Vol 44 Nº1 (2019) 45 – 65

Meteorologica

TÉCNICAS DE CLASIFICACIÓN SUPERVISADA PARA LA DISCRIMINACIÓN ENTRE ECOS METEOROLÓGICOS Y NO METEOROLÓGICOS USANDO INFORMACION DE UN RADAR DE BANDA C

Sofia Ruiz Suarez[1,2], Mariela Sued[2,7], Luciano Vidal[1], Paola Salio[3,4,5,7], Daniela Rodriguez[2,7], Stephen Nesbitt[6] y Yanina Garcia Skabar[1,5,7]

[1]Servicio Meteorológico Nacional, Buenos Aires, Argentina
[2]Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales - UBA
[3]Centro de Investigaciones del Mar y la Atmósfera- UBA
[4]Departamento de Ciencias de la Atmosfera y los Océanos - UBA
[5]UMI-Instituto Franco Argentino sobre Estudios del Clima y sus ImpactosCNRS 3351, Buenos Aires, Argentina
[6]Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana-Champaign, USA
[7]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

(Manuscrito recibido el 8 de agosto de 2017, en su versión final el 2 de enero de 2018)

RESUMEN

Los datos provenientes de los radares meteorológicos son de suma importancia para el diagnóstico y monitoreo de los sistemas que producen precipitación y sus posibles fenómenos severos asociados. Los ecos causados por objetivos no meteorológicos introducen errores en la información por lo que es necesario detectar la presencia de los mismos previo a la utilización de los datos. Este trabajo presenta cuatro técnicas de clasificación supervisada basadas en diferentes modelos estadísticos que buscan dar una respuesta a este problema.

Asimismo como parte importante de este trabajo, se aplicaron técnicas de remuestreo estadísticas sobre los datos de entrenamiento, las que permitieron hacer un análisis más completo sobre los resultados. En la actualidad, las técnicas de remuestreo son herramientas fundamentales en la estadística moderna. Las mismas, a partir de simulaciones sobre los datos, permiten obtener información adicional sobre los modelos planteados.

Para este trabajo se realizó un estudio de caso con datos provenientes del radar meteorológico Doppler banda C de doble polarización ubicado en la Estación Experimental Agropecuaria INTA Anguil (La Pampa). Partiendo de la clasificación manual de un experto, se aplicaron cuatro métodos de clasificación supervisada de diferentes grados de flexibilidad en su estructura: Modelo lineal, Modelo Cuadrático, Modelo Logístico y Modelo de Bayes Naive. Luego se compararon los resultados y se evaluó el desempeño de cada uno de ellos. Si bien se encontraron dificultades a la hora de clasificar las zonas de frontera entre clases, los resultados obtenidos fueron adecuados,mostrando el mejor desempeño el modelo menos flexible, el modelo lineal. Se considera necesario seguir avanzando en esta línea de investigación a fin de incorporar una mayor cantidad de casos y tener una mayor significancia de los resultados.

Dirección Electrónica: sofi.ruizsuarez@gmail.com.ar

Palabras clave: clasificación supervisada, remuestreo, eco no meteorológico, radar meteorológico.

SUPERVISED CLASSIFICATION TECHNIQUES FOR DISCRIMINATION BETWEEN METEOROLOGICAL AND NON-METEOROLOGICAL ECHOES USING A C-BAND RADAR

ABSTRACT

Data coming from meteorological radars is of the utmost importance for the diagnosis and monitoring of precipitation systems and their possible associated severe phenomena. The echoes caused by objectives that are not meteorological introduce errors in the information. Therefore, it is necessary to detect their presence before using this data. This paper presents four supervised classification techniques based on different models which seek to give an answer to this problem.

In addition, as an important part of this work, resampling techniques were implemented on the training set in order to further asses the results. Resampling methods are an indispensable tool in modern statistics. Those techniques provide additional information about the model of interest by repeatedly drawing samples from then data.

Based on data from a C-band Dual-PolarizationDoppler weather radar located in Anguil and from a previous expert's manual classification, four supervised classification methods with different degrees of flexibility in their structure were implemented: Lineal Model, Quadratic Model, Logistic Model and Bayes Naive Model. Finally, the results of each of them were assessed and compared. Although difficulties were encountered in classifying boundary zones between classes, the results obtained were adequate, showing the best performance in the least flexible model, the linear one. It is considered necessary to keep working in this line of research in order to include more cases in the analysis and allow a better inference on the results.

Keywords: supervised classification, resampling methods, non-meteorological echoes, weather radar.

1. INTRODUCCIÓN

Los radares meteorológicos son una herramienta fundamental para el diagnóstico y monitoreo de los sistemas precipitantes y los posibles fenómenos meteorológicos severos asociados como granizo, ráfagas destructivas de viento e incluso tornados.

Para poder utilizar la información obtenida a partir de los radares meteorológicos es necesario en primera instancia aplicar metodologías de control de calidad tendientes a minimizar las fuentes de error (Zawadzki, 1984) presentes en los datos con el objetivo de poder emplear los mismos rutinariamente en aplicaciones relacionadas con estimaciones cuantitativas de precipitación, detección de granizo, elaboración de pronósticos a muy corto plazo, asimilación de datos en modelos numéricos de pronóstico del tiempo, diagnóstico de tiempo severo, entre otras aplicaciones. Uno de los problemas más recurrentes en los datos de radar meteorológico, es la presencia de ecos no meteorológicos tales como ecos biológicos (insectos, pájaros), ecos de terreno (edificios, montañas), o incluso presencia de propagación anómala (Rico-Ramirez

y Cluckie, 2008; Berenguer y otros, 2006).

Esta problemática ha sido abordada por varios autores a lo largo de las últimas décadas de diferentes maneras. Se han desarrollado distintas técnicas, las cuales pueden dividirse en las que implementan filtro directamente en el procesador de señal I\Q del radar (Siggia y Passarelli, 2004) y aquellas que lo hacen sobre los momentos obtenidos (reflectividad, doppler, coeficiente de correlación, entre otras variables). Sobre estas últimas a su vez, se pueden distinguir tres tipos de abordajes: las que utilizan técnicas de árboles de decisión (Stein y Smith, 2001), las que lo hacen en base a redes neuronales (Lakshmanan y otros,2010; Greku y Krajewski, 2000), y las que utilizan las llamadas técnicas de lógica difusa (Cho y otros, 2006; Gourley y otros, 2006; Hubbert y otros, 2009; Bo Young Ye y otros, 2015; Berenguer y otros, 2006; Rico-Ramírez y Cluckie, 2008; entre otros). Es posible notar en los últimos años una leve tendencia a utilizar preferentemente los modelos basados en lógica difusa. En su mayoría los trabajos realizados hacen un análisis pixel por pixel y se basan, según la frecuencia electromagnética de observación utilizada por el radar y las variables disponibles, en distinguir comportamiento de la informaciónen dos clases: meteorológica y no meteorológica.

Numerosos autores, entre ellos Moszkowiczy otros (1993), Berenguer y otros (2006), Rico-Ramirez y Cluckie (2008), Gourley y otros (2006) buscan estimar para cada pixel la probabilidad de pertenecer a la clase “eco no meteorológico” y proponen clasificadores basados en la regla de clasificación de Bayes. En particular, Moszkowicz y otros (1993) a fin de identificar la presencia de ecos de terreno asociados a propagación anómala propusieron un método de clasificación basado en la regla de Bayes tomando como campos de entrada en su modelo la elevación de la estrategia de escaneo con mayor valor de reflectividad, el valor de ese máximo y el gradiente horizontal de la reflectividad, entre otros. Los autores suponen que las funciones de probabilidad condicional de cada campo sujeto a cada tipo de eco son gaussianas y luego estiman los parámetros de dichas distribuciones con datos previamente seleccionados y clasificados de forma manual por un experto.

Más recientemente Berenguer y otros (2006) presentan un trabajo focalizado hacia la eliminación de propagación anómala. Para cada pixel evalúan la posibilidad de que la medición haya sido afectada por dicho fenómeno asociando un valor entre 0 y 1. Con este objetivo, analizan las distribuciones de las frecuencias de cada campo estudiado condicional al tipo de eco. A partir de la regla de clasificación de Bayes, derivan la probabilidad condicional de cada pixel de estar afectado según el valor de los campos. Propone dos configuraciones distintas del algoritmo dependiendo si se trata de zonas cercanas al mar o no.

Estudios previos han demostrado la gran utilidad de disponer de información polarimétrica para identificar áreas con presencia de ecos no meteorológicos. Trabajos como Schur y otros (2003) y Cho y otros (2006) presentan clasificadores basados en técnicas de lógica difusa para radares de doble polarización. Los autores definen las variables predictoras a partir del comportamiento de las variables polarimétricas tanto a nivel espacial como temporal en los distintos tipos de ecos.

En el estudio presentado por Gourley y otros (2006) también se trabaja sobre datos provenientes de radares de doble polarización con técnicas de lógica difusa. En este caso los autores también tienen en cuenta el comportamiento de la velocidad radial y de la continuidad en el campo de la reflectividad horizontal. Sugieren considerar funciones de pertenencia derivadas a partir de una estimación no paramétrica de la densidad de cada campo en cada eco. Rico-Ramirez y Cluckie (2008) presentan un clasificador en donde la Regla de Bayes y la lógica difusa son protagonistas. Trabajan con radares de banda C de doble polarización. Este modelo se diferencia de los algoritmos presentados por

Gourleyy otros (2006) y por Schury otros (2003) en cuanto a la forma de calcular la textura de las variables.

A diferencia de muchos de los trabajos previamente mencionados, el presente trabajo busca abordar la problemática de la diferenciación entre tipos de ecos a partir de técnicas de estadística clásica que modelan el problema según las leyes de la probabilidad. Según Kosko (1994) la lógica difusa y la probabilidad difieren tanto en lo conceptual como en lo teórico, pero a la vez coinciden en varios puntos. Ambos sistemas combinan conjuntos y proposiciones de manera asociativa, distributiva y conmutativa, y a la vez describen incertidumbres a partir de una cierta cantidad perteneciente al intervalo [0,1]. La diferencia entre ambos enfoques radica en cómo son considerados los conjuntos y sus complementos. Para el enfoque clásico la intersección entre un conjunto y su complemento es vacía, por lo tanto la probabilidad de que esto ocurra es cero. En cambio, en la teoría de la lógica difusa esto no siempre es verdadero, es decir podría pasar que existiera algún elemento en la intersección entre un conjunto y su complemento.

Actualmente en Argentina se cuenta con radares Doppler de doble polarización banda C (frecuencia de 5.6 GHz) que funcionan de forma operativa y son utilizados a diario por el Servicio Meteorológico Nacional. Además, se está llevando a cabo el proyecto SiNaRaMe (Sistema Nacional de Radares Meteorológicos) que tiene como fin expandir la red actual de radares a partir de la incorporación de instrumentos de fabricación nacional. Este crecimiento vertiginoso de la disponibilidad de datos de radar en nuestro país implica la necesidad de avanzar en la implementación de metodologías de control de calidad de la información generada para poder contar con datos más precisos que puedan ser utilizados en las distintas aplicaciones.

El objetivo del presente trabajo es introducir y evaluar el desempeño de cuatro métodos de clasificación supervisada a fin de mejorar la diferenciación entre ecos meteorológicos y no meteorológicos como un primer paso dentro del desarrollo de un sistema de control de calidad de la información de radar a ser implementada por el Servicio Meteorológico Nacional. Los métodos presentados están basados en la Regla de clasificación de Bayes. Este trabajo propone hacer un análisis del comportamiento de dichos métodos, para luego compararlos e inferir sobre la capacidad discriminante de cada uno de ellos. Para la validación y testeo de los diferentes procedimientos, se utilizaron técnicas estadísticas de remuestreo de manera de poder hacer un análisis más completo de los resultados obtenidos.

El trabajo se organiza de la siguiente forma. En la Sección 2 se describen los datos utilizados, en la Sección 3 se presenta la metodología a seguir y se exponen los métodos de clasificación utilizados. La discusión de los resultados obtenidos se expone en la Sección 4. Por último en la Sección 5 se dan las conclusiones del trabajo y se proponen las líneas de trabajo futuro.

2. DATOS

Se utilizaron datos del radar meteorológico Selex SI Gematronik Doppler de doble polarización en banda C (5.6 GHz) instalado en el predio de la Estación Experimental Agropecuaria INTA Anguil (La Pampa), ubicado en 36º 22’ 22,9”S y 63º 58’ 58”O (Figura 1). Los datos analizados corresponden a la estrategia de escaneo que genera un volumen de datos cada 10 minutos conformado por un total de 10 elevaciones de antena que varían entre 0,5 y 19,2 grados, con una resolución en rango de 0,25 km y de 1º en azimut (ancho del haz), y un alcance máximo de 120 kilómetros. Se seleccionaron los volúmenes con el mayor tiempo de emisión entre pulsos a fin de tener una velocidad nyquist del radar alta, en este caso de 40 metros por segundo para el escaneo seleccionado. Esto se corresponde con un valor de frecuencia de repetición de pulso de 1000 y 750 Hz considerando un procesamiento de señal staggered.

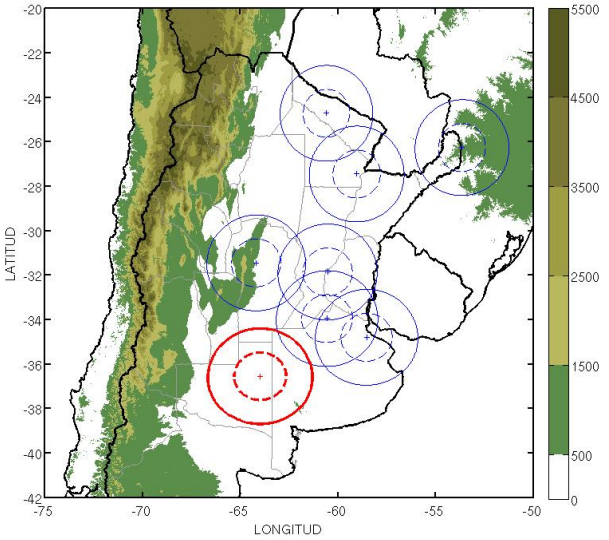

Figura 1: Ubicación del radar meteorológico Doppler banda C de doble polarizaciónen la Estación Experimental Agropecuaria INTA Anguil. En círculos rojos llenos se indica el área de cobertura de 240 km, mientras que en punteado el de 120 km. En círculos azules se muestra la posición de los radares de banda C existentes en Argentina. En sombreado se muestra la topografía en metros.

Para este trabajo de las variables generadas por el radar se consideraron: coeficiente de correlación co-polar (ρ_{HV}), velocidad radial Doppler (V), y reflectividad diferencial (ZDR), para la primera y segunda elevación de antena del radar (0,5 y 1,3 grados). Al considerar la velocidad Doppler como variable de entrada es necesario tener en cuenta la posible presencia de aliasing en los datos (Battan, 1973). Dado que en los casos trabajados no se detectó este efecto, no se realizaron correcciones previas a los datos. No obstante, si existieran errores por aliasing se deberían corregir con el fin de que los resultados sean consistentes.

Tal como se explicará en la próxima sección, los algoritmos de clasificación supervisada que aquí se presentan se construyen utilizando una muestra de entrenamiento donde, para cada observación, se dispone del valor de las variables observadas y la categoría a la cual la observación pertenece(en este caso entre eco meteorológico y no meteorológico). Esta información es fundamental para la construcción del procedimiento de clasificación que determinará la clase a la que pertenece una nueva observación en función del valor de las variables disponibles medidas en la nueva observación. Con este objetivo, se analizó un conjunto de casos y se realizó una clasificación manual asistida por un meteorólogo experto en el área. El experto meteorólogo realizó una inspección visual donde se consideró la variabilidad temporal y la estructura espacial de los ecos de radar para determinar la pertenencia a una u otra categoría. Para el caso de ecos meteorológicos, se buscó ecos con un apreciable desarrollo vertical (observable al menos en las tres primeras elevaciones de la antena) en la variable reflectividad horizontal, que al mismo tiempo presenten velocidades Doppler distintas de cero y valores de coeficientes de correlación HV cercanos a la unidad. En contraste, los ecos no meteorológicos no presentan un desarrollo vertical apreciable, ya que mayormente se ubican en el primer kilómetro de la atmósfera y también presentan una señal con una textura espacial no homogénea tanto en el caso de ecos biológicos como en ecos originados por propagación anómala. De estos casos, se seleccionaron algunos de ellos para ser utilizados como muestra de entrenamiento, y otros para ser utilizados como datos de testeo.

A fin de determinar los casos de entrenamiento se tuvo en consideración que estuvieran disponibles todas las variables antes mencionadas. Además se buscó que estos casos representaran diferentes configuraciones en relación a la distribución espacial de los ecos meteorológicos y no meteorológicos. Se consideraron entonces las siguientes fechas para formar parte de la muestra de entrenamiento:

- 20 de junio de 2009
- 22 de noviembre de 2009
- 1 de enero de 2010
- 6, 4, 9, 20, 21 y 27 de febrero 2010

El número total de píxeles con información fue

de 1.187.967. De ellos, el 47,1 % correspondiente a señal meteorológica y un 52,9 % a señal no meteorológica.

Todos los clasificadores se testearon en un mismo conjunto de imágenes provenientes de cuatro escenarios correspondientes a cuatro eventos distintos (Tabla 1). Se contó con tres imágenes de cada uno de estos escenarios, resultando así un total de 12 imágenes. Al momento de la selección de estas imágenes se procuró que las mismas resulten representativas de los escenarios típicos encontrados: situaciones donde los ecos se encuentren separados, situaciones con los ecos mezclados, situaciones con eco no meteorológico sobre el sitio del radar, situaciones con presencia de eco meteorológico intenso, entre otros.

3. METODOLOGIA

En vista de los trabajos discutidos y según los resultados obtenidos por los mismos, para el presente trabajo se consideró trabajar con métodos de clasificación netamente estadísticos, intentando en principio comprender el funcionamiento de los mismos, para luego poder comparar los resultados provenientes de otras metodologías propuestas.

Junto con la técnica de clasificación a utilizar, son las variables de entrada, las que determinan la bondad de un método. Como es de esperar la selección de las mismas depende en parte de su capacidad discriminante, pero a la vez es necesario considerar su disponibilidad. Varios autores han remarcado la capacidad predictiva que posee la variable ρ_{HV} por si sola en cuanto a identificación de ecos no meteorológicos frente a los meteorológicos: valores bajos de ρ_{HV} se suelen relacionar con ecos no meteorológicos (Ryzhkov y Zrnic, 1998; Gourley y otros, 2006). No obstante, objetivos como el granizo grande podrían llegar a tener valores bajos de este factor haciendo que una región con presencia de eco meteorológico sea catalogada como no meteorológica.

El presente trabajo propone considerar el desvío estándar en una ventana de 3 x 3 en rango y azimut de la reflectividad diferencial (SZDR). Es decir, para cada punto se calcula el desvío estándar de Z_{DR} sobre la región de 9 vecinos cuyo centro es el punto en cuestión. Finalmente, las variables empleadas como variables de entrada (variables predictoras) para los algoritmos desarrollados son:

- Coeficiente de correlación co-polar (ρ_{HV})
- Velocidad Radia lDoppler (V)
- Desvío de la reflectividad diferencial (SZDR)

Si bien la variable SZDR otorga una noción de la estructura espacial al análisis, se consideró necesario incorporar un proceso final a la clasificación obtenida por los modelos (Postproceso), de manera de darle mayor importancia a la caracterización espacial de los datos. Partiendo de la clasificación inicial del modelo, se reasignaron algunos puntos teniendo en cuenta la clase de sus vecinos. Más específicamente, sobre cada píxel se analizó una ventana de 3×3 (en rango y azimuth), si la mayoría de sus vecinos (la mitad +1) fueron asignados a su misma clase, entonces no se hizo ningún cambio, sino, se cambió la clase previamente asignada.

Uno de los inconvenientes encontrados fue que muchas veces los datos con los que se contaba tenían variables faltantes, es decir a algunos puntos les faltaba información de alguna de las variables (ρ_{HV}, V, y/o $SZDR$). A fin de salvar este inconveniente se procedió de la siguiente manera: si la falta de dato se daba en la muestra de entrenamiento, simplemente aquellos puntos fueron eliminados. En cambio, cuando el punto con una variable faltante era parte del conjunto de testeo, se clasificó dicho punto considerando el modelo con las variables disponibles. Es decir, para cada metodología se ajustaron cuatro modelos, uno con las tres variables y los otros tres con las variables tomadas de a dos. Luego dependiendo de la disponibilidad de variables en cada pixel se predijo su clase según alguno de estos cuatro modelos ajustados. Por ejemplo si para el píxel p, se contara con las variables ρ_{HV} y V, y SZDR no estuviese disponible, se clasificaría el punto a partir del modelo que utiliza como

Fecha	Ventana Horaria	Características
20 de junio 2009	Desde 04:00 hasta 06:00 UTC	- Presencia de ecos meteorológicos extendidos espacialmente. - Sobre el sitio del radar se observa eco no meteorológico.
22 de noviembre 2009	Desde 08:30 hasta 09:30 UTC	- Presencia de eco meteorológico moderado. - Alrededor del radar: eco no meteorológico
9 de febrero 2010	Desde 21:00 hasta 23:00 UTC	- Mezcla de eco meteorológico y no meteorológico. - Eco meteorológico intenso.
27 de febrero de 2010	Desde 00:00 hasta 02:00 UTC	- Presencia de ecos no meteorológicos alrededor del sitio del radar aislados de ecos meteorológicos.

Tabla I: Casos para la validación

variables (ρ_{HV},V). Si la falta de dato fuera en dos o más variables, se clasificaría al punto como sin clase (SC). En la Figura 2 esquematizamos de manera sencilla la metodología a seguir.

Según los métodos de clasificación que se desarrollaran en la sección 3.1 se clasificaron las imágenes del conjunto de testeo (Tabla I). Se asignó a cada pixel en una de las tres categorías: meteorológica (M), no meteorológica (NM) o sin clase (SC). Estos resultados fueron utilizados para evaluar la performance de cada método, comparando la verdadera clase de cada uno de ellos con la asignada por el método de clasificación. Una vez desarrollados los métodos y obtenidos los resultados para los casos de testeo, se buscó hacer un análisis más completo sobre la performance de los clasificadores. Fue necesario entonces darle mayor variabilidad a los datos. Con este objetivo se implementaron técnicas de remuestreo sobre los datos de entrenamiento. El detalle de este procedimiento se explica en la Sección 3.2.

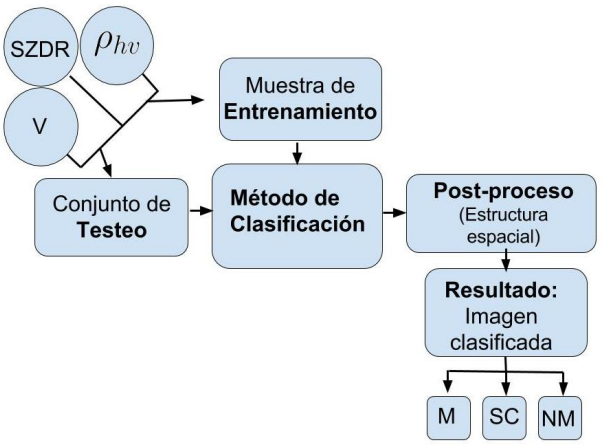

Figura 2: Diagrama de flujo sobre la metodología a seguir para el problema de clasificación en imágenes de radar propuesto en el presente trabajo. La letra M corresponde a la categoría Eco Meteorológico, la NM a Eco No Meteorológico y SC a datos sin clasificar.

3.1 METODOS DE CLASIFICACION SUPERVISADA

Las técnicas de clasificación supervisada parten de muestras pre-clasificadas con las que se

aspira aprender cómo discriminar (o clasificar) nuevas observaciones. Los métodos abordados en este trabajo toman como premisa la Regla de Clasificación de Bayes, la cual se basa en asignar la nueva observación a la clase con mayor probabilidad condicional. Más específicamente, si se denota con $x = (x_v, x_\rho, x_{szdr})$ a los valores de las variables V, ρ_{HV} y $SZDR$ en cierto punto p (información disponible), la Regla de Bayes asignará al punto p el tipo de eco con mayor probabilidad condicional. Es decir, si

$$P(M|x) > P(NM|x) \tag{1}$$

se asignará p a la clase meteorológica, de lo contrario se lo asignará a la clase no meteorológica, siendo que $P(M|x)$ y $P(NM|x)$ denotan la probabilidad de Eco Meteorológico (M) y de Eco No Meteorológico (NM) respectivamente, condicional a observar x.

La expresión (1) puede ser reformulada en términos de las probabilidades $P(M)$ y $P(NM)$ de cada clase, combinadas con las funciones de densidad ($f_M(x)$ y $f_{NM}(x)$) del vector $x = (x_v, x_\rho, x_{szdr})$ en las mismas, obteniéndose así que (1) resulta equivalente a

$$P(M)f_M(x) > P(NM)f_{NM}(x) \tag{2}$$

Las representaciones (1) y (2) sugieren dos enfoques diferentes para construir reglas de clasificación. Por un lado, existen aquellas basadas en (1), que modelan probabilidades condicionales. Por otra parte, la representación (2) da origen a clasificadores mediante la estimación de las funciones de densidad condicional en cada categoría, siendo que en muchas aplicaciones las probabilidades a priori, es decir P(M) y P(NM), suelen suponerse iguales (ambas con valor 0,5).

Existe una gran variedad de técnicas, que según el problema y los datos disponibles son más o menos adecuadas. En lo que resta de esta sección se describirán brevemente las metodologías que fueron utilizadas en este trabajo.

3.1.1. Modelo Lineal Discriminante (LDA)

El modelo discriminante lineal parte de la base que las funciones de densidad $f_M(x)$ y $f_{NM}(x)$ son gaussianas, ambas con distintos vectores de medias (μ_M y μ_{NM}) pero con la misma matriz de varianzas (Σ). Es decir,

$$f_C(x) = \frac{exp\left(\frac{-1}{2}(x-\mu_C)'\Sigma^{-1}(x-\mu_C)\right)}{(2\pi)^{3/2}\left|\Sigma\right|^{1/2}} \tag{3}$$
$$\text{con } C \in \{M, NM\}$$

Los parámetros del modelo (μ_M, μ_{NM} y Σ) se estimaron como:

$$\hat{\mu}_C = \frac{1}{n_C}\sum_{i=1}^{n_C} x_{Ci} \tag{4}$$

$$\hat{\Sigma} = \frac{n_M - 1}{n-2}\hat{\Sigma}_M + \frac{n_{NM}-1}{n-2}\hat{\Sigma}_{NM}$$

$$\hat{\Sigma}_C = \frac{1}{n_C - 1}\sum_{i=1}^{n_C}(x_iC - \hat{\mu}_C)(x_iC - \hat{\mu}_C)' \tag{5}$$
$$\text{con } C \in \{M, NM\}$$

donde x_{Mi} (x_{NMi}) denota la observación i-ésima perteneciente a la clase Meteorológica (No Meteorológica) de la muestra de entrenamiento, mientas que n_M (n_{NM}) denota el número total de observaciones de esa clase.

3.1.2. Modelo Cuadrático Discriminante (QDA)

En el método anterior se supuso que las observaciones dentro de cada una de las clases se distribuían según una ley Gaussiana, con distintas medias pero con la misma matriz de varianzas y covarianzas. Es posible ahora "relajar" este supuesto considerando el caso en que la matriz de varianzas y covarianzas sea diferente para cada clase. El razonamiento es el mismo que antes, simplemente que ahora será necesario estimar por separado la matriz de varianzas y covarianzas de cada clase, además de cada una de las medias. Es decir, en este nuevo escenario los parámetros desconocidos que debemos estimar son: un vector de medias (μ_M y μ_{NM}) y una matriz de varianzas (Σ_M y Σ_{NM}) para cada clase.

Las estimaciones para los vectores de medias son las mismas que para el caso anterior (4), se distinguen estimaciones diferentes de las matrices de covarianzas (5).

Uno se podría preguntar por qué se elegiría el método lineal si es posible utilizar un método más flexible como el cuadrático. La respuesta viene de la mano de la relación sesgo-varianza presente en cada uno de los métodos: Si bien se tiene que el clasificador cuadrático es más flexible, a la vez necesita de la estimación de una mayor cantidad de parámetros que el clasificador lineal. Por otro lado, el modelo lineal si bien es menos flexible, es más estable en lo que refiere a cambios en el resultado por modificaciones en la muestra de entrenamiento.

Para obtener más información sobre las técnicas de clasificación dadas por el análisis del discriminante (a) y (b), se puede consultar en Peña (2002).

3.1.3. Modelo Bayes Naive (BN)

Este modelo busca estimar las funciones de densidad $f_M(x)$ y $f_{NM}(x)$ con aún menos suposiciones, partiendo únicamente de la muestra de entrenamiento. Se propuso entonces utilizar lo que se denomina Estimador Núcleo de la Densidad (en este caso núcleo gaussiano):

$$f_C(x) = \frac{\sum_{i=1}^{n_C} exp\left(\frac{-(x-\mu_C)'(x-\mu_C)}{2h^2}\right)}{n_C h^3 (2\pi)^{3/2}} \quad (6)$$
$$\text{con } C \in \{M, NM\}$$

En donde x es un vector de tres coordenadas (correspondientes a V, ρ_{HV} y $SZDR$) y h el parámetro de suavizado. La elección del parámetro de suavizado es esencial para el buen funcionamiento del estimador. Como h controla la concentración de peso alrededor de cada punto de la muestra de entrenamiento, se tiene que valores chicos de h darán lugar a que únicamente las observaciones más cercanas al punto donde se quiere estimar la función de densidad sean relevantes en la estimación. De lo contrario, al tomar valores grandes de h, observaciones más lejanas influirán también en la estimación. Se han desarrollado varias técnicas para la selección del parámetro de suavizado, que no sólo pueden ser utilizadas para el caso del estimador núcleo, sino que también pueden ser fácilmente aplicables para otros estimadores. Para este trabajo se tomó como h el dado por la Regla de Referencia a la normal. Para más información sobre esta técnica y sobre este método de clasificación en general se puede consultar Silverman (1986) y Delicado (2008).

En general, el estimador tal como está en la expresión (6) no resulta práctico de aplicar, debido al alto nivel de cómputo que este implica, sumado a que cuando la cantidad de variables explicativas del modelo aumenta, la precisión del estimador disminuye. Como el objetivo final está en la clasificación, conocer de forma precisa la densidad de los datos resulta innecesario. Una forma de eludir este problema es construir un estimador bajo la hipótesis de que las variables (en este caso: ρ_{HV}, V y $SZDR$) son independientes. Se consideró entonces como estimador:

$$\hat{f}_C(x) = \prod_{k=1}^{3} \hat{f}_{Ck}(x_k) \quad (7)$$
$$k \in \{\rho_{HV}, V, SZDR\}$$

donde ahora $\hat{f}_{Ck}(x)$ denota el estimador no paramétrico de la densidad de la variable k en la clase C, y se calcula como en (6), pero utilizando solamente la variable x_k. Este método se lo denomina “Método de Bayes Naive”. Si bien el mismo parte de una suposición que no suele ser correcta en la mayoría de los casos, no sólo simplifica de forma evidente el estimador sino que también en la práctica da muy buenos resultados.

Hasta aquí los tres modelos presentados modelan y estiman las funciones de densidad de cada una de las clases. Una vez obtenidos los valores de los estimadores, al suponer que las probabilidades a priori son iguales para ambas clases, la regla de clasificación para un cierto punto (x) estará

dada por:

$$r(x) = \begin{cases} M \text{ si } \hat{f}_M(x) \geq \hat{f}_{NM}(x) \\ NM \text{ sino} \end{cases} \quad (8)$$

3.1.4. Modelo de Regresión Logística (LG)

Este método sugiere modelar $P(M|x)$ y $P(NM|x)$, para luego clasificar según (1). El modelo de regresión logística propone considerar:

$$P(C|x) = \frac{1}{1 + e^{-\beta_{0C} + \beta_{1C}x}} \quad (9)$$

En donde β_{0C} es un parámetro univariado, $\beta_{1C} = (\beta_{1Cv}, \beta_{1C\rho_{HV}}, \beta_{1Cszdr})$ es un parámetro de tres coordenadas y $C = M$ ó NM. Supongamos que $C = M$, entonces $P(NM|x) = 1 - P(C|x)$, análogamente se definiría $P(M|x)$, si se considerara $C = NM$.

Fue entonces necesaria la estimación de dichos parámetros, para la cual se utilizó el método de máxima verosimilitud. La idea esencial es la de estimar los parámetros desconocidos de forma tal que para todos los elementos de la muestra de entrenamiento la probabilidad estimada de pertenecer a la verdadera clase de procedencia (entiéndase la clase M o la clase NM) sea, los más posible, cercana a uno. Obtener este valor, es decir obtener el valor que maximiza la verosimilitud no resulta ser una tarea trivial. De todas formas, fue posible aproximar estos estimadores a partir de métodos iterativos muchos de los cuales han sido implementados en varios paquetes estadísticos como R (`http://cran.r-project.org/`). Más información sobre el modelo de regresión logística se puede encontrar en Peña (2002) y Hastie y Tibshirani (2009), entre otros.

3.2 TECNICAS DE REMUESTