Performance Analysis of Clustering Internal Validation Indexes with Asymmetric Clusters

J. Rojas-Thomas, M. Santos, M. Mora, and N. Duro

***Abstract*—The present work evaluates the performance of a set of internal clustering indexes in artificial and real data sets regarding a specific structural characteristic. In particular, it deals with data sets whose clusters present asymmetric characteristics in their geometries. With this objective, the concept of symmetry is formalized with respect to the axes of maximum variance of the clusters by the definition of a new index. Then a novel methodology is proposed to evaluate the performance of the internal indexes for crisp clustering in sets with this specific structural characteristic. The new defined index is combined with the correlation analysis, allowing to evaluate dynamically the performance of 11 internal indexes well known in the literature, and of a recently proposed index, the Representative Tree Index (RTI). In this way, this methodology allows us to measure not only the absolute error of the indices in relation to a particular configuration of clusters, but also the degree to which the structural characteristic of interest influences the performance of the index, obtaining a more generic understanding of their behaviors.**

***Index Terms*—Clustering, Internal indexes, Performance evaluation, Symmetry.**

I. Introducción

El clustering es una de las técnicas más útiles dentro del área de minería de datos, con vastas aplicaciones, abarcando dominios como la ingeniería, la medicina, negocios y las ciencias sociales [1-5]. El proceso del clustering de datos consiste en clasificar de una manera no supervisada un conjunto de patrones (observaciones o datos) en grupos (clústeres) y, de esta forma, descubrir patrones y correlaciones de interés en la distribución de los datos [6-7]. El objetivo final es que en la partición que se obtiene los datos asignados a un mismo clúster sean similares y los de clústeres diferentes sean muy distintos [8-9].

La aplicación del proceso del clustering de datos consta de cuatro pasos principales: a) selección del espacio de características, b) aplicación del algoritmo de clustering sobre el conjunto de datos obteniendo una partición, c) validación de los resultados y, finalmente d) interpretación de los mismos.

JC Rojas is with the National Distance Education University, 28040-Madrid, Spain (correorojas@gmail.com).

M. Santos is with the University Complutense of Madrid, 28040-Madrid, Spain (msantos@ucm.es).

M. Mora is with Universidad Católica de Maule, Chile (marcomoracofre@gmail.com).

N. Duro is with the National Distance Education University, 28040-Madrid, Spain (nduro@dia.uned.es).

Específicamente, el principal objetivo de la etapa de validación de los resultados es la de encontrar la partición que mejor encaja con la estructura subyacente de los datos, para lo que se utilizan los índices de evaluación internos. Este tipo de índices no requieren información extra acerca de los datos, se basan solamente en el cálculo de propiedades como el nivel de compactación de los clústeres, sus grados de separación y niveles de redondez [10].

Es de gran importancia en problemas reales medir y comparar los rendimientos de los índices internos bajo diferentes escenarios. Para poder realizar este análisis se suelen utilizar conjuntos de datos en los cuales la partición ideal y el número óptimo de clústeres son previamente conocidos. Para ello se utilizan tanto conjuntos de datos artificialmente generados como extraídos de problemas reales de clasificación, generalmente de repositorios públicos.

Una carencia bastante común en los estudios sobre el rendimiento de los índices es que no establecen una medida formal del valor de las características estructurales de los conjuntos de datos utilizados, basándose solamente en una apreciación visual de los mismos, de forma que el análisis de la influencia de estos factores en el rendimiento de los índices internos es poco riguroso. Otra limitación es que la evaluación de los índices se basa en escenarios que representan la característica estructural de interés a partir de la cual se generan rankings absolutos de los índices, obviando el hecho de que estos índices se ven afectados también por otras propiedades estructurales de estos conjuntos no consideradas en el estudio, tales como las posiciones de los clústeres, tamaños, etc., lo cual no permite extrapolar los rendimientos de los índices a otros escenarios. Esto se puede ver en la diversidad de resultados que presentan los estudios de los rendimientos de estos índices.

Por ese motivo en este trabajo se propone una nueva metodología, la cual se basa en la definición de un estadístico para medir la presencia de la característica estructural de interés en los conjuntos de datos. Además permite realizar un análisis dinámico de cómo el grado de presencia de la característica estructural de interés afecta al rendimiento del índice en diferentes casos. Esto, junto con el estudio bidimensional de sus rendimientos, permite agrupar los índices en categorías más flexibles.

En este artículo la aplicación de esta metodología se ha centrado en una propiedad estructural no cubierta por otros trabajos previos, la existencia de asimetrías en el interior de

los mismos clústeres. Características estructurales tradicionalmente consideradas en estos estudios han sido las densidades, el número de clústeres, sus grados de separación, el nivel global de ruido y las diferencias de tamaño, entre otras [11].

Este trabajo se estructura de la siguiente forma: la sección II resume el estado del arte respecto a la evaluación de los rendimientos de los índices internos de clustering. La sección III describe el nuevo índice de simetría definido y la metodología de análisis de rendimientos propuesta. La sección IV detalla las definiciones de los índices internos evaluados en este estudio. La sección V presenta los resultados experimentales y su análisis. Las conclusiones finalizan el articulo.

II. Estado del Arte

Los trabajos que han analizado la problemática de los rendimientos de los índices internos comparan los más utilizados en la literatura sobre clustering [8, 10, 12, 13] y/o proponen un nuevo índice interno [14-17]. Se diferencian principalmente en las características estructurales sobre las cuales los índices son evaluados y el alcance del estudio (número total de índices internos, de conjuntos de datos, etc). Algunos estudios utilizan varios algoritmos de clustering para validar los resultados [18].

Por ejemplo, en [13] se evalúan los rendimientos de 11 índices internos. Las características estructurales de los conjuntos de datos artificiales de dos dimensiones analizadas por este trabajo son: ruido, densidad, sub-clústeres y tamaños. El estudio realiza además un análisis de monotonicidad de los índices y concluye que sólo un índice (S_Dbw) obtiene buen rendimiento en todas las características consideradas, y las que afectaron en mayor medida los rendimientos de los índices fueron la presencia de ruido y de sub-clústeres.

