ESTUDIO DE INUNDACIONES COSTERAS APLICANDO EL TRATAMIENTO DE DATOS SEGÚN LA TEORÍA DE LOS CONJUNTOS APROXIMADOS

Gisela del Valle Rodríguez [1], María Josefina Codorníu Pujals [1], Águeda L. García Martín [1], Ida Mitrani Arenal [2]

[1]Departamento de Meteorología. Instituto Superior de Tecnologías y Ciencias Aplicadas, Universidad de La Habana, Cuba. gisela@instec.cu
[2] Centro de Física de la Atmósfera. Instituto de Meteorología, Cuba. ida.mitrani@insmet.cu

RESUMEN

El desarrollo tecnológico ha permitido, especialmente en el ámbito científico del Tiempo y el Clima, disponer de bases de datos de volumen y dimensionalidad creciente. Estas bases de datos están compuestas por numerosas variables meteorológicas y/o climatológicas provenientes de fuentes heterogéneas (instrumentos meteorológicos, modelos numéricos, percepciones visuales de los observadores meteorológicos, testimonios de pobladores, etc.) con tipologías de datos de diversa naturaleza (cuantitativos, cualitativos y/o semánticos con caracteres numéricos, ordinales, nominales y lingüísticos) que generan nuevos retos para el adecuado tratamiento. La Teoría de los Conjuntos Aproximados permite tratar los datos originales sin transformación previa e identificar los vínculos existentes entre las variables en forma de reglas de asociación del tipo IF-THEN. En este trabajo se muestran los resultados de aplicar dicha teoría a una base de datos de inundaciones costeras provocadas por sistemas frontales que ocurrieron en el litoral habanero en el período 1980-2010, utilizando el modelo clásico de Pawlak y algunas de sus extensiones, así como sus correspondientes algoritmos proporcionados por la herramienta computacional ROSE2. Para diferentes casos de estudio de tipo exploratorio, se distinguen las variables indispensables, se reduce la dimensionalidad y se comparan las reglas extraídas, todo lo cual constituye una peculiar representación del conocimiento en el campo de las inundaciones costeras. Finalmente, se muestran los resultados de un estudio de valor predictivo, y se fundamenta la contribución del sistema de reglas para un sistema de alerta temprana para inundaciones costeras en el litoral habanero. **Palabras clave:** Teoría de los Conjuntos Aproximados, descubrimiento del conocimiento en bases de datos, inundaciones costeras

ABSTRACT

Developments in technology has enabled accumulation of large databases and most of the meteorological and climatological databases consist of large quantity of real time observatory datasets of high dimension space. These databases are constructed with variables obtained from heterogeneous sources (meteorological instruments, numerical model outputs, visual perceptions and other non-authoritative data) which should be represented in different ways (quantitative, qualitative and/or semantic) with continuous, numerical, ordinal, nominal and/or linguistic characters. It has been an important issue to reduce huge objects and large dimensionality in databases. Attribute reduction, also called feature selection, finds a subset of attributes to reduce dimensionality. The application of the Rough Set Theory in complex databases has its own advantages; the rough set based attribute reduction finds particular subsets of conditional attributes providing the same information for classification purpose with the original set. Moreover, Rough Set Theory does not need any preliminary or additional information about data, and permits to discover interesting and useful patterns among these stored data in the databases as rules or logical expressions in the following form: IF (conditions) THEN (decision class). This article reports two

examples – an exploratory study and a predictive study - that applies the Rough Set Theory to a database of coastal flooding happened in Havana's shoreline from 1980 through 2010 using the computational tool ROSE2. The rough set analysis supplies some useful elements of knowledge discovery: the indispensable variables, the relevance of attributes and/or criteria, the quality of lower and upper approximations, the minimal subsets of attributes or criteria (reducts), the set of the non-reducible attributes or criteria (core) and the induced association/decision rules. The system of meaningfully rules obtained from a predictive study should be important to implement monitoring and early warning system. **Key Words:** Rough Sets Theory, Knowledge Discovery in Databases, coastal flooding.

Introducción

Las investigaciones sobre las inundaciones costeras en Cuba, iniciadas en la última década del siglo XX, se han visto incrementadas en las dos primeras décadas de este siglo. En ellas se han estudiado las inundaciones costeras en relación a los eventos hidrometeorológicos que las originan, los diferentes tramos costeros del país, así como sus tendencias climáticas (Hidalgo, 2016). En un detallado análisis bibliográfico, Hidalgo identifica la amplia diversidad de variables utilizadas y la falta de homogeneidad que, desde el punto de vista metodológico, impiden realizar una adecuada generalización de criterios a partir de los resultados de los diferentes autores, aun cuando en todos los estudios se utiliza un *enfoque estadístico* efectuándose así una *descripción precisa* de los datos.

Una de las vías que permite la *descripción no precisa* de los datos, se logra mediante la aplicación de la Teoría de los Conjuntos Aproximados (en inglés, *RST-Rough Set Theory*) creada por Pawlak en 1981 (Pawlak, 1981a) (Pawlak, 1981b) en la que se concibe la *indiscernibilidad* (Zhao *et al.*, 2007) entre objetos matemáticos describiendo la incertidumbre del *conocimiento aproximado* como *vaguedad* a través de una *relación de equivalencia* (Yao, 2015a) (Li & Hong, 2017). La posibilidad de tratar los datos sin requerir de una transformación o preparación inicial de los mismos hacen atrayente esta teoría que, respetando el contenido semántico de los valores de las variables, logra encontrar relaciones o asociaciones entre las mismas y que permite dar solución, desde otra perspectiva, al problema de la reducción de la dimensionalidad (JeraldBeno & Karnan, 2012). Las autoras de este trabajo realizan el análisis de las inundaciones costeras provocadas por sistemas frontales que ocurrieron en el litoral habanero en el período 1980-2010 a través de la aplicación de la Teoría de los Conjuntos Aproximados. Los objetivos de este trabajo radican en mostrar los pasos requeridos para la aplicación de esta teoría y la forma en que se interpretan los resultados ilustrando así el *enfoque no estadístico* del estudio.

