Representación de series de tiempo utilizando lógica difusa

Antonio Moreno-Garcia, Juan Moreno-Garcia, Luis Jimenez-Linares, and Luis Rodriguez-Benitez

Departamento de Tecnologías y Sistemas de Información,
Universidad de Castilla-La Mancha, España,
antmorgarcia@gmail.com, {juan.moreno,luis.jimenez,luis.rodriguez}@uclm.es

Abstract. El objetivo de este trabajo consiste en presentar una representación difusa de series de tiempo, denominada "fuzzy piecewise linear segment", que represente de una forma simplificada y eficiente la serie. Esta representación también recoge la incertidumbre asociada debido al error en la generación de los segmentos. Para la obtención de nuestra representación son necesarios dos pasos, primeramente se obtendrá una representación de la series de tiempo basada en segmentos denominada comúnmente en la literatura *piecewise linear segment*, y posteriormente, se convertirá a una representación difusa que podrá ser utilizada en diferentes aplicaciones. Se presentan algunos ejemplos de cómo nuestra propuesta puede representar apropiadamente la serie de entrada.

Keywords: series de tiempo; representación difusa, consultas en bases de datos

1 Introducción

Las series de tiempo (TS) se utilizan en una gran cantidad de aplicaciones, por ejemplo en: procesamiento de imagen [1], economía [2, 3], ciencias sociales [4, 5] o deporte [6]. Los datos son tomados en bruto por sensores o sistemas de captura de información para ser almacenados y posteriormente ser consultados. Por todo esto se realiza en la actualidad una investigación intensiva en esta línea.

Las TS representan los datos como "datos en bruto", es decir, un conjunto de valores tomados consecutivamente de algún dispositivo o sistema y que se toma una muestra igualmente espaciada en el tiempo. Formalmente, se pueden representar mediante la Ecuación 1.

$$Y = \{y_1, y_2, \ldots, y_n\} \tag{1}$$

donde y_i es el valor de la TS en el instante i y $1 \leq i \leq n$.

Esta forma de representación de los datos tiene varios inconvenientes. Uno de los más importantes es que necesitan una gran cantidad de memoria para ser almacenados, por ejemplo, un sensor que tome una muestra cada cierto intervalo de tiempo pequeño y que esté tomando datos constantemente necesitará una

importante cantidad de memoria de almacenamiento, y esto únicamente para un sensor. Tal cantidad de información dificultará las operaciones sobre la TS, incrementando la dificultad de la misma y su tiempo de ejecución.

Por ello, se necesita otro tipo de representación que consuma menos memoria y permita operaciones sobre ella de forma más rápida y eficiente. Una de las formas más usadas consiste en la denominada "*piecewise linear segment*" que consiste en representar la serie mediante un conjunto de segmentos cada uno de los cuales representa un intervalo temporal de la serie. Una primera revisión de la literatura relativa se presenta en [7, 8] (Grupo dirigido por Keogh y que sin duda es uno de los grupos más importantes que investigan en este campo) y una más actualizada en [9]. La primera propuesta para segmentar una serie fue de Shatkay [10, 11] utilizando una secuencia de funciones matemáticas. Posteriormente Keogh y otros [7] presentaron el algoritmo SWAB (*Sliding Window and Bottom-up*) que utiliza la técnica Bottom-Up y el mecanismo de ventana deslizante. Fuchs y otros [12] crean una técnica denominada *SwiftSeg* mediante aproximación polinomial de TS utilizando utilizando bases de polinomios ortogonales en ventana deslizante y/o creciente. Huang y otros [9] hacen uso de la interpolación como herramienta. García-Treviño y Barria [13] presentan un marco de representación de TS que utiliza la dependencia de datos de las mismas junto con algoritmos de clasificación estadística.

Esta forma de representación es muy efectiva y rápida en cuanto a tiempo de ejecución. Pero tiene el problema de que aumenta el error debido a la propia transformación de los datos en bruto al conjunto de segmentos obtenido. Además, siempre se presenta error asociado a la propia toma de datos debido al ruido. Una forma de solucionar esto es utilizando la lógica difusa que permite tratar la incertidumbre introducida en la captura de datos y la posterior transformación a segmentos.

