
S824
Mean
P06-00
P06-01
P06-02
P06-03
P06-04
P06-05
P06-06
P06-07
P06-08
P06-09
pat7
TR
1058
901.00
744.00
587.00
430.00
273
t
S82
S84
S86
S812
S824
Mean
P07-00
P07-01
P07-02
P07-03
P07-04
P07-05
P07-06
pat8
TR
1058
901.00
744.00
587.00
430.00
273
t
S82
S84
S86
S812
S824
Mean
P08-00
P08-01
P08-02
P08-03
P08-04
P08-05
P08-06
P08-07
P08-08

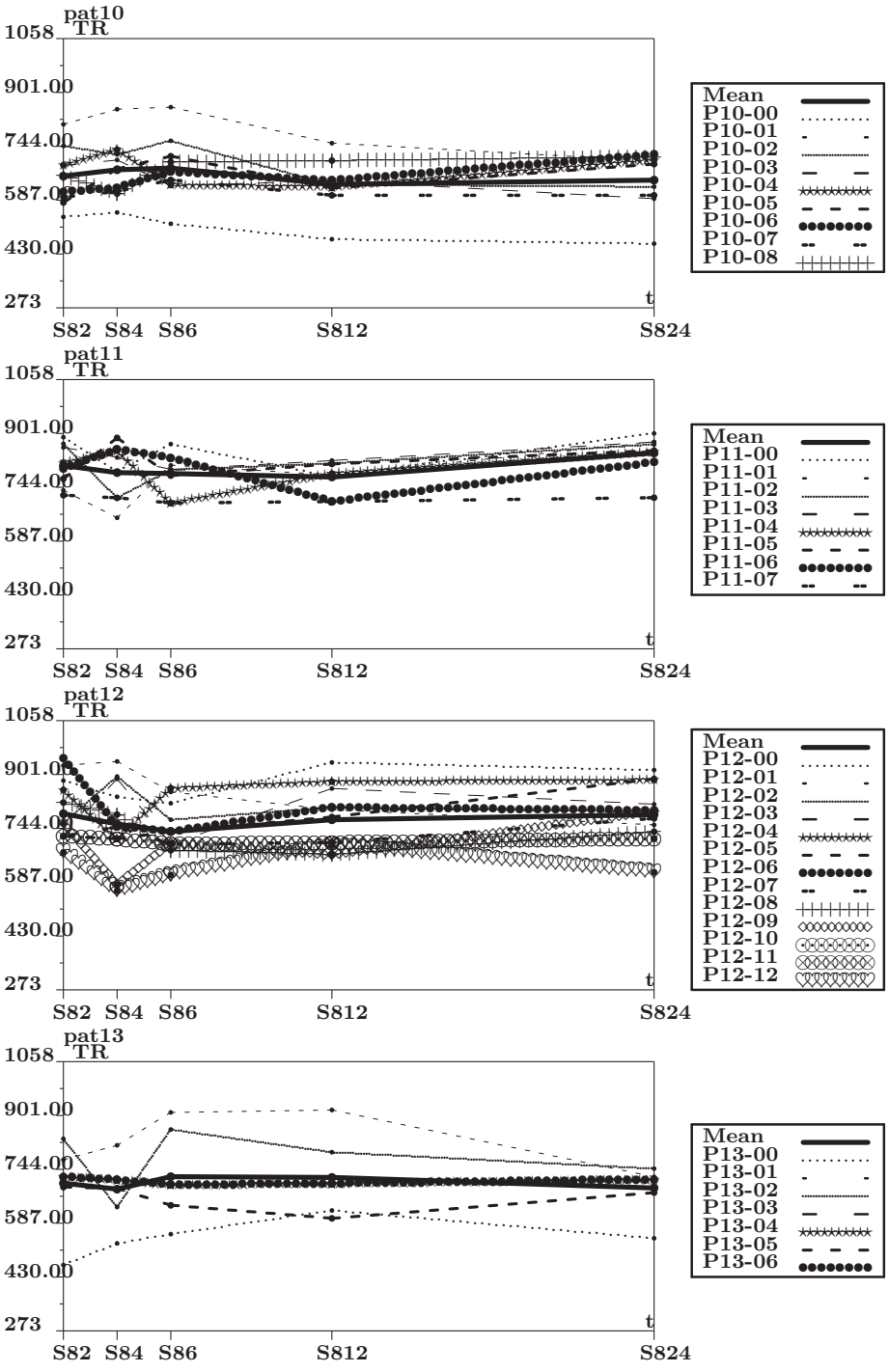


Figura 3: Variabilidad interna de las clases (prueba S8).

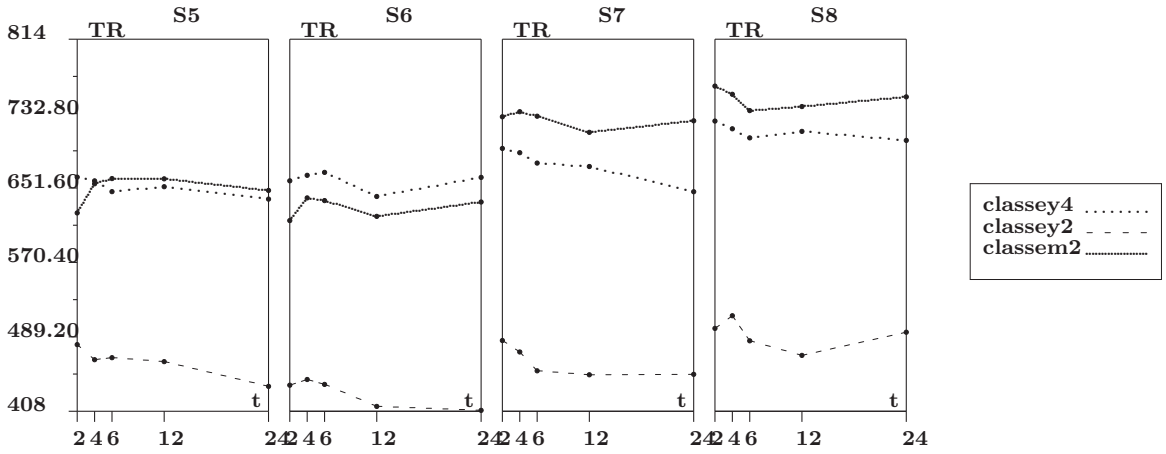


Figura 4: Variabilidad entre las clases (pruebas S5-S8).