Materiales y métodos

Ideas básicas de la Teoría de los Conjuntos Aproximados

Existe una excelente, actualizada y accesible bibliografía que presenta rigurosamente las bases teóricas matemáticas de la Teoría de los Conjuntos Aproximados (Pawlak & Skowron, 2007) (Zhao *et al.*, 2007) (Yao, 2015a). Fundamentada en la *discernibilidad* y *aproximación* (Pawlak, 1994) opera con cierta metodología propia. Se ha considerado pertinente realizar los siguientes comentarios:

¿Qué es un *rough set*?

- Los objetos a describir pueden no ser *discernible*s en términos de descriptores (valores de atributos) debido a la *información imprecisa* o *información imperfecta*, que es la causa de la *no diferenciación* de los objetos en términos de datos disponibles y evita, en consecuencia, su asignación precisa a un conjunto.

- Cualquier conjunto de objetos que no sean discernibles se llama *conjunto elemental*, y forma un *"gránulo"* básico del conocimiento acerca del universo. Cualquier conjunto de objetos que sea la unión de algunos conjuntos elementales se considera como *"preciso"* y en el caso contrario, puede describirse como *"rough"* (impreciso, vago).
- De manera intuitiva, un *rough set* es un conjunto de objetos que, en general, no pueden ser caracterizados de manera precisa en términos de valores de un conjunto de atributos.

Sistema de información y tabla de decisión

- La percepción de cada objeto del universo está dada por la información accesible acerca de ellos a través de los valores de ciertos *atributos*. Los objetos caracterizados por la misma información no son *discernibles*.
- Los datos que generan la información se ordenan en el *sistema de información* y si se colocan en las columnas los *atributos de condición* y en las filas los objetos a describir, se construye la *tabla de decisión* si fuera posible involucrar al menos un *atributo de decisión* que corresponderá a la clase de decisión.

Aproximaciones (alta, baja y región límite)

- Los conjuntos elementales que poseen idénticos valores en los *atributos de condición* se denominan *conjuntos elementales* (*átomos*). Para cada clase del *sistema de información* es posible determinar la *aproximación alta* y la *aproximación baja*.
- Cuando el *número cardinal* de un conjunto elemental es más de uno, es probable que sus objetos pertenezcan a diferentes clases de decisión, y en este caso existe ambigüedad.
- En el modelo clásico Rough Set, la *aproximación baja* contiene todos los conjuntos elementales incluidos en la clase de decisión y la *aproximación alta* contiene todos los conjuntos elementales cuya intersección con la clase de decisión no da el conjunto vacío.
- Se determinan la *precisión de cada aproximación*, la *precisión de la clasificación* y la *calidad de la clasificación*.

Reducción de los atributos

- La reducción de atributos determina si existen *atributos redundantes* en el *sistema de información*, identificando el *mínimo conjunto de atributos* que mantenga la misma *calidad* de clasificación, aproximen los datos de la misma manera o sea un subconjunto de ellos y se denominan *reductos.*
- La intersección de los reductos se denomina *núcleo*, son los atributos comunes a todos los reductos y de ellos no se debe prescindir.

Inducción de reglas

- El conocimiento analizado se expresa en forma de *reglas* o sentencias lógicas tipo "IF <antecedentes> THEN <consecuentes>" o "IF<atributos de condición>THEN<atributos de decisión>".
- La metodología es útil cuando el conjunto de datos es *inconsistent*e y objetos descritos por los mismos valores de los *atributos de condición* son asignados a diferentes *clases de decisión*. Las *reglas de decisión* pueden generarse de las aproximaciones o de las fronteras de las clases de decisión.
- Las reglas obtenidas pueden ser *reglas determinísticas* (consistentes, precisas, exactas) o *no determinísticas* cuando las decisiones no se determinan unívocamente por las condiciones.
- Cada regla se caracteriza por el soporte-SOP o total de casos que la cumplen, la cobertura-COB que representa las razones de la decisión y la fortaleza-FOR o su respaldo en la *tabla de decisión*.

- Existen otras medidas que describen diferentes aristas del *conocimiento aproximado* (Yao, 2010) y que han sido comparados con ciertos análogos estadísticos (Tan *et al.*, 2004).

Tanto las reglas obtenidas (con sus respectivas medidas) como las salidas parciales de los pasos requeridos para extraerlas (aproximaciones, atributos no dispensables, reductos y núcleo) constituyen un sistema de resultados que debe ser analizado integradamente según el modelo RST seleccionado.

Pasos requeridos para la aplicación de la Teoría de los Conjuntos Aproximados

Para aplicar con éxito esta teoría, es imprescindible cumplir de forma detallada y cuidadosa una secuencia de pasos, cuyos objetivos se resumen en la Tabla 1 y que deberán ser del dominio del investigador, requiriendo de este un mayor enfoque a los contenidos que a los algoritmos (Yao, 2015b).

Tabla 1. Pasos a tener en cuenta para aplicar adecuadamente la Teoría de los Conjuntos Aproximados

No.	Pasos y objetivos
1	**Selección de la base de casos**: filtrar la base de datos y formar la base de casos
2	**Identificación del sistema de información:** estructurar casos seleccionados en forma matricial (filas con casos y columnas con valores de los atributos)
3	**Asignación de roles a los atributos:** identificar atributos de condición y atributos de decisión para construir la tabla de decisión.
4	**Discretización de atributos:** discretizar variables pendientes, de ser necesario.
5	**Aplicación de la metodología RST:** seleccionar el modelo RST y la herramienta computacional conveniente.
5.1	**Determinación del factor de consistencia:** verificar que sea desigual de 1 para continuar el proceso y si es igual a 1 no es necesario continuar.
5.2	**Establecimiento de las relaciones de indiscernibilidad**
5.3	**Determinación de las aproximaciones:** determinar la aproximación alta, la aproximación baja y la región límite. Si esta última es vacía no se debe continuar.
5.4	**Reducción de atributos:** obtener los reductos y analizar las variantes.
5.5	**Intersección de reductos:** encontrar el conjunto de atributos indispensables o núcleo.
5.6	**Inducción de reglas:** encontrar las asociaciones entre los atributos de condición que determinan los distintos tipos de decisión en forma de reglas.
6	**Análisis integrado:** analizar los resultados (atributos dispensables, núcleo, reglas IF-THEN con sus respectivas medidas) como nuevo conocimiento.