Kacprzyk y otros [14] presentaron un método para la creación de resúmenes lingüísticos a partir de TS. Es un trabajo muy interesante y detalla una propuesta de proceso que consiste en los siguientes pasos:

1. Generación de tendencias. Calcula el conjunto de segmentos con un método creado y que se basa en el algoritmo de Sklansky y Gonzalez [15].
2. Representación de las características de la dinámica de la TS. Presentan una metodología que se basa en la lógica difusa y que se detallará más adelante.
3. Generación de resúmenes lingüísticos. Utiliza protoformas y cálculo de proposiciones cuantificadas lingüísticamente para la obtención de los resúmenes. Es la parte principal del trabajo y se sale de los objetivos de este trabajo.

Dentro del objetivo de este trabajo nos interesa la segunda fase que representa los segmentos mediante lógica difusa. Los autores consideran tres aspectos que son:

– Dinámica del cambio: Lo definen como la velocidad de cambio. Para ello utilizan la pendiente de la tendencia. Definen funciones de pertenencias difusas para obtener una granularidad difusa. Por ejemplo, presentan un esquema que puede clasificar la pendiente como *rápidamente decreciente*, *de-*

creciente, *lentamente decreciente*, *constante*, *lentamente creciente*, *creciente* y *rápidamente creciente*.
- Duración: La definen como "la longitud de una tendencia". Utilizan una variable lingüística que la denominan "*long trend*".
- Variabilidad: Es la "extensión vertical de un grupo de datos". Se refiere a si la tendencia está bien modelada con el segmento o hay mucha incertidumbre, mucho ruido. Proponen cinco medidas estadísticas para medirla.

Nuestra propuesta va en otro sentido, y modela los segmentos de manera difusa para poder realizar operaciones entre segmentos, valor crisp-segmento, etc. Se propone el uso de números difusos como salida ante un valor de entrada. Estos números son obtenidos automáticamente a partir de los segmentos basándose en una medida del error que hace uso de los valores originales de la TS en el dominio que el segmento cubre. Los aspectos que Kacprzyk propone pueden ser obtenidos desde los segmentos difusos que hace nuestra propuesta. La duración como tal no se utiliza, aunque puede ser calculada en base a los instantes inicial y final de nuestra propuesta. Respecto a la dinámica se puede calcular utilizando el ángulo que define el segmento.

Por otro lado, ya existe en la bibliografía el concepto de "segmento difuso" que fue presentado por Hoover [16]. Estos autores buscaban la detección de nervios ópticos en una imagen del fondo del ojo mediante el uso de estos segmentos. Éstos fueron diseñados para detectar las líneas de convergencia en las formas de una imagen. Utilizan doce imágenes para los test obteniendo un 65% de tasa de éxito.

El documento está estructurado de la siguiente forma. En la Sección 2 se detalla nuestra propuesta. A continuación se muestra un ejemplo de uso y algunos ejemplos y pruebas realizadas con nuestra aproximación en la Sección 3, también se presentará una discusión (Sección 3.1). Finalmente, la Sección 4 muestra las conclusiones y trabajos futuros.

2 Fuzzy piecewise linear segment

Antes de detallar nuestra propuesta en necesario indicar que el primer paso consiste en la transformación de la TS a un conjunto ordenado de segmentos. Para realizar esta conversión se puede utilizar cualquier método de la literatura. Nosotros hemos utilizado un método original propuesto por nuestro grupo que se basa en un mecanismo de ventana deslizante y que tiene un coste computacional bajo. No ha sido publicado todavía. Este método obtiene como salida un conjunto de segmentos que será utilizado como entrada para el desarrollo presentado aquí. El conjunto de segmentos se representa mediante la Ecuación 2.

$$S = \{s_{f_1,l_1}, s_{f_2,l_2}, \ldots, s_{f_m,l_m}\} \tag{2}$$

donde cada segmento $s_{f_k,l_k} = (m_{f_k,l_k} * x) + c_{f_k,l_k}$ se define en el intervalo de tiempo desde el primer instante f_k al último instante l_k. Como puede verse, se representa mediante la ecuación de la recta.

Un segmento s_{f_k,l_k} será formalmente representado mediante la siguiente 2-tupla: $s_{f_k,l_k} = \{m_{f_k,l_k}, c_{f_k,l_k}\}$.