el efecto del tratamiento a lo largo del tiempo ya que no podemos conocer la influencia real que ejerce cada ES aplicado, si hemos utilizado la media de cada paciente.

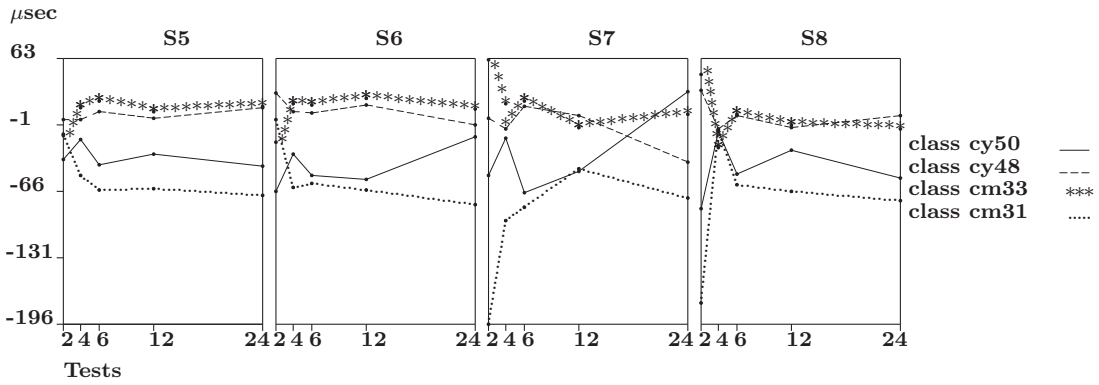


Figura 5: Four class curves for tests S5 to S8.

En cambio, cuando no utilizamos la media de las medidas seriadas, el gráfico de variabilidad entre clases Figura 5 si indica como evoluciona el tiempo de reacción en determinada clase de forma veraz a lo largo del tratamiento.

6 Conclusiones

En la Tabla tal realizamos el cruce de las particiones obtenidas (propuesta de Matthews y sin ella), para valorar la asociación entre ambas clases.

De la tabla 2 obtenemos la conclusión mas importante de este estudio. En el caso de la partición sin utilizar la propuesta de Matthews (cy50, cy48, cm33 y cm31) se ve que los ES no se asocian a una sola clase. Lo que indica que cada ES tiene un efecto sobre la evolución del paciente a lo largo de la terapia. Al utilizar la propuesta de Matthews esta información tan importante se pierde.

Partición	Class cy50	Class cy48	Class cm33	Class cm31
Class y4	y4: p1,5,8,10,13 y50: p1,3,5,6,8,10,13	y4: p1,5,8,10,13 y48: p1,3,5,6,8,10,13		
Class y2	y2: p3,6 y50: p1,3,5,6,8,10,13	y2: p3,6 y48: p1,3,5,6,8,10,13		
Class m2			m2: p2,7,11,12 m33: p2,7,11,12	m2: p2,7,11,12 m31: p2,7,11,12

Tabla 2: *Relación entre las particiones.*

Esto no indica que la propuesta de Matthews no es válida, sino que no en todos los casos donde se presenten medidas seriadas en bloque será conveniente utilizar alguna función de síntesis.

Como reflexión final, podría ser interesante iniciar el estudio realizando una Tabla como la Tabla X para determinar si la propuesta Matthews sería útil o no.

Referencias

[Gib94] K. Gibert. *L'us de la Informació Simbòlica en l'Automatització del Tractament Estadístic de Dominis poc Estructurats.* In the statistics and operations research phd. thesis, Universitat Politecnica de Catalunya, Barcelona, Spain, 1994.

[Lin99] J.K. Lindsey. *Models for Repeated Measurements.* Oxford University Press, Great Britain, second edition, 1999. ISBN: 0-19-850559-0.

[Mat93] J.N.S. Matthews. A refinament to the analysis of serial data using summary measures. *Statistics in Medicine*, 12:27–37, 1993. Wiley.

[RGR01a] J. Rodas, K. Gibert, and J. Rojo. Electroshock Effects Identification Using Classification Techniques. *Springer's Lecture Notes of Computer Science Series*, Crespo, Maojo and Martin (Eds.):238–244, 2001. Second International Symposium, ISMDA 2001.

[RGR01b] J. Rodas, K. Gibert, and J. Rojo. Influential factors determination on an ill-structured domain response. Research LSI-01-6-R, Technical University of Catalonia, Barcelona, Spain, March 2001. http://www.lsi.upc.es/dept/techreps/html/R01-6.html.

[RGRC01] J. Rodas, K. Gibert, J. Rojo, and U. Cortés. A methodology of knowledge discovery in serial measurement applied to psychiatric domain. Research LSI-01-53-R, Technical University of Catalonia, Barcelona, Spain, December 2001. http://www.lsi.upc.es/dept/techreps/html/R01-53.html.