En [8] se evalúa la capacidad de 30 índices internos para reconocer la partición que más se aproxima a la distribución objetivo usando índices externos. Las características estructurales consideradas en los conjuntos de datos artificiales son seis: número mínimo de datos por clúster, número de clústeres, solapamiento entre clústeres, número de dimensiones del espacio de características, densidades y presencia de ruido. No se encuentra evidencia de la superioridad clara de un índice en particular sobre el resto. También se concluye que el nivel de ruido y grado de solapamiento entre clústeres son los factores que más afectaron a los rendimientos de los índices internos.

En [10] se analizan seis índices internos en combinación con 12 diferentes técnicas de clustering, considerando las variaciones de un mismo algoritmo. Las características estructurales de los conjuntos artificiales de prueba destacadas son: el número de clústeres, el número de dimensiones, el tipo de distribución de los datos de los clústeres en cada dimensión (gaussiana, circular), grado relativo de separación de los centros de los clústeres. El estudio concluye que se obtienen resultados satisfactorios en condiciones apropiadamente restringidas, tales como la distribución gaussiana de los datos y en combinación con el uso de algoritmos de clustering orientados a encontrar clústeres compactos. Del total de índices analizados el que obtiene el mejor rendimiento es Silhouette.

En [14] se propone un nuevo índice, I. Las principales características estructurales de los conjuntos de datos fueron la dimensionalidad del espacio de características (hasta diez dimensiones), la superposición parcial entre las clases, y la distribución estadística que siguen los datos de cada clúster. En todas las pruebas el índice propuesto (I) reconoce el número óptimo de clústeres, superando a los otros índices analizados.

En [15] se analiza los rendimientos de 12 índices internos y se compara el rendimiento del índice propuesto (WB) con dos índices similares (CH y Xu). Entre las características estructurales de los conjuntos de datos artificiales destacan la función de distribución utilizada para generar los clústeres, densidades, geometrías y asimetrías, aspectos sobre los que sin embargo el trabajo no ofrece ninguna descripción cualitativa o cuantitativa.

En [16] se compara el índice propuesto (RTI) con siete índices internos. Las características estructurales de los conjuntos de datos artificiales fueron el número de clústeres óptimo, las geometrías de los clústeres (circular, elíptica, arco, anillo), sus grados de dispersión, densidades, el grado de solapamiento entre los clústeres y el nivel de ruido global presente. Las geometrías de los clústeres se representan de manera visual. Para medir los grados de dispersión, densidades, solapamiento y nivel de ruido el trabajo propone una serie de índices basados en un enmallado que se realiza en el espacio de características.

III. Metodología Propuesta

La propuesta de este artículo abarca dos aspectos: primero, la definición formal del concepto de simetría de un clúster respecto a su eje de máxima varianza y, segundo, una nueva metodología de evaluación de los rendimientos de los índices internos de clustering a través de un análisis de correlación.

En este trabajo se considera la simetría de un clúster sobre su eje de máxima varianza como el grado de simetría en la distribución de los datos, tomando como referencia el centroide del clúster, sobre el vector de máxima varianza. Esta medida se obtiene al dividir un clúster por su centroide usando un hiperplano perpendicular al vector de máxima varianza del mismo. Si el clúster es simétrico, el centroide debería estar localizado en el centro de la geometría del clúster en la dimensión determinada por el vector. Por el contrario, si existen diferencias significativas en la distribución de los datos y en las densidades involucradas, el centroide aparece desplazado hacia uno u otro extremo de la geometría.

En la Figura 1 (arriba) se muestra un clúster con un alto grado de asimetría respecto a su eje de varianza máxima. El promedio de las proyecciones de los datos sobre el eje de máxima varianza a partir del centroide se representa como dos rectas que nacen a partir de este último. Se puede observar el alto grado de disimilitud en las longitudes de las dos proyecciones. En la Figura 1 (abajo) se muestra un clúster con un alto grado de simetría, donde las proyecciones tienen

idéntica longitud.

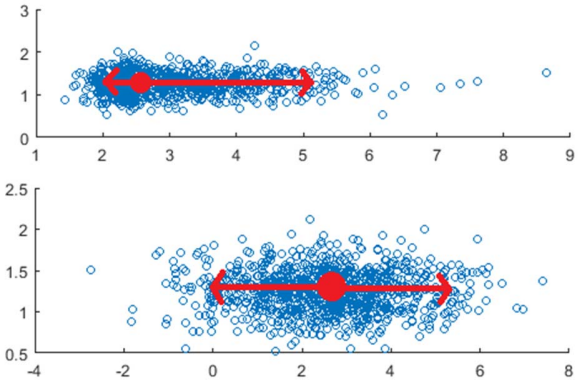

Fig. 1. Clústeres con un alto grado de asimetría (arriba) y simétrico (abajo) respecto a la distribución de los datos en su eje de máxima varianza.

Para obtener una medida del grado de simetría se propone utilizar las proyecciones de los vectores distancia de los datos del clúster respecto del centroide sobre el vector de máxima varianza. Estas proyecciones se promedian separadamente para los dos subconjuntos a ambos lados del hiperplano, constituyéndose en un estimador de la longitud de la geometría del clúster. Luego se calcula la razón entre las longitudes de ambas proyecciones, usando como denominador el mayor valor. Una perfecta simetría daría un valor de 1 e irá progresivamente disminuyendo hasta cero con clústeres de menor simetría.

En la Figura 2, de izquierda a derecha y de arriba a abajo se muestra el proceso de obtención de las longitudes del clúster sobre el vector de máxima varianza. En a) obtención del centroide del clúster (circulo negro), b) obtención del vector de máxima varianza, c) división del clúster a través de un hiperplano perpendicular al vector (línea segmentada), d) obtención de los vectores distancia de los datos al centroide, e) proyecciones de estos vectores distancia sobre el vector de máxima varianza del clúster y, finalmente, f) cálculo de las longitudes promedio de estas proyecciones a ambos lados del hiperplano.

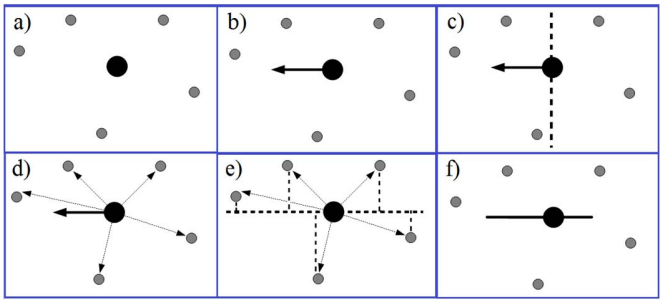

Fig. 2. Calculo de las longitudes del clúster sobre el vector de máxima varianza: a) a f).