La idea de nuestra propuesta consiste en obtener una representación que consta de un conjunto de segmentos difusos. Esto quiere decir que ante una entrada $t_i \in \mathbb{R}$ se obtendrá un número triangular difuso de salida fn_i: $fs(t_i) = fn_i$.

Nuestra propuesta de representación se denomina "Fuzzy Piecewise Linear Segment" ($FPLS$). Una $FPLS$ se modeliza formalmente mediante la Ecuación 3.

$$FPLS(T) = \{fpls_{f_0,l_0}, fpls_{f_1,l_1}, \dots fpls_{f_{|FPLS|-1},l_{|FPLS|-1}}\} \quad (3)$$

donde cada $fpls_{f_k,l_k} = \{m_{f_k,l_k}, c_{f_k,l_k}, p_{f_k,l_k}\}$ y p_{f_k,l_k} es la media del ratio de error que viene definido por la Ecuación 5.

Un segmento difuso $fpls_{f_k,l_k}$ se modela mediante una 3-tupla $\{m_{f_k,l_k}, c_{f_k,l_k}, p_{f_k,l_k}\}$ que contiene la pendiente m_{f_k,l_k}, la ordenada c_{f_k,l_k} y la media del ratio de error p_{f_k,l_k} de ese segmento. Con estos tres valores se puede calcular el conjunto difuso de salida fn_i ante un valor de entrada t_i. La Ecuación 4 muestra cómo un $fpls_{f_k,l_k}$ genera el número difuso triangular de salida fn_i cuyo soporte viene definido por los puntos fn_i^a, fn_i^b y fn_i^c.

$$fpls_{f_k,l_k}(t_i) = \begin{cases} si\ t_i < f_k: & no\ salida \\ si\ f_k \leq t_i \leq l_k: & \{fn_i^a, fn_i^b, fn_i^c\} = calcular(s_{f_k,l_k}, t_i) \\ si\ l_k < t_i: & no\ salida \end{cases} \quad (4)$$

donde $calcular(s_{f_k,l_k}, t_i)$ es una función que calcula fn_i^{DOWN}, fn_i^s y fn_i^{UP}; y finalmente genera el conjunto difuso triangular de salida $fn_i = \{fn_i^a, fn_i^b, fn_i^c\}$.

Para valores de t_i más pequeños que f_k no ofrece salida, ya que el segmento no está definido para esos valores del dominio. Para valores mayores que l_k ocurre lo mismo. Y para valores comprendidos entre f_k y l_k se calcula utilizando dos segmentos paralelos a s_{f_k,l_k} denominados UP_{f_k,l_k} y $DOWN_{f_k,l_k}$ (Figura 1). También hace uso de una medida del error que se ha denominado "media del ratio de error" (Ecuación 5). Esta ecuación mide el valor promedio del ratio del error estimado para cada segmento.

$$p_{f,l} = \frac{\sum_{i=f}^{l} \frac{|s_{f,l}(t_i) - y_i|}{y_i}}{l - f + 1} \quad (5)$$

donde f y l son los instantes de comienzo y fin del segmento y $s_{f,l}(t_i)$ es el valor del segmento $s_{f,l} \in S$ en el instante t_i.

p_{f_k,l_k} se utiliza para calcular los segmentos UP_{f_k,l_k} y $DOWN_{f_k,l_k}$ que están desplazados hacia arriba y abajo en el eje de ordenadas respecto a s_{f_k,l_k} respectivamente, es decir:

- $UP_{f_k,l_k} = (m_{f_k,l_k} * x) + c_{f_k,l_k} + p_{f_k,l_k}$.

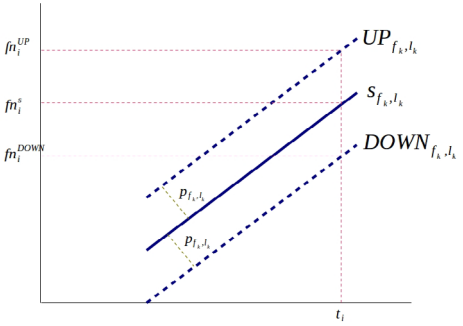

Fig. 1. Método de cálculo de fn_i.

– $DOWN_{f_k,l_k} = (m_{f_k,l_k} * x) + c_{f_k,l_k} - p_{f_k,l_k}$.