Los primeros cuatro pasos se dedican al análisis de la base de datos, del tipo de estudio a realizar y de la doble relación: *forma-contenido* y *naturaleza de la variable-tipo de dato* (Yao, 2015a) incorporándose además la discretización de las variables numéricas continuas. Constituyen un primer bloque que formaría parte del diseño metodológico de la investigación.

El segundo bloque (paso 5) es de tipo operativo donde se concentran los procedimientos de la *metodología RST* y se determinan las aproximaciones, los atributos dispensables, los reductos y finalmente se extraen las reglas, es decir, los resultados.

Y el tercer y último bloque (paso 6) está reservado al análisis de los resultados.

Herramienta computacional seleccionada

Del variado conjunto de herramientas computacionales disponibles actualmente (Riza *et al.*, 2014) fue seleccionada la aplicación informática ROSE2 creada en el Instituto de Ciencias de la Computación de la Universidad de Poznan, Polonia (Slowinski *et al.*, 1999), reconocida institución que con más de 40 años de experiencia continúa actualmente siendo líder en esta disciplina. La versión libre ROSE2-2.2 (Predki *et al.*, 2004), desarrollada en el lenguaje de programación principal C++, es un software modular que corre sobre el Sistema Operativo WINDOWS y tiene implementado tres modelos de exploración de datos, cuatro variantes para determinar los reductos, así como tres esquemas de inducción de reglas, todo ello con el riguroso respaldo matemático y computacional.

En esta investigación, fue seleccionado el modelo clásico de Pawlak (para el cual la *aproximación baja* contiene todos los conjuntos elementales incluidos en la clase de decisión y la *aproximación alta* todos los conjuntos elementales cuya intersección con la clase de decisión no es un conjunto vacío) y el esquema de inducción de reglas de descripción mínima (conjunto mínimo de reglas que cubren todos los objetos de un conjunto) que utiliza el algoritmo LEM2(Slowinski & Vanderpooten, 1997) .

Sistema de información y tabla de decisión para el estudio de inundaciones costeras

Caracterización de la intensidad de la inundación. A partir de los criterios precisados en el Manual de Procedimientos de Meteorología Marina del INSMET (INSMET, 2015) relativos a la altura significativa de la ola y la conciliación con la cronología de inundaciones costeras confeccionada por Hidalgo y colaboradores (Hidalgo *et al.*, 2016), las inundaciones costeras fueron clasificadas en tres clases: LIGERA, MODERADA y FUERTE.

Clasificación de los sistemas frontales. Los sistemas frontales bajo estudio (frentes fríos y bajas extratropicales que transitaron por el Golfo de México) requirieron la clasificación de los frentes fríos según su intensidad, utilizando el reconocido criterio de González explicado por (Hernández, 2013) según el viento máximo medio sostenido (Vmm). En la Tabla 2 se precisa esta clasificación además de la asociación *naturaleza de la variable-tipo de dato*.

Tabla 2. Caracterización de las situaciones meteorológicas bajo estudio, sus categorías correspondientes y la asociación de la variable con el tipo de dato

Sistemas frontales	**Clasificación de los sistemas frontales bajo estudio**			**Asociación variable-tipo de dato**	
	Intervalo	**Categoría**	**Símbolo**	**Variable**	**Dato**
Frentes fríos	Vmm ≤ 9.7 m/s	Débil	ffd	Cuantitativa continua	Códigos según criterios
	9.7<Vmm<15.3 m/s	Moderado	ffm		
	Vmm ≥ 15.3 m/s	Fuerte	fff		
Baja extratropical en tránsito por el Golfo de México			b	Cualitativa nominal	Códigos simbólicos

Variables seleccionadas de la base de datos. La Tabla 3 muestra el sistema de variables especificándose los valores o categorías de cada una según su naturaleza cuantitativa o cualitativa, numérica por intervalos o nominales, en su asociación *naturaleza de la variable-tipo de dato* lo cual es necesario tener en cuenta para el ulterior procesamiento.

Se dispuso de dos variables relativas a la boya 42003 NDBC, NOAA (26,044 LN y 85,612 LW). En este caso, para la altura significativa de la ola en la boya (HsigBoya), la partición fue efectuada según los criterios del Manual de Procedimientos de Meteorología Marina del INSMET (INSMET, 2015).

Tabla 3. Variables cualitativas y cuantitativas identificadas en el estudio con sus categorías clasificatorias en la relación variable-dato

Variables	**Criterios para la partición**				**Asociación variable-tipo de dato**	
	Símbolo	**Categorías**			**Variable**	**Dato**
Situación meteorológica	SitMET	Frente Frío	Fuerte	fff	Cuantitativa continua	Códigos según criterios
			Moderado	ffm		
			Débil	ffd		
		Baja extratropical		b	Cualitativa nominal	Códigos simbólicos
Tipología por el giro de los vientos	Tipo	Clásico		C	Cualitativa nominal	Códigos simbólicos
		Reversino		R		
		Secundario		S		
Velocidad media del viento [m/s]	VmV	VmV ≤ 9.7 m/s			Cuantitativa continua	Códigos según criterios
		9.7<VmV<15.3 m/s				
		VmV ≥ 15.3 m/s				
Dirección predominante del viento	DirPred	W, NW, NNW, NE, NNE			Cualitativa nominal	Códigos simbólicos
Duración de la inundación [horas]	Persist	6, 12, 18, 24, 30 h			Cuantitativa discreta	Códigos según Criterios
Altura significativa de la ola [m]	HsigBoya	HsigBoya ≤ 4 m			Cuantitativa continua	Códigos según criterios
		4< HsigBoya <5 m				
		HsigBoya ≥ 5 m				
Índice ENOS	IndENOS	IndENOS ≤ 48		N	Cuantitativa continua	Códigos según criterios
		48< IndENOS ≤ 150		D		
		150< IndENOS ≤ 350		M		
		350< IndENOS ≤ 800		F		
		IndENOS>800		MF		
Intensidad de la inundación	INUND	LIGERA			Cuantitativa continua	Códigos según criterios
		MODERADA				
		FUERTE				