A. *Definición Formal del Índice de Simetría*

La metodología usada para crear este índice está parcialmente basada en la propuesta por los mismos autores en [19] para mejorar el índice Davies-Bouldin. Sea D un conjunto de datos particionado en M clústeres. Sea $C_k=\{x_1,x_2,\ldots,x_n\}$ el conjunto de n datos perteneciente al clúster *k-ésimo*. Sea c_k su centroide, y v_k^{max} el vector que captura la mayor varianza del clúster (vector principal). El cálculo del índice requiere los siguientes pasos:

a) Obtener los vectores distancia $D_k=\{d_1,d_2,\ldots,d_n\}$ entre el centroide c_k y el dato *x-iésimo* perteneciente al clúster.

b) Calcular el producto vectorial de las distancias recién calculadas, D_k, y el vector de máxima varianza, v_k^{max}, representado como,

$$p_i = \left| \vec{d}_i \bullet \vec{v}_k^{\max} \right| \tag{1}$$

Donde p_i es la proyección del vector distancia d_i sobre la recta determinada por el vector de máxima varianza v_k^{max}.

c) Agrupar las proyecciones obtenidas en dos conjuntos, $P+$ y $P-$, dependiendo del signo de las proyecciones. Esto implica dividir el clúster por un hiperplano perpendicular al vector de máxima varianza que pasa a través del centroide del clúster.

$$P^+ = \{\, p_i \mid p_i > 0 \,\} \qquad P^- = \{\, p_i \mid p_i < 0 \,\} \tag{2}$$

d) Calcular las cardinalidades de los conjuntos P^+ y P^-.

$$c^+ = \left| P^+ \right| \qquad c^- = \left| P^- \right| \tag{3}$$

e) Obtener el promedio de los valores absolutos de los conjuntos, L^+ y L^-.

$$L^+ = \frac{\sum_{j=1}^{c+} \left| p_j \right|}{c^+} \qquad L^- = \frac{\sum_{j=1}^{c-} \left| p_j \right|}{c-} \tag{4}$$

f) Calcular el valor del índice de simetría para el clúster *k-ésimo* como la división del menor valor por el mayor valor de entre los L^+ y L^- recién calculados.

$$\text{Simetría} = \frac{\min\left[L^-, L^+\right]}{\max\left[L^-, L^+\right]} \tag{5}$$

B. *Metodología*

Se usan dos características principales para evaluar los índices internos sobre los conjuntos de datos de prueba. Para ello en una primera etapa se mide el valor de la característica estructural de interés en cada conjunto de datos. Luego, en una segunda etapa, se generan N particiones de cada uno de los conjuntos de datos aplicando un algoritmo de clustering previamente seleccionado, cubriendo desde un mínimo de dos hasta un máximo de $N+1$ clústeres. Finalmente, se aplican los índices internos sobre las N particiones de cada conjunto de datos, registrando el valor del número de clústeres óptimo calculado para cada uno de ellos.

A partir de estos datos se obtiene la primera variable, que corresponde al error promedio de cada índice respecto al número de clústeres óptimo sobre todos los conjuntos de prueba. Para ello se define una función de error con la cual se evalúan los valores dados por los índices en cada conjunto de datos.

La segunda variable mide la evolución de los rendimientos de los índices internos en relación con las magnitudes de la característica estructural de interés. Esta relación se obtiene calculando la correlación entre la función de error de los índices y el grado de la característica estructural para cada conjunto de datos de prueba.

Por último se realiza un análisis bidimensional de los índices en un espacio de características basado en estas dos variables, donde se agrupan a los índices en categorías predefinidas de acuerdo al comportamiento combinado que presentan respecto a estas dos variables. El esquema global de esta metodología se representa en la Figura 3.

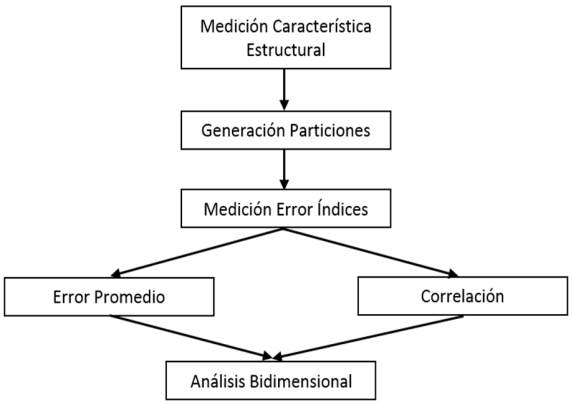

Fig. 3. Representación de la metodología propuesta.

IV. Índices de Validación Internos para el Benchmark

Este trabajo evalúa 11 índices internos bien conocidos en la literatura. En la tabla I se detallan las definiciones de cada uno de ellos respecto a cómo miden el grado de compactación de los clústeres, el grado de separación entre ellos, y la definición del índice que combina ambos factores usando la notación definida en [19]: *D* es el conjunto de datos de entrada, *N* es el total del número de datos en *D*, *g* es la media de todo el conjunto de datos, *P* es el número de dimensiones de *D*, *k* es el número de clústeres, *Ci* es el *i*-ésimo clúster, n_i es el total de datos del clúster C_i , c_i es el centroide del clúster *Ci*, σ(*Ci*) es el vector varianza de *Ci*, $d(x, y)=||x-y||^2$ es la distancia entre *x* e *y*, y n_w corresponde al total de pares de datos dentro de los clústeres, definido como:

$$n_W = \sum_{C_i \in C} \binom{n_i}{2} = \sum_{i=1}^{k} \frac{n_i(n_i-1)}{2} \tag{6}$$

El índice RTI (Representative Tree Index) se basa en la construcción de un árbol de extensión mínimo que captura la geometría de los clústeres, y en la estimación de las densidades de los datos alrededor de los arcos generados. El proceso de subdivisión del clúster se realiza de forma iterativa hasta una profundidad "*t*".