Los valores fn_i^a, fn_i^b y fn_i^c que definen fn_i se calculan como el valor de salida de $DOWN_{f_k,l_k}$ ($fn_i^a = DOWN_{f_k,l_k}(t_i)$), s_{f_k,l_k} ($fn_i^b = s_{f_k,l_k}(t_i)$) y UP_{f_k,l_k} ($fn_i^c = UP_{f_k,l_k}(t_i)$) para el valor t_i.

Gracias a este mecanismo el soporte de fn_i está calculado basándose en el ratio de error medio que ofrece el segmento calculado, con lo que la medida de la incertidumbre es apropiada, a más error mayor soporte para el número difuso obtenido.

Esta forma de representación permite realizar comparaciones del número difuso con valores crisp, otros números difusos o cualquier otra representación aprovechando la potencia de las operaciones de la lógica difusa y el adecuado tratamiento de la incertidumbre que ésta ofrece.

3 Pruebas

En esta sección se presenta un ejemplo de cómo nuestra aproximación modela una TS y se detallarán algunas características de los resultados obtenidos. También se plantearán algunas ideas para posibles aplicaciones de la misma.

Como se ha comentado anteriormente se necesita un conjunto S obtenido mediante algún método que genere los segmentos a partir de la TS. En este trabajo hemos utilizado un método propio para obtener el conjunto de segmentos. La Figura 2 muestra la representación gráfica de la TS de entrada y del posicionamiento de los segmentos. Como puede verse, consta de cinco segmentos que visualmente se observan bien ajustados.

S y la TS original se han utilizado como entrada para generar el conjunto de segmentos difusos $FPLS$. La Tabla 1 muestra el conjunto $FPLS$ obtenido. La

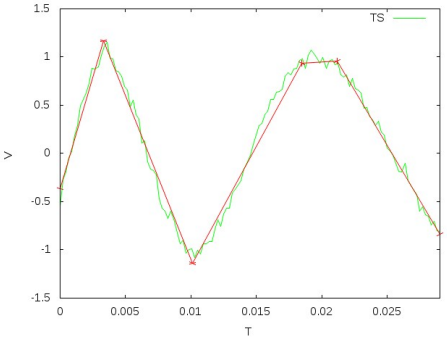

Fig. 2. Conjunto S utilizado como entrada.

primera y segunda columna muestran el identificador del segmento obtenido que indican los instantes iniciales y finales (f_k y l_k), mientras que las tres últimas indican la pendiente, ordenada y la media del ratio de error de cada segmento.

Tabla 1. Conjunto $FPLS$ utilizado en el ejemplo (Figura 2).

$fpls_{f_k,l_k}$	m_{f_k,l_k}	c_{f_k,l_k}	p_{f_k,l_k}
$fpls_{0.0,\ 0.0033}$	458.9375	−0.3616	0.2574
$fpls_{0.0033,\ 0.0101}$	−339.4396	2.3022	0.2144
$fpls_{0.0101,\ 0.0185}$	248.9599	−3.6608	0.2433
$fpls_{0.0185,\ 0.0212}$	6.9336	0.8103	0.0517
$fpls_{0.0212,\ 0.0290}$	−228.6086	5.79461	0.1421

Para comprobar cómo el conjunto de segmentos difusos obtenido representa la salida de la serie se han generado los números difusos de salida para un conjunto de valores de la TS igualmente espaciados (Tabla 2). Las primera y segunda columnas muestran t_i e y_i de cada muestra seleccionada de la TS de entrada. La tercera el número difuso obtenido fn_i. La cuarta muestra la pertenencia obtenida al número difuso fs_i por v_i que permitirá mostrar cómo modela cada fs_i a su valor representado de la TS y_i. Finalmente, la última columna muestra la posición k del $fpls_{f_k,l_k} \in T$ utilizado para obtener el número difuso de salida fs_i.