Para el resto de las variables, fueron utilizados los mismos criterios del estudio estadístico de Hernández (Hernández, 2013), es decir:

- Para la velocidad media del viento, tanto en la estación (VmVEstación) como en la boya (VmVBoya), se utilizaron umbrales análogos a los utilizados en la clasificación de los frentes fríos.

- Para la persistencia de la inundación (Persist) fueron utilizados umbrales múltiplos de 6 horas.

- Respecto al Índice ENOS (IndENOS) como indicador adimensional (calculado a partir de la media trimestral de los últimos tres meses de la anomalía de la temperatura superficial del mar y la media de estos tres meses del índice de Oscilación del Sur) fueron establecidas cinco categorías: N-neutro, D-débil, M-moderado, F-fuerte y MF-muy fuerte.

Características de los estudios realizados: Se realizaron dos tipos de estudio, uno exploratorio y otro predictivo; en ambos el *atributo de decisión* fue la intensidad de la inundación (INUND).

En el estudio exploratorio fueron utilizadas todas las variables mostradas en la Tabla 3, con el objetivo de encontrar las diferentes relaciones entre ellas.

Para el estudio predictivo fueron seleccionadas aquellas variables que podían ser conocidas de antemano a la ocurrencia de la inundación.

En la Figura 2 se comparan las semejanzas y diferencias de los *sistemas de información* para ambos tipos de estudio, destacándose el necesario reajuste de los valores/categorías de (SitMET) que, para el estudio predicitvo se reducen a dos (bajas extratropicales que transitaron por el Golfo de México y frentes fríos) eliminándose los atributos (VmVEstación) y (Persist).

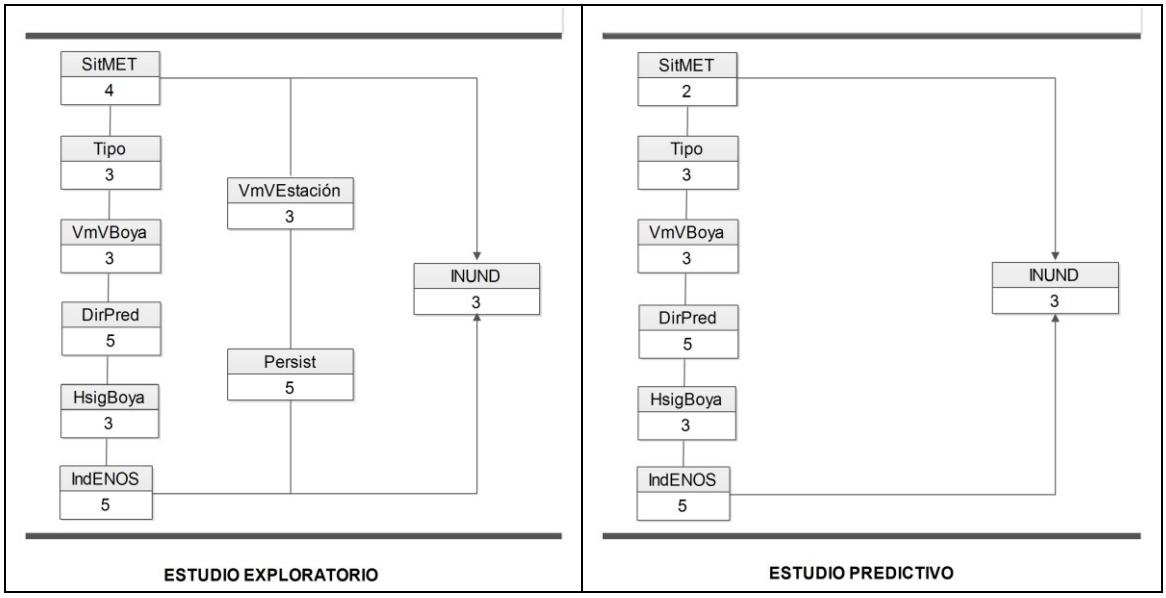

Figura 2. Comparación entre los *sistemas de información* de los dos tipos de estudio

Resultados y discusión

El total de inundaciones motivadas por sistemas frontales y que ocurrieron en el litoral habanero en el período 1980-2010 fueron 39 (26 ligeras, 9 moderadas y 4 fuertes).

1. Resultados y análisis del estudio exploratorio

Determinación de la aproximación alta y la aproximación baja. La Tabla 4 muestra que el conjunto a describir estuvo perfectamente discriminado.

Tabla 4. Resultado de determinar las aproximaciones al caso de estudio

Clase (inundación)	Total de eventos	Aproximación baja	Aproximación alta	Calidad de la aproximación
LIGERA	26	26	26	1,00
MODERADA	9	9	9	1,00
FUERTE	4	4	4	1,00

Determinación de los reductos y del núcleo. Fueron determinados dos subconjuntos del conjunto total de atributos requeridos para una descripción adecuada del *sistema de información* sometido a análisis. En la Tabla 5 se representa el conjunto intersección compuesto por (SitMET, Tipo, VmVEstación, Persist y HsigBoya) que resultaron ser los atributos indispensables para la descripción de las inundaciones costeras los cuales constituyen el núcleo.

Tabla 5. Descripción de los dos reductos obtenidos para el estudio exploratorio

Red	SitMET	Tipo	VmVEstación	VmVBoya	Persist	HsigBoya	IndENOS
1	X	X	X	X	X	X	
2	X	X	X		X	X	X

Tanto el atributo (VmVBoya), al igual que (IndENOS), complementaron por separado el conjunto mínimo de atributos requeridos para la adecuada descripción de las inundaciones costeras. El atributo (DirPred), por su ausencia en ambos reductos, no posee relevancia y sería por ello dispensable.

Inducción de reglas según el modelo clásico de Pawlak (descripción mínima).

Partiendo del modelo clásico de Pawlak y del esquema de inducción de reglas de descripción mínima, fueron generadas 20 reglas exactas. Se requirieron 9 reglas para describir las inundaciones ligeras, 8 reglas para las inundaciones moderadas y bastaron 3 reglas para las inundaciones fuertes. La diversidad de situaciones que propician inundaciones es mayor en las inundaciones ligeras y moderadas que las que motivan inundaciones fuertes.

Selección de reglas relevantes. Los mayores valores de cobertura-COB *[coverage]* dentro de cada clase de decisión, y dentro de estas los mayores valores de fortaleza-FOR *[strengh]* permitieron agrupar las reglas por su relevancia. En la Tabla 6 se detallan estas reglas relevantes con sus respectivas medidas y se añade una columna con el total de *atributos de condición* que se denomina longitud de la regla (LON).

Ejemplo de interpretación de una regla. La interpretación de una regla requiere sus medidas y debe analizar tanto la regla directa como la regla inversa. Por ejemplo, para la regla R1 de la Tabla 6, se interpreta: Si (VmVEstación<9.7m/s) y (Persist=12h), entonces (INUND=LIGERA) lo cual está respaldada por 8 casos en la *tabla de decisión* que representan el 20,5% del total de casos. La regla inversa de R1 se interpreta como: el 30,7% de las inundaciones ligeras (INUND=LIGERA) son motivadas por sistemas frontales en las que (VmVEstación<9.7m/s) y (Persist=12h), siendo coincidentes estos dos *atributos de condición* en un 20,5% del total de los casos. Esta regla es corta (LON=2) por lo que tiene un alto grado de generalidad.

Asociaciones de atributos en las inundaciones fuertes. Con solamente cuatro inundaciones costeras fuertes se obtienen tres reglas, las mejor representadas y de cobertura suficiente respecto al resto de las clases. Los peligros asociados a las inundaciones fuertes estuvieron determinados por ellas (R18, R20 y R19) y tienen el mismo alto grado de generalidad (LON=2). A partir del análisis correspondiente, se puede inferir que:

- Aun cuando la dirección predominante del viento (DirPred) fue un atributo dispensable, éste puede formar parte de las reglas, incluso de las relevantes.

- El 50% de las inundaciones fuertes (INUND=FUERTE) fueron causadas por sistemas frontales con dirección predominante del viento del Oeste (DirPred=W) aun cuando en la boya manifestaron velocidad media no tan alta (VmVBoya <9,7 m/s).

- El 50% de las inundaciones fuertes (INUND=FUERTE) fueron motivadas por bajas extratropicales que transitan por el Golfo de México en circunstancias de (IndENOS=MODERADO).

Tabla 6. Reglas de decisión más relevantes que representan las tres clases de inundaciones para el estudio exploratorio

REGLAS RELEVANTES **Estudio exploratorio-método clásico de Pawlak**		**LON**	**SOP**	**FOR**	**CER**	**COB**
R1	Si (VmVEstación < 9,7m/s) Y (Persist=12h) ENTONCES (INUND=LIGERA)	2	8	0,205	1,00	0,307
R6	Si (Tipo=C) Y (VmVBoya<9,7 m/s) ENTONCES (INUND=LIGERA)	2	6	0,154	1,00	0,231
R2	Si (VmVEstación< 9,7m/s) Y (Persist=18h) E (IndENOS=NEUTRO) ENTONCES (INUND=LIGERA)	3	5	0,129	1,00	0,192
R12	Si (Tipo=C) Y (9.7m/s<VmVBoya<15,3m/s) Y (Persist= 6h) ENTONCES(INUND=MODERADA)	3	2	0,051	1,00	0,222
R16	Si (SitMET=fffuerte) Y (DirPred=NW) Y (HsigBoya< 4m) ENTONCES(INUND=MODERADA)	3	2	0,051	1,00	0,222
R18	Si (VmVEstación< 9,7m/s) Y (DirPred=W) ENTONCES(INUND=FUERTE)	2	2	0,051	1,00	0,50
R20	Si (SitMET=baja) E (IndENOS=MODERADO) ENTONCES(INUND=FUERTE)	2	2	0,051	1,00	0,50
R19	Si (VmVBoya>15,3 m/s) Y (HsigBoya>5m) ENTONCES(INUND=FUERTE)	2	1	0,025	1,00	0,25

Significatividad de los atributos en el sistema de reglas. No todos los atributos de condición aparecen en el sistema de reglas con la misma frecuencia, siendo la misma una medida de su significatividad. Aun en el caso en que no se dispusiera de información acerca de la variable asociada el evento ENOS, fueron confeccionados los histogramas correspondientes de cada atributo de condición respecto al total de reglas que se muestran en la Figura 3 y de cuyo análisis se concluye lo siguiente:

- El atributo (IndENOS) no llegó a rebasar la significatividad de los cuatro atributos de 3 categorías como lo son (Tipo, VmVEstación, VmVBoya, HsigBoya), todos de frecuencias muy cercanas.

- Independientemente de si se incluye o se excluye (IndENOS) son características de ambas situaciones la elevada presencia del atributo de condición (Persist) seguidos de (SitMET) y la menor incidencia de aquel atributo dispensable (DirPred) que no aparecía en los reductos.

- En la situación para la que se excluye (IndENOS), las frecuencias del resto de los atributos, a excepción de (Persist) y (Tipo), se refuerzan en sus valores, siendo significativo que las variables asociadas a la boya (VmVBoya, HsigBoya) lo hacen en mayor grado.

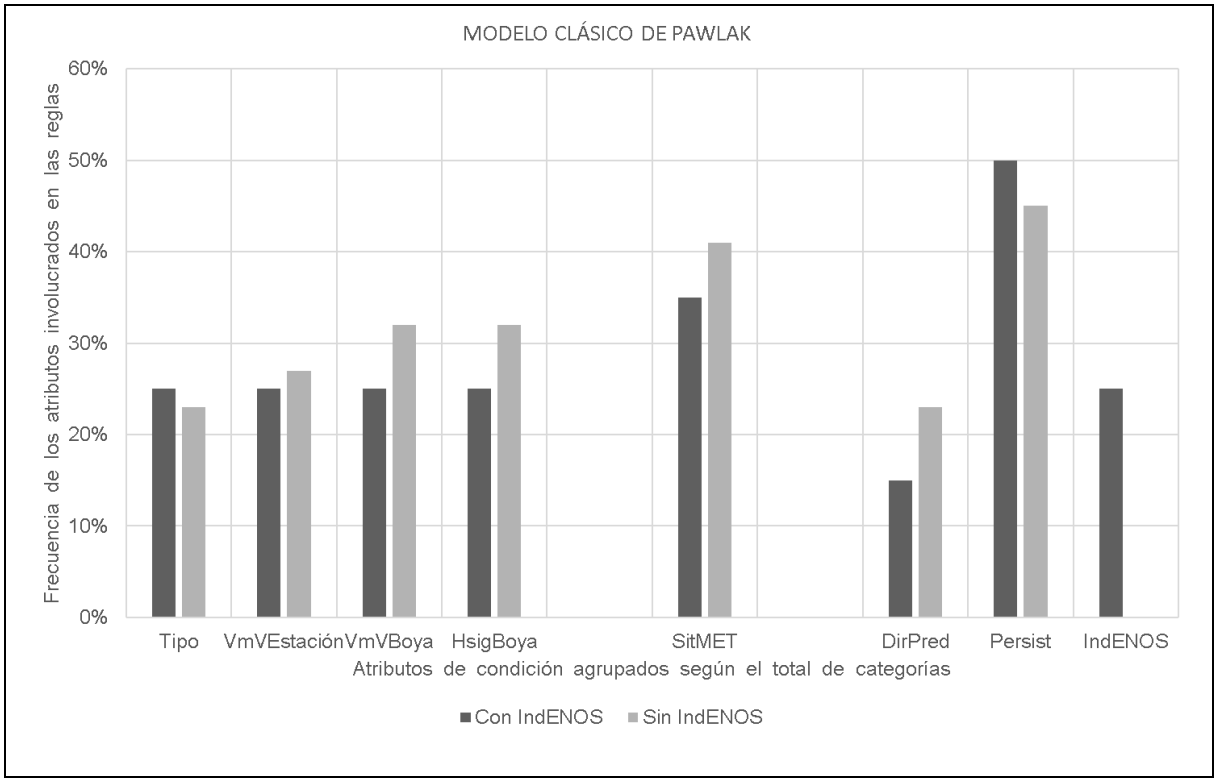

Figura 3. Histogramas comparativos de la diferente significación de los atributos de condición en las reglas obtenidas agrupados por total de categorías

La posibilidad de exclusión del (IndENOS) del sistema de información en el estudio exploratorio, permitió analizar los efectos de la no disponibilidad de este indicador específico en el sistema de información atendiendo al sistema de reglas relevantes se muestran en la Tabla 7.

De excluir (IndENOS), los cambios en las dos únicas reglas relevantes que incluían el (IndENOS), R2 para la clase (INUND=LIGERA) y R20 para (INUND=FUERTE) dan paso a otros atributos, pero mantienen la generalidad por poseer igual longitud.

Tabla 7. Modificaciones en las reglas relevantes al excluir el atributo (IndENOS) del *sistema de información* en el estudio exploratorio

	Estudio exploratorio Con IndENOS	**SOP**		**Estudio exploratorio Sin IndENOS**
R2	Si (VmVEstación< 9.7m/s) Y (Persist=18h) E (IndENOS=NEUTRO) ENTONCES (INUND=LIGERA)	5	8	Si (VmVEstación<9,7m/s), Y (DirPred=NW) Y (4m<HsigBoya <5m) ENTONCES (INUND=LIGERA)
R20	Si (SitMET=baja) E (IndENOS=MODERADO) ENTONCES (INUND=FUERTE)	2	2	Si (SitMET=baja) Y (Persist=12h) ENTONCES (INUND=FUERTE)

2. Resultados y análisis del estudio predictivo

Aproximación alta y aproximación baja. La aproximación alta y la aproximación baja para cada clase de decisión en el estudio predictivo se muestran en la Tabla 8, disminuyendo la calidad de la aproximación general en relación al estudio exploratorio, evidenciándose también marcadas diferencias en relación a la calidad de cada clase. La clase (INUND=FUERTE) está bien delimitada, no siendo así en el caso de (INUND=LIGERA) y siendo muy mala en la (INUND=MODERADA).

Tabla 8. Resultados de determinación de las aproximaciones en el estudio predictivo

Clase (inundación)	**Total de eventos**	**Aproximación baja**	**Aproximación alta**	**Calidad de la aproximación**
LIGERA	26	20	32	0,6250
MODERADA	9	3	15	0,2000
FUERTE	4	4	4	1,0000

Reductos y núcleo. Solamente fue identificado un reducto: (SitMET, Tipo, VmVBoya, DirPred, HsigBoya, IndENOS) que está compuesto por el total de atributos de condición por lo que también este subconjunto de atributos constituye el núcleo. Todos los atributos fueron esenciales, es decir, ninguno fue dispensable.

Inducción de reglas según el modelo clásico de Pawlak (descripción mínima).

Fueron generadas un total de 18 reglas de decisión, 16 exactas y 2 aproximadas que se muestran en la Tabla 9.

Las reglas aproximadas explican la mala calidad de las aproximaciones tanto para (INUND=LIGERA) como para (INUND=MODERADA) lo cual ya se había mostrado en la Tabla 8.

Tabla 9. Reglas aproximadas que se generan en el estudio predictivo, siendo R18 la de mayor generalidad

REGLAS APROXIMADAS Estudio predictivo-método clásico de Pawlak		LON
R17	Si (Tipo=C) Y (9,7m/s<VmVBoya <15.3m/s) Y (DirPred=NW) Y(HsigBoya<4m) E (IndENOS=NEUTRO) ENTONCES (INUND=LIGERA) O (INUND=MODERADA)	5
R18	Si (DirPred=NW) E (IndENOS=FUERTE) ENTONCES (INUND=LIGERA) O (INUND = MODERADA)	2

Del total de reglas exactas, resultaron 9 para (INUND=LIGERA), 4 para (INUND=MODERADA) y 3 para (INUND=FUERTE).

Para el estudio predictivo, al seleccionar las reglas exactas de mayores valores de cobertura-COB *[coverage]* y los mayores valores de fortaleza-FOR *[strengh]* dentro de cada clase de decisión se determinan las reglas de relevancia mostradas en la Tabla 10. Debido a sus bajos valores de cobertura-COB *[coverage]* no fue posible seleccionar ninguna regla relevante dentro de la clase (INUND=MODERADA), estando agrupadas junto a (INUND=LIGERA) en las reglas aproximadas.

Tabla 10. Reglas de decisión relevantes que representan solamente dos de las tres clases de inundaciones para el estudio predictivo

REGLAS RELEVANTES Estudio predictivo-método clásico de Pawlak		LON	SOP	FOR	CER	COB
R4	Si (Tipo=C) Y (VmVBoya<9,7 m/s) ENTONCES (INUND=LIGERA)	2	6	0,153	1,00	0,231
R1	Si (4m<HsigBoya<5m) E (IndENOS=NEUTRO) ENTONCES (INUND=LIGERA)	2	5	0,128	1,00	0,192
R14	Si (SitMET=baja) Y (DirPred= W) ENTONCES (INUND=FUERTE)	2	2	0,051	1,00	0,500
R16	Si (SitMET=baja) E (IndENOS=MODERADO) ENTONCES (INUND=FUERTE)	2	2	0,051	1,00	0,500
R15	Si (VmVBoya>15.3m/s) Y (HsigBoya> 5m) ENTONCES (INUND=FUERTE)	2	1	0,025	1,00	0,250

Con solamente cuatro inundaciones costeras fuertes (INUND=FUERTE), se obtienen las reglas mejor representadas y que a su vez mejor fundamentan este tipo de inundación. Los peligros asociados a las inundaciones fuertes estuvieron determinados por la relevancia de las tres reglas que las describen (R14, R16 y R15) y a partir de las cuales se pudo conocer que los mayores peligros de (INUND=FUERTE) se encuentran vinculados a:

- El 25% de las inundaciones fuertes (INUND=FUERTE) son provocadas por sistemas frontales donde en la boya hayan sido detectados vientos fuertes (VmVBoya>15.3m/s) y olas altas (HsigBoya> 5m).

- El 50% de las inundaciones fuertes (INUND=FUERTE) están motivadas por bajas extratropicales que transitan por el Golfo de México con dirección predominante del viento del Oeste (DirPred=W).

- El 50% de los casos de inundaciones fuertes (INUND=FUERTE) están motivadas por bajas extratropicales que transitan por el Golfo de México en circunstancias de (IndENOS=MODERADO).

Las dos últimas interpretaciones permitieron identificar que, para el caso de las bajas extratropicales que transitan por el Golfo de México, tanto las condiciones de (IndENOS=MODERADO) como la dirección de viento predominante del Oeste (DirPred=W) representan un alto peligro de inundaciones fuertes.

3. Observaciones comparativas de ambos estudios para las inundaciones fuertes

Las tres reglas que caracterizan las inundaciones fuertes, en ambos tipos de estudio, poseen igual longitud (LON=2) y tienen por tanto gran generalidad. Dos de ellas, reflejadas en la Tabla 11, coinciden plenamente en cuanto a atributos.

Tabla 11. Reglas de asociación coincidentes, en ambos tipos de estudio, para las inundaciones fuertes

SitMET	VmVBoya	HsigBoya	IndENOS	INUND	SOP	FOR	COB
baja			MODERADO	FUERTE	2	0,051	0,50
	>15,3m/s	>4m		FUERTE	1	0,025	0,25

La tercera de ellas coincide solamente en el atributo (DirPred=W) pero vinculan de forma separada un segundo atributo, tal y como se muestra en la Tabla 12, en el caso exploratorio (VmVEstación) y en el estudio predictivo (SitMET).

Tabla 12. Reglas de asociación que resultan diferentes para inundaciones fuertes en los dos estudios

Tipo de estudio	SitMET	DirPred	VmVEstación	INUND	SOP	FOR	COB
Exploratorio		W	<9,7 m/s	FUERTE	2	0,051	0,50
Predictivo	baja	W		FUERTE	2	0,051	0,50

Estas consideraciones son coincidentes con los criterios de expertos que, mediante otras vías, han llegado a identificar la significatividad de las diferentes variables (Mitrani *et al.*, 2011) (Mitrani *et al.*, 2014) y que motivan inundaciones costeras fuertes en el litoral habanero.

Para el diseño de sistemas de alerta temprana, el valor de los resultados del estudio predictivo es indudablemente superior al del estudio exploratorio, pues se utilizan solamente atributos cuyos valores/categorías pueden ser conocidos de antemano por parte de los especialistas.

Conclusiones

Desde una *perspectiva no estadística*, la aplicación de la Teoría de los Conjuntos Aproximados en el estudio de las inundaciones costeras permitió identificar asociaciones importantes entre valores de algunas variables meteorológicas y la intensidad de la inundación que pueden complementar los estudios de tipo estadístico.

Los resultados obtenidos permiten recomendar que, para el caso de los sistemas frontales – tanto frentes fríos como bajas extratropicales que transitan oor el Golfio de México – es importante atender los

registros previos de la velocidad del viento y la altura significativa de la ola en la boya 42003 situada en el Golfo de México por sus posibles efectos en el litoral habanero.

Para las bajas extratropicales que transitan por el Golfo de México, se identifican situaciones extremas de inundaciones en el litoral habanero tanto si la dirección predominante del viento fuera del Oeste como en aquellos casos de circunstancias moderadas en relación a la presencia del evento ENOS.

Referencias bibliográficas

1. Hernández, I. 2013. *Las inundaciones costeras generadas por sistemas frontales en el malecón habanero*. Trabajo de Diploma-Licenciatura en Meteorología, La Habana, Cuba: Instituto Superior de Tecnologías y Ciencias Aplicadas, 97 p.

2. Hidalgo, A. 2016. *Metodología para el estudio climático de las inundaciones costeras en Cuba*. Tesis de Maestría en Ciencias Meteorológicas, La Habana, Cuba: Instituto Superior de Tecnologías y Ciencias Aplicadas, 79 p.

3. Hidalgo, A.; Pérez, G.; Mitrani, I.; Hernández, N.; Córdova, O. L.; Regueira, V.; Ramírez, W. & Rodríguez, C. M. 2016. *Cronología de las inundaciones costeras en Cuba-Programa Meteorología y Desarrollo Sostenible del País*. Proyecto 'Procedimiento para la ejecución y uso de las observaciones del estado de la superficie marina desde estaciones costeras, en la predicción de oleaje e inundaciones costeras en territorio cubano', no. P211LH007-015, La Habana, Cuba: Instituto de Meteorología, Cuba, p. 43.

4. INSMET 2015. *Manual de Procedimientos de Meteorología Marina*. La Habana, Cuba: INSMET, 131 p.

5. JeraldBeno, T. R. & Karnan, M. 2012. "Dimensionality Reduction: Rough Set Based Feature Reduction". *International Journal of Scientific and Research Publications*, 2(9): 1–6.

6. Li, S. & Hong, Z. 2017. "An Approach of Rough Set for Data Fusion". In: *Proceedings APETC 2017*, *Asia-Pacific Engineering and Technology Conference*, pp. 1809–1812, ISBN: 978-1-60595-443-1.

7. Mitrani, I.; García, E.; Hidalgo, A.; Hernández, I.; Salas, I.; Pérez, R.; Rodríguez, C. M. & Pérez, A. L. 2011. *Las inundaciones costeras en Cuba y sus tendencias climáticas*. (ser. Segunda Comunicación Nacional sobre el Cambio Climático), CITMA-Agencia Medio Ambiente, La Habana, Cuba: INSMET.

8. Mitrani, I.; García, E.; Hidalgo, A.; Hernández, I.; Salas, I.; Pérez, R.; Rodríguez, C. M. & Pérez, A. L. 2014. *Inundaciones costeras en Cuba. Estructura termohalina y su influencia en las inundaciones*. La Habana, Cuba: Agencia del Medio Ambiente. Cuba.

9. Pawlak, Z. 1981a. *Classification of Objects by Means of Attributes*. Reports, no. 429, Warsaw, Poland: Institute of Computer Science, Polish Academy of Sciences.

10. Pawlak, Z. 1981b. “Information systems – theoretical foundations”. *Information Systems*, 6: 205–218.

11. Pawlak, Z. 1994. *Vaguenes and Uncertainy-A Rough Set Perspective*. ICS Research Report, no. 19/94, Warsaw: Warszaw Technical University.

12. Pawlak, Z. & Skowron, A. 2007. “Rudiments of Rough Sets”. *Information Sciences*, 177: 3–27, DOI: 10.1016/j.ins.2006.06.003.

13. Predki, B.; Slowinski, R.; Stefanowski, J.; Susmaga, R. & Wilk, S. 2004. *ROSE 2. Rough Set Data Explorer*. version 2.2, [ROSE], Poznan University, Poznan, Poland, Available: <http://www-idss.cs.put.poznan.pl/rose>.

14. Riza, L. S.; Janusz, A.; Bergmeir, C.; Cornelis, C.; Herrera, F.; Slezak, D. & Benitez, J. M. 2014. “Implementing algorithms of rough set theory and fuzzy rough set theory in the R package ‘RoughSets’”. *Information Sciences*, 287: 68–89, DOI: 10.1016/j.ins.2014.07.029.

15. Slowinski, K.; Slowinski, R.; Stefanowski, J. & Fibak, J. 1999. *ROSE2. Rough Set Data Explorer-User’s Guide*. Laboratory of Intelligent Decision Support Systems, Poznan University, Available: <http://www-idss.cs.put.poznan.pl/rose>, [Consulted: July 15, 2016].

16. Slowinski, R. & Vanderpooten, D. 1997. “Similarity relation as a basis for rough approximations”. In: Wang, P. P. (ed.), *Advances in Machine Intelligence and Soft Computing*, vol. NC, Bookwrights Raleigh, pp. 17–33.

17. Tan, P.-N.; Kumar, V. & Srivastava, J. 2004. “Selecting the right objective measure for association analysis”. *Information Systems*, 29: 293–313.

18. Yao, Y. Y. 2010. “Notes on Rough Set Approximations and Associated Measures”. *Journal of Zhejiang-Ocean University*, 29(5): 399–410.

19. Yao, Y. Y. 2015a. “Rough Set Approximations: A concept analysis point of view”. In: Ishibuch, H. (ed.), *Encyclopedia of Life Support Systems (EOLSS)*, vol. Computational Intelligence-Volume I, pp. 282–296.

20. Yao, Y. Y. 2015b. “The two sides of the theory of rough sets”. *Knowledge-Based Systems*, 80: 67–77, DOI: 10.1016/j.knosys.2015.01.004.

21. Zhao, Y.; Yao, Y. Y. & Luo, F. 2007. “Data Analysis Based on Discernibility and Indiscernibility”. *Information Sciences*, 177: 4959–4976.