TABLA I
Definiciones de los 11 Índices Internos Utilizados

Ind	Compactación	Separación	Definición
CH	$SSW = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\|x_j - c_i\|^2$	$SSB = \sum_{i=1}^{M} n_i\|c_i - g\|^2$	$CH = \frac{SSB/(k-1)}{SSW/(N-k)}$
I	$E_i = \sum_{i=1}^{k}\sum_{j=1}^{N}\|x_i - c_i\|^2$ $E_1 = \sum_{1}^{N}\|x_i - g\|^2$	$D_i = max_{i,j}\|c_i - c_j\|$	$I(k) = \left(\frac{1}{k}\times\frac{E_1}{E_i}\times D_i\right)^p$
Dunn	$diam(C_i) = max_{x,y\in C_i} d(x,y)$	$dist(C_i, C_j) = min_{x\in C_i, y\in C_j} d(x,y)$	$D = \frac{min\left(dist(C_i,C_j)\right)}{max(diam(C_i))}$
D-B	$\mathrm{disp(C_i)} = \frac{1}{n_i}\sum_{x\in C_i} \mathrm{d(x, c_i)}$	$\mathrm{dist(C_i, C_j)} = \mathrm{d(c_i, c_j)}$	$\mathrm{DB} = \frac{1}{k}\sum_{i} \max_{j\neq i}\frac{\mathrm{disp(C_i)} + \mathrm{disp(C_j)}}{\mathrm{dist(C_i, C_j)}}$
SD	$\mathrm{Scat(k)} = \frac{1}{k}\sum\frac{\|\sigma(c_i)\|}{\|\sigma(D)\|}$	$\mathrm{Dis(k)} = \frac{\max d(c_i,c_j)}{\min d(c_i,c_j)}\sum(d(c_i,c_j))^{-1}$	$SD(k) = Dis(k_{max})\cdot Scat(k) + Dis(k)$
XB	$Comp(C_i) = \sum_{x\in C_i} d^2(x, c_i)$	$dist(C_i, C_j) = d^2(c_i, c_j)$	$XB = \frac{\sum_{i=1}^{k} Comp(C_i)}{n\cdot \min_{i,j}\cdot dist(C_i,C_j)}$
Sil	$a(x) = \frac{\sum_{y\in C_i, x\neq y} d(x,y)}{n_i - 1}$	$b(x) = \min_{j\neq i}\frac{1}{nj}\sum_{y\in Cj} d(x,y)$	$\mathrm{S} = \frac{1}{N}\left(\sum_{C_i\in C}\left(\sum_{x\in C_i}\frac{\mathrm{b(x)} - \mathrm{a(x)}}{\max[\mathrm{b(x)}, \mathrm{a(x)}]}\right)\right)$
C	$\mathrm{S(C)} = \sum_{C_i\in C}\sum_{x_i x_j\in C_i} \mathrm{d_e(x_i, x_j)}$	$\mathrm{S_{min}(C)} = \sum \min(n_W)_{x_i x_j \in D}\{\mathrm{d_e(x_i, x_j)}\}$ $\mathrm{S_{max}(C)} = \sum \max(n_W)_{x_i x_j \in D}\{\mathrm{d_e(x_i, x_j)}\}$	$\mathrm{CI(C)} = \frac{\mathrm{S(C)} - \mathrm{S_{min}(C)}}{\mathrm{S_{max}(C)} - \mathrm{S_{min}(C)}}$
B-HG	S-: nº veces dist. par de datos ∈ mismo clúster > dist. par de datos ∈ distinto clúster	S+: nº veces distancia par de datos ∈ mismo clúster < distancia par de datos ∈ distinto clúster	$\mathrm{G} = \frac{(\mathrm{S}+) - (\mathrm{S}-)}{(\mathrm{S}+) + (\mathrm{S}-)}$
G+			$\mathrm{G} = \frac{2(\mathrm{S}-)}{\mathrm{n_w}(\mathrm{n_w} - 1)}$
CS	$\mathrm{diam(C_k)} = \frac{1}{\|C_k\|}\cdot \sum_{x_j\in C_k}\max_{x_j\in C_k}\{\mathrm{d_e(x_i, x_j)}\}$	$\mathrm{separacion(C_k)} = \min_{C_l\in C, l\neq i}\{\mathrm{d}(\overline{C_k}, \overline{C_l})\}$	$\mathrm{CS(C)} = \frac{\sum_{C_k\in C}\mathrm{diam(C_k)}}{\sum_{C_k\in C}\mathrm{separacion(C_k)}}$

Sea $RT_k=(SC_k,E_k)$ el árbol representativo del clúster C_k, donde $SC_k=\{sc_1,sc_2,\ldots,sc_m\}$ son los sub-centroides generados a partir del mismo clúster y $E_k=\{e_1,e_2,\ldots,e_{m-1}\}$ el conjunto de arcos del árbol. *Edge_Dispersion*(e_i) es la función que calcula el grado de dispersión de los datos del *i-ésimo* arco del mismo árbol.

El grado de compactación de un clúster se define entonces como el arco del árbol representativo que presenta el mayor nivel de dispersión, y la distancia entre dos clústeres como el nivel de dispersión de los sub-centroides más cercanos entre los respectivos árboles representativos. Para el cálculo de la separación entre clústeres se definen $RT_g=(SC_g,E_g)$ y $RT_h=(SC_h,E_h)$ como los respectivos árboles representativos de los clústeres C_g and C_h, y sean sc_g y sc_h los sub-centroides más cercanos entre ambos clústeres de acuerdo a la distancia euclidiana. La función *Edge_Dispersion*(sc_g, sc_h) calcula el grado de dispersión en el arco que une a ambos sub-centroides (Tabla II).

La definición del índice RTI es la siguiente:

$$RTI = \frac{\sum_{i=1}^{k} FACTOR(C_i)}{k} \tag{7}$$

Donde *Factor(C_i)* corresponde a la contribución del clúster *i*-ésimo al valor del índice, dado por el termino:

$$\frac{\text{CLUSTER\_DISCOHESION}(C_i)}{\text{MIN}\left(\text{INTERCLUSTER\_DISTANCE}(C_i,C_j)\right)}, \forall\, C_j \in P,\ i \neq j \tag{8}$$

TABLA II
COMPACTACIÓN Y DISTANCIAS DE LOS CLÚSTERES EN EL ÍNDICE RTI

Comp	$CLUSTER\_DISCOHESION(C_k) = \max(EDGE\_DISPERSION(e_i)), \forall e_i \in E_k$
Sep	$INTERCLUST\ ER\_DISTANCE\ (C_g, C_h) = EDGE\_DISPERSION\ (sc_g, sc_h)$ $sc_g \in SC_g \wedge sc_h \in SC_h \wedge d(sc_g, sc_h) \leq d(sc_i, sc_j), \forall sc_i \in SC_g, \forall sc_j \in SC_h$

Existen dos versiones del índice RTI, dependiendo de la implementación de la función que calcula el grado de dispersión de un arco (Tabla III). La primera versión, denominada RTI^{avg}, lo calcula como la distancia promedio al punto medio del arco de todos los datos que pertenecen a los dos sub-clústeres conectados por éste. La segunda versión, RTI^{min}, calcula el nivel de dispersión como la menor distancia de los mismos datos al punto medio del arco.

Formalmente, sean $C^{sub}{}_g$ y $C^{sub}{}_h$ dos sub-clústeres cuyos respectivos sub-centroides, sc_g y sc_h, están unidos por un arco, y *n* y *m* la cardinalidad de los respectivos sub-clústeres. El punto medio del arco se define como:

$$MP = \frac{(sc_g + sc_h)}{2} \tag{9}$$

TABLA III
DISPERSIÓN DE UN ARCO PARA LAS DOS VERSIONES DEL ÍNDICE RTI

Version RTI	EDGE_DISPERSION
RTImin	$MIN(d(p_i, MP)), \forall p_i \in (C^{sub}{}_g \cup C^{sub}{}_h)$
RTIavg	$\frac{\sum_{i=1}^{n+m} d(p_i, MP)}{n+m}, \forall p_i \in (C^{sub}{}_g \cup C^{sub}{}_h)$

La versión del índice que arroja los mejores resultados es la RTI^{min} con valores del parámetro *t* iguales a uno y dos [19], que será con la que se trabajará en este estudio.

V. RESULTADOS EXPERIMENTALES Y DISCUSIÓN

A. Experimentos con Datos Artificiales

Se generaron ocho conjuntos de datos de prueba, compuestos de cuatro clústeres con 70 datos cada uno (Figura 4). Estos conjuntos presentan diversos grados de simetría global, la cual se define como el promedio de las simetrías de los clústeres individuales:

$$Simetria\_global(D) = \frac{1}{k}\sum_{i=1}^{k} Simetria(C_i) \tag{10}$$

Para la creación de los clústeres de estos conjuntos se utilizó un generador pseudo-aleatorio con una distribución normal estándar. Este generador de números aleatorios permite configurar los parámetros de los dos ejes de coordenadas de la distribución normal. Así, cada uno de los clústeres se crea centrado en el origen del sistema de coordenadas (0,0) y con el eje "x" como el eje de mayor varianza del clúster. Luego se desplaza a la posición deseada en el espacio de coordenadas y se rota para darle su orientación final. Inicialmente en el eje "y" todos los valores de los datos de un clúster cualquiera tienen la misma media (0) y desviación estándar (0.25). Para los valores de la coordenada "x", la desviación estándar de los valores positivos es diferente de la de los valores negativos generados. La mayor o menor diferencia entre ambas desviaciones determina el mayor o menor grado de simetría del clúster. Adicionalmente también se incrementa la cardinalidad de los datos de la mitad del clúster con menor desviación estándar (valores negativos) para disminuir aún más la simetría.

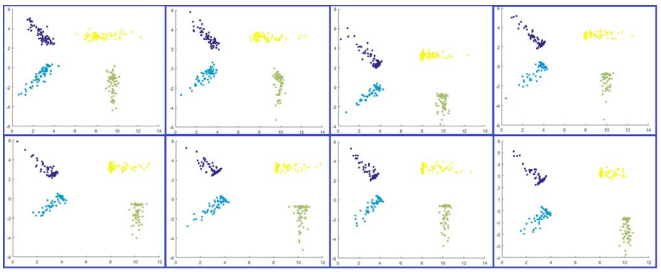

Fig. 4. Conjuntos de prueba ordenados de mayor a menor grado de simetría (arriba izquierda a inferior derecha).

Se generan $N = 9$ particiones con el algoritmo de clustering *k-means* para cada conjunto, abarcando un rango de dos a diez clústeres.

La medida de error utilizada es el error absoluto entre el número objetivo de clústeres buscado (cuatro en todos los casos) y el valor obtenido por el índice interno:

$$error = |numero\_objetivo - prediccion\_indice| \tag{11}$$

La tabla IV muestra los resultados de los índices internos (filas) sobre cada uno de los conjuntos de prueba (columnas, del uno al ocho) en términos del error absoluto.

La segunda fila, "Simetría", muestra el valor de simetría global de cada conjunto. La última fila, "Error", muestra el promedio de la función de error de todos los índices por cada conjunto, mientras que la última columna de la derecha, "Error", muestra el promedio de la función de error en todos los conjuntos de cada uno de los índices.

Para el análisis de correlación se utiliza el coeficiente de Pearson, abarcando valores entre -1 a +1, donde +1 significa correlación lineal positiva total, y -1 una correlación negativa total. Se considera que valores > |0.8| implican una fuerte correlación y valores < |0.5| una débil correlación.

El análisis de correlación se dividió en dos fases: en la primera etapa se realizó un análisis general del efecto de la simetría sobre el rendimiento de los índices internos.

TABLA IV
RESULTADOS DE LOS ÍNDICES PARA CADA CONJUNTO DE DATOS DE PRUEBA

Conjunto	1	2	3	4	5	6	**7**	8	**Error**
Simetría	,7221	,6419	,6130	,5950	,5313	,5033	,4904	,4667	
SD	0,00	0,00	0,00	1,00	1,00	0,00	2,00	0,00	0,50
XB	0,00	0,00	0,00	1,00	1,00	0,00	4,00	0,00	0,75
Silo	0,00	0,00	0,00	2,00	1,00	0,00	2,00	0,00	0,63
I	0,00	4,00	2,00	4,00	6,00	6,00	4,00	4,00	3,75
Dunn	0,00	2,00	2,00	2,00	2,00	2,00	2,00	2,00	1,75
CH	6,00	4,00	6,00	4,00	6,00	6,00	4,00	5,00	5,13
D-B	0,00	0,00	1,00	1,00	1,00	0,00	2,00	0,00	0,63
B-H	0,00	0,00	3,00	4,00	3,00	6,00	4,00	6,00	3,25
C	0,00	4,00	3,00	6,00	6,00	6,00	4,00	6,00	4,38
G+	0,00	0,00	0,00	1,00	1,00	0,00	4,00	0,00	0,75
CS	0,00	0,00	2,00	2,00	1,00	6,00	5,00	4,00	2,50
$RTI_{min}1$	0,00	0,00	0,00	1,00	0,00	0,00	0,00	1,00	0,25
$RTI_{min}2$	0,00	0,00	0,00	1,00	1,00	0,00	0,00	1,00	0,38
Error	0,46	1,08	1,46	2,31	2,31	2,46	2,85	2,23	1,89

Para esto se calculó la correlación lineal entre los valores de error promedio de todos los índices para cada conjunto (última fila) vs el valor de simetría de cada uno de estos conjuntos (fila "Simetría"), dando un valor de -0.9. Con esto se comprueba que existe una muy alta correlación entre el rendimiento promedio de los índices internos y la simetría de los conjuntos. Al ser negativa implica que a medida que el índice de simetría disminuye, el error absoluto promedio crece (Figura 5).

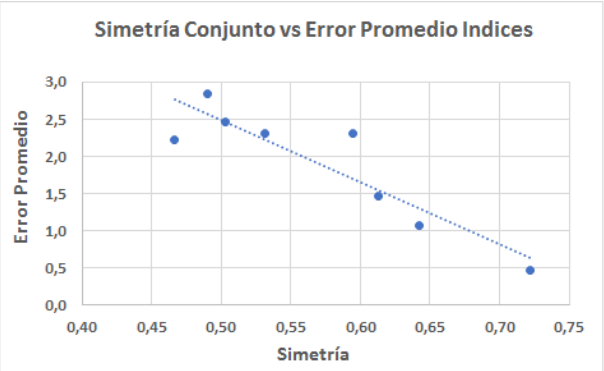

Fig. 5. Correlación entre la simetría de los 8 conjuntos de prueba (eje horizontal) y el error absoluto promedio de todos los índices internos para cada conjunto.

En una segunda etapa se realizó un análisis individual para cada índice del efecto de la simetría de los conjuntos sobre sus rendimientos particulares. Para esto se calculó la correlación lineal entre los valores de error individuales de los índices por cada conjunto (filas 3-15) vs el valor de simetría de cada uno de estos conjuntos (fila "Simetría"). La Figura 6 muestra estos valores ordenados de mayor a menor (positivo a negativo).

Como se puede observar, la mayoría de los índices dan valores de correlación negativos, a excepción del índice CH, cubriendo un amplio rango de valores.

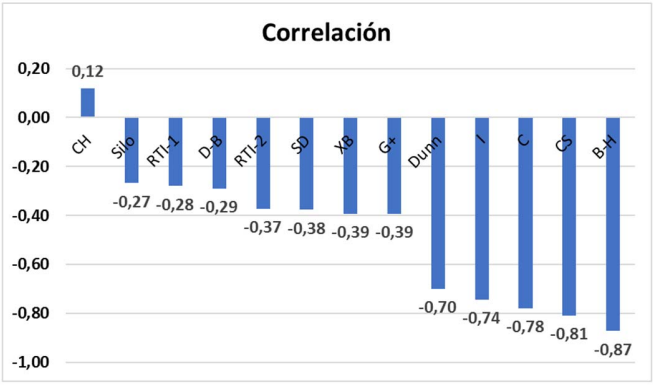

Fig. 6. Resultados de los valores de correlación para cada índice interno, de mayor a menor considerando el signo.

Finalmente se realiza al análisis bidimensional utilizando los valores de correlación recién calculados y los errores absolutos promedios de cada índice. En la Figura 7 se puede observar que la relación entre tasa de error y correlación no es directa, y que los índices se agrupan en tres clústeres bien definidos.

Considerando una baja tasa de error si es ≤ 1, y baja correlación si es < |0.5|, los índices se agrupan en las siguientes categorías:

a) Baja correlación – Bajo error promedio: el rendimiento ideal. Los rendimientos están muy poco afectados por las características de simetría de los conjuntos, y son capaces de detectar el numero correcto de clústeres con alto grado de certidumbre. En esta categoría quedan clasificados los índices internos SD, XB, G+, Davies-Bouldin, Silohuette y RTI, donde el índice $RTI_{min}1$ muestra el mejor rendimiento en general respecto a ambas variables.
b) Baja correlación – Alto error promedio: índices cuyo rendimiento no se ve muy afectado por los valores de simetría, pero sí por otros factores como, por ejemplo, la distribución espacial de los clústeres, o la compatibilidad del índice con el algoritmo de clustering utilizado. En esta categoría solamente queda clasificado el índice CH.
c) Alta correlación – Alto error promedio: índices con los peores rendimientos. La alta correlación indica rendimientos muy dependientes de los valores de simetría de los conjuntos, y el alto error promedio indica que sus predicciones se alejan en gran medida del número correcto de clústeres con los conjuntos de prueba utilizados. En esta categoría quedan clasificados los índices internos B-H, CS, C, I y Dunn, donde este ultimo muestra el mejor rendimiento del grupo.

B. Experimentos con Datos Reales

Los conjuntos de datos reales pueden presentar diversas características estructurales combinadas, tales como ruido, solapamiento, etc., y cada una de ellas va a afectar al rendimiento de los índices en diferente grado. Por lo tanto, no se podrían extraer conclusiones del funcionamiento de un índice respecto a una característica aislada, como se ha hecho con los conjuntos de datos artificiales generados "ex profeso" respecto a la simetría, donde se controla que esta característica sea la dominante.

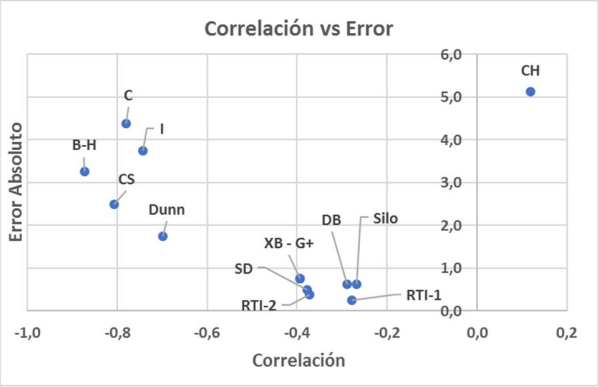

Fig. 7. Posición de los índices evaluados en el espacio de coordenadas determinado por el promedio de los errores absolutos de cada índice (eje vertical) y el valor de correlación entre los valores de simetría global de cada conjunto de datos y los errores absolutos del índice en cada uno de estos (eje horizontal).

Sin embargo, sí se puede hacer un análisis global del rendimiento de los índices para encontrar una tendencia, si existe, utilizando la correlación entre el grado de simetría presente en los datos de casos reales y el rendimiento promedio de todos los índices combinados. Para hacer este análisis se han utilizado 14 conjuntos de datos reales extraídos del repositorio público "UCI Machine Learning Repository" [20]. En concreto, estos conjuntos son: Iris (3 clases, 4 atributos, 150 muestras), Breast Cancer Wisconsin (Diagnostic) (2 clases, 30 atributos, 569 muestras), Wine (3 clases, 13 atributos, 178 muestras), Ecoli (8 clases, 7 atributos, 336 muestras), Haberman's Survival (2 clases, 3 atributos, 306 muestras), Breast Tissue (6 clases, 9 atributos, 106 muestras), Seeds (3 clases, 7 atributos, 210 muestras), Spectf Heart (2 clases, 44 atributos, 80 muestras), Steel Plates (7 clases, 27 atributos, 1941 muestras), Connectionist Bench (Sonar, Mines vs. Rocks) (2 clases, 60 atributos, 208 muestras), Fertility (2 clases, 10 atributos, 100 muestras), Parkinson (2 clases, 23 atributos, 197 muestras), Statlog (Vehicle Silhouettes) (4 clases, 18 atributos, 846 muestras), y finalmente Yeast (10 clases, 8 atributos, 1484 muestras).

La Figura 8 muestra los valores de simetría de cada uno de estos conjuntos. Se puede ver que los valores van desde muy poca simetría (Steel, 0.58) hasta conjuntos con un alto grado de esta característica (mayor que 0.9).

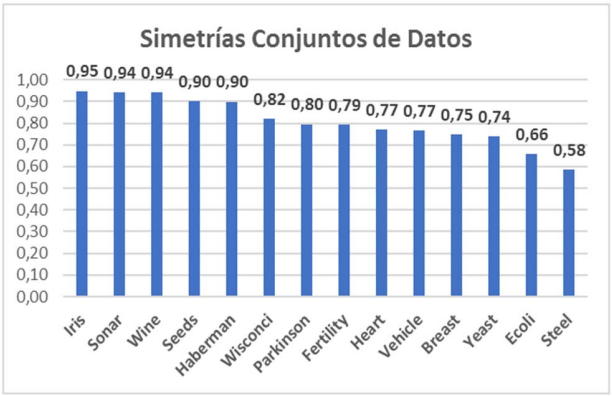

Fig. 8. Valores de simetría de los conjuntos de datos reales, de mayor a menor.

En la Figura 9 se representa el error promedio de todos los índices para cada conjunto de datos. La Figura 10 presenta la relación entre el error promedio de todos los índices y el valor de simetría medido. El valor de correlación obtenido entre ambas medidas fue de 0.79, lo que confirma que existe una correlación significativa entre el nivel de simetría de los conjuntos de datos reales analizados y el rendimiento de los índices.

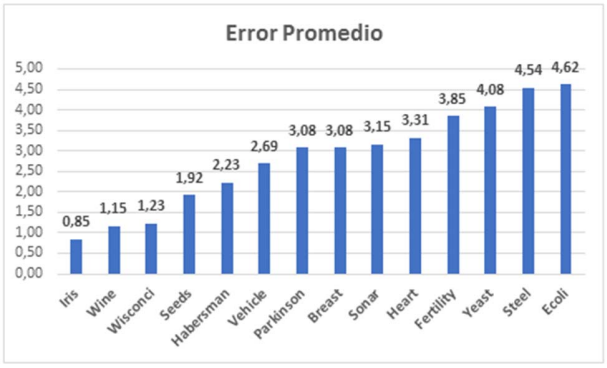

Fig. 9. Error promedio de los índices para cada conjunto de datos reales, de menor a mayor.

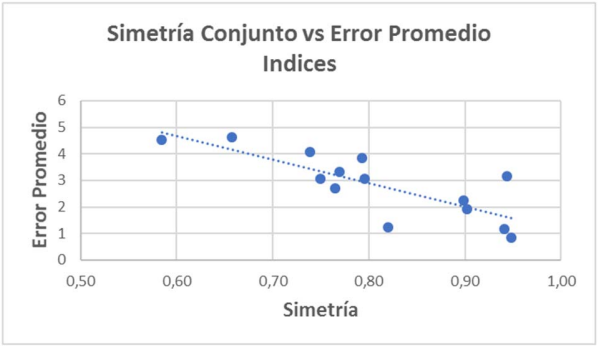

Fig. 10. Grado de simetría de los conjuntos de datos reales frente al error promedio de los índices internos. La recta de puntos representa la tendencia de los datos.

La propuesta y los experimentos se implementaron con el software científico Matlab R 2016a sobre el sistema operativo Windows 8 de 64 bits. El hardware utilizado es un laptop HP modelo Envy m6 con un procesador Intel de 2 núcleos y 2,5 GHz. La memoria RAM del equipo es de 8 GB.

VI. Conclusiones y Trabajos Futuros

En este trabajo se ha descrito formalmente el concepto de simetría en la geometría de los clústeres respecto de su eje de máxima varianza, y se ha propuesto una nueva metodología para evaluar los rendimientos de los índices internos respecto a la presencia y grado de esta característica estructural en los conjuntos de datos.

Con esta metodología se han evaluado 12 índices internos sobre conjunto artificiales y sobre conjuntos reales, midiéndose el grado en que la simetría de los clústeres influye en sus rendimientos. Se han catalogado los índices en tres grupos de acuerdo al rendimiento mostrado en un espacio de dos variables que considera tanto sus rendimientos relativos como absolutos.

Como trabajos futuros se proponen formalizar otras características estructurales de los datos a través de la definición de nuevos índices, así como explorar y analizar los rendimientos de los índices internos sobre estas nuevas características. También se propone aplicar esta metodología de evaluación de rendimiento de índices internos usando otros algoritmos de clustering [21].

REFERENCIAS

[1] M. Halkidi, Y. Batistakis, M. Vazirgiannis. "On clustering validation techniques". *J Intell Inf Syst*, vol. 17, no. 2-3, pp. 107-145, 2001.

[2] A.J.O. Reyes, A.O. Garcia, Y.L. Mué. "System for processing and analysis of information using clustering technique". *IEEE Lat Am T*, vol. 12, no. 2, pp. 364-371, 2014.

[3] M.C. Sergio, J.A. de Souza, A.L. Goncalves. "Idea Identification Model to Support Decision Making". *IEEE Lat Am T*, vol. 15, no. 5, pp. 968-973, 2017.

[4] T.R. Botelho, D. Soprani, C. Rodrigues, A. Ferreira, A. Frizera. "New approach to the EEG signals classification using the variance of the difference between the classes of a Bayesian classifier". *Rev Iberoam Autom In,* vol. 14, no. 4, pp. 362-371, 2017.

[5] F.A.R. Silva. "Analytical Intelligence in Processes: Data Science for Business". *IEEE Lat Am T*, vol. 16, no. 8, pp. 2240-2247, 2018.

[6] A. K. Jain. "Data clustering: 50 years beyond K-means". *Pattern Recogn Lett*, vol. 31, no. 8, pp. 651-666, 2010.

[7] E. Hancer, D. Karaboga. "A compehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number". *Swarm Evolutionary Comp*, vol. 32, pp. 49-67, 2017.

[8] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona. "An extensive comparative study of cluster validity indices". *Pattern Recogn*, vol. 46, no. 1, pp. 243-256, 2013.

[9] J. Cervantes, J. Taltempa, F.G. Lamont, J.S.R. Castilla, A.Y. Rendon, L.D. Jalili. "Análisis comparativo de las técnicas utilizadas en un sistema de reconocimiento de hojas de planta". *Rev Iberoam Autom In,* vol. 14, no. 1, pp. 104-114, 2017.

[10] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, E.R., Dougherty. "Model-based evaluation of clustering validation measures". *Pattern Recogn*, vol. 40 no. 3, pp. 807-824, 2007.

[11] K.R. Žalik, B. Žalik. "Validity index for clusters of different sizes and densities". *Pattern Recogn Lett*, vol. 32, no. 2, pp. 221-234, 2011.

[12] L.J. Deborah, R. Baskaran, A. Kannan. "A survey on internal validity measure for cluster validation". *Int J Comp Sci Eng.* vol. 1, no. 2, pp. 85-102, 2010.

[13] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu. "Understanding of internal clustering validation measures". In *Data Mining* (ICDM), *2010 IEEE 10th Int Conf* on (pp. 911-916). IEEE, 2010.

[14] U. Maulik, S. Bandyopadhyay. "Performance evaluation of some clustering algorithms and validity indices". *IEEE Trans Pattern Anal Mach Intell*, vol. 24, no. 12, pp. 1650-1654, 2002.

[15] Q. Zhao, P. Fränti. "WB-index: A sum-of-squares based index for cluster validity". *Data Knowl Eng.* vol. 92, pp. 77-89, 2014.

[16] J.C. Rojas-Thomas, M. Santos, M. Mora. "New internal index for clustering validation based on graphs". *Expert Syst Appl.* vol. 86, pp. 334-349, 2017.

[17] Z. Wang, Z. Yu, C.P. Chen, J. You, T. Gu, H.S. Wong, J. Zhang. Clustering by local gravitation. *IEEE Trans Cyber*. vol. 48, no. 5, pp. 1383-1396, 2018.

[18] A. Starczewski. "A new validity index for crisp clusters". *Pattern Anal Appl*, vol. 20, no. 3, pp. 687-700, 2017.

[19] J.C. Rojas-Thomas, M. Santos, M. Mora. "New Version of Davies-Bouldin index for clustering validation based on hyper rectangles". In *6th Chilean Conference on Pattern Recognition*, pp. 1-6, 2014.

[20] Asuncion, Arthur, and David Newman. "UCI machine learning repository." (2007). Asuncion, A., & Newman, D. (2007). [https://archive.ics.uci.edu/ml/index.php]

[21] K. Gajowniczek, T. Ząbkowski. "Simulation Study on Clustering Approaches for Short-Term Electricity Forecasting". *Complexity*, vol 2018, Art ID 3683969, 21 pp, 2018.

Juan Carlos Rojas-Thomas nació en Concepción, Chile. Recibió su grado de Ingeniero Civil en Informática en 2003, y el de Máster en Ciencias de la Computación en 2009, ambos por la Universidad de Concepción, Chile. Actualmente está cursando el Programa de Doctorado en Ingeniería de Sistemas y Control, en la Universidad Nacional de Educación a Distancia, España.

Ha publicado artículos y ha sido revisor en diversas revistas y conferencias. Sus intereses de investigación son el Clustering de Datos, Redes Neuronales y el Procesamiento Digital de Imágenes.

Prof. Dr. Matilde Santos Peñas was born in Madrid, Spain. She received her B.Sc., M.Sc. degrees and her Ph.D in Physics from the University Complutense of Madrid (UCM). She is with the Department of Computer Architecture and Systems Engineering at the UCM, where she is currently Full Professor in System Engineering and Automatic Control.

She has published many papers in international scientific journals and several book chapters. She has supervised 10 Ph.Ds. She has worked on several national, European and international research projects, leading some of them.

Her major research interests are: Artificial Intelligence (mainly in the automatic control field), Pattern recognition, Modelling and simulation, Engineering applications of Soft Computing techniques, Renewable energy.

Dr. Marco Mora es profesor del Departamento de Computación de la Universidad Católica del Maule, Talca, Chile. Es director e investigador senior del Laboratorio de Investigaciones Tecnológicas en Reconocimiento de Patrones LITRP (www.litrp.cl). Es Ingeniero Electrónico y Magister en Eléctrica de la Universidad de Concepción, Chile. Es Doctor en Ciencias de la Computación por la Université de Toulouse, Francia.

Sus intereses de investigación son Procesamiento Digital de Imágenes, Redes Neuronales, Biometría y Aplicaciones Industriales de Reconocimiento de Patrones.

Dr. Natividad Duro es licenciada en Ciencias Físicas por la Universidad Complutense de Madrid y doctora en Ciencias por la UNED desde 2002. Es profesora Titular de Universidad en el Dpto. de Informática y Automática de la UNED.

Sus áreas de investigación principales se centran en el control y modelado de procesos, diseño de nuevos sistemas orientados a la educación, minería de datos, procesamiento de señales y big data.