Tabla 2. Muestra de los números difusos obtenidos.

t_i	y_i	fn_i	$\mu_{fplsf_k,g_k}(y_i)$	k
0.0000	-0.5424	$[-0.619, -0.3616, -0.1042]$	0.2975	
0.0013	0.2897	$[-0.0043, 0.2531, 0.5105]$	0.8577	1
0.0027	0.8796	$[0.6104, 0.8678, 1.1252]$	0.954	
0.004	0.9803	$[0.724, 0.9384, 1.1527]$	0.8047	
0.0054	0.4847	$[0.2694, 0.4837, 0.6981]$	0.9956	
0.0067	-0.0879	$[-0.1853, 0.0291, 0.2435]$	0.4541	2
0.008	-0.5996	$[-0.6399, -0.4255, -0.2111]$	0.1878	
0.0094	-0.8971	$[-1.0945, -0.8801, -0.6657]$	0.9209	
0.0107	-1.0406	$[-1.2366, -0.9933, -0.75]$	0.8055	
0.0121	-0.6873	$[-0.9032, -0.6599, -0.4166]$	0.8872	
0.0134	-0.3716	$[-0.5697, -0.3264, -0.0832]$	0.8143	3
0.0147	0.0908	$[-0.2363, 0.007, 0.2503]$	0.6556	
0.0161	0.5588	$[0.0972, 0.3404, 0.5837]$	0.1024	
0.0174	0.8367	$[0.4306, 0.6739, 0.9172]$	0.3309	
0.0188	0.8866	$[0.8886, 0.9404, 0.9921]$	0.0	4
0.0201	0.9952	$[0.8979, 0.9496, 1.0014]$	0.1196	
0.0214	0.8154	$[0.7535, 0.8957, 1.0378]$	0.4355	
0.0228	0.6605	$[0.4473, 0.5895, 0.7316]$	0.5001	
0.0241	0.2898	$[0.1412, 0.2833, 0.4254]$	0.9545	5
0.0254	-0.017	$[-0.165, -0.0229, 0.1193]$	0.9585	
0.0268	-0.331	$[-0.4712, -0.3291, -0.1869]$	0.9864	
0.0281	-0.6484	$[-0.7774, -0.6353, -0.4931]$	0.9081	

Como puede verse, cada valor de entrada obtiene un número difuso de salida. La $FPLS$ obtenida puede ser utilizada en diferentes aplicaciones. Por ejemplo para realizar comparaciones entre TS. A continuación se describirá una primera aproximación de cómo se puede realizar la comparativa de dos segmentos. La idea para comparar sería similar a los algoritmos propuestos en [17, 18]. La Figura 3 muestra el proceso de comparación de $fpls_{f_a,l_a}$ y $fpls_{f_b,l_b}$ que devolverá un valor de similaridad de los segmentos. Se deben comparar los números difusos de salida para los valores en el intervalo de tiempo que cubren ambos segmentos ($[t_{first}, t_{last}]$) igualmente espaciados (indicado en la figura mediante líneas verticales punteadas). Para realizar la comparación entre los números difusos se puede utilizar cualquiera de los diferentes métodos de comparación propuestos en la literatura. Finalmente, la similaridad será la media de entre las comparaciones realizadas. Actualmente estamos formalizando este método.

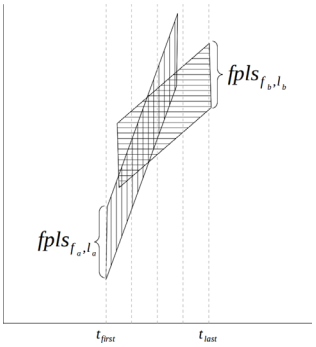

Fig. 3. Comparando dos segmentos difusos.

Además, destacar que nuestra representación de los segmentos puede ser utilizada para representar la TS de una forma similar a lo que Kacprzyk propone.

3.1 Discusión

El trabajo más parecido a nuestra propuesta viene por parte de Kacprzyk y otros [14] que proponen obtener un conjunto de conjuntos difusos que representen el segmento de manera completa con el fin de obtener una visión global del segmento. Posteriormente se utiliza para realizar una descripción de la TS. Kacprzyk y otros presentan los conceptos de dinámica del cambio y duración (medidas mediante conjuntos difusos), y variabilidad del segmento utilizando

medidas estadísticas. Nosotros pretendemos poder representar cada segmento para poder obtener información en cada instante de tiempo. Por ello proponemos el concepto de segmento difuso que devolverá un número difuso ante un valor de entrada.

El uso de números difusos aporta la posibilidad de capturar la incertidumbre de la representación mediante segmentos debida al error y ofrecer una serie de métodos para comparar los segmentos difusos obtenidos con la TS directamente, con segmentos lineales, o con otros segmentos difusos.

Por todo ello, creemos que es una representación con muchas posibilidades y que puede ser utilizada en distintas aplicaciones que necesiten obtener datos de alto nivel a partir de una TS. Por ejemplo, en consultas o descripción de TS.

4 Conclusiones y trabajos futuros

Este trabajo ha presentado una forma de representación de una TS basada en la lógica difusa. Representa la serie como "fuzzy piecewise linear segment" que se basa en una definición de segmento difuso. Cada segmento difuso permite obtener un número difuso de salida ante un instante temporal de entrada. También se esbozan algunas ideas de cómo poder realizar comparación de segmentos difusos.

Esta representación es apropiada para su utilización en distintas aplicaciones que hagan uso de TS como pueden ser descripción y consulta de TS. Se ha presentado brevemente una primera aproximación que queda pendiente como trabajo para futuras investigaciones.

References

1. Romani, L., De Avila, A., Chino, D., Zullo, J., Chbeir, R., Traina, C., Traina, A.: A new time series mining approach applied to multitemporal remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing. 51(1) (2013) 140-150.
2. Wright, M., Stern, P.: Forecasting new product trial with analogous series. Journal of Business Research. 68(8) (2015) 1732-1738.
3. Wei., L-Y.: A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting. Applied Soft Computing, 42 (2016) 368-376.
4. Liu, L., Peng, Y., Wang, S., Liu, M., Huang, Z.: Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors. Information Sciences, 340-341 (2016) 1-17.
5. Sanchez-Valdes, D., Alvarez-Alvarez, A., Trivino, G.: Dynamic linguistic descriptions of time series applied to self-track the physical activity. Fuzzy Sets and Systems, 285 (2016) 162-181.
6. Moreno-Garcia, J., Abián-Vicén, J., Jimenez-Linares, L., Rodriguez-Benitez, L.: Description of multivariate time series by means of trends characterization in the fuzzy domain. Fuzzy Sets and Systems, 285 (2016) 118-139.
7. Keogh, E., Chu, S., Hart, D., Pazzani, M.: An Online Algorithm for Segmenting Time Series. Proc. IEEE Int'l Conf. Data Mining, (2001) 289-296.
8. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting Time Series: A Survey and Novel Approach. Data Mining in Time Series Databases, M. Last, A. Kandel, and H. Bunke, eds., World Scientific Publishing, 57, ch. 1, (2004) 1-22.

9. Huang, H., Matija, M., Suykens, J.A.K.: Hinging Hyperplanes for Time-Series Segmentation. IEEE Trans. Neural Networks and Learning Systems, (2013) 24(8) 1279-1291.
10. Shatkay, H.: Approximate Queries and Representation for Large Data Sequences. Technical Report CS-95-03, Brown University, February 1995.
11. Shatkay, H., Zdonik, S.: Approximate queries and representations for large data sequences. Proc. 12th Int'l Conf. on Data Engineering, (1996) 536-545.
12. Fuchs, E., Gruber, T., Nitschke, J., Sick, B.: Online Segmentation of Time Series Based on Polynomial Least-Squares Approximations. IEEE Pattern Analysis and Machine Intelligence. 32(12) 2232-2245.
13. Garcia-Treviño, E.S., Barria, J.A.: Structural generative descriptions for time series classification. IEEE Transactions on Cybernetics. 44(10) (2014) 1978-1991.
14. Kacprzyk, J., Wilbik, A.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets and Systems, 159 (2008) 1485-1499.
15. Sklansky, J., Gonzalez, V.: Fast polygonal approximation of digitized curves. Pattern Recognition. 12 (1980) 327–331.
16. Hoover, A., Goldbaum, M. : Fuzzy convergence. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (1998) 716-721.
17. Moreno-Garcia, J., Castro-Schez. J.J., Jimenez, L.: A New Method to Compare Dynamical Systems Systems. P. Melin et al. (Eds.), Springer-Verlag Berlin Heidelberg 2007, LNAI 4529. (2007) 533-542, 2007.
18. Moreno-Garcia, J., Castro-Schez. J.J., Jimenez, L.: A New Method to Compare Dynamical Systems Modeled Using Temporal Fuzzy Models. Uncertainty and Intelligent information Systems, 22 (2008), 307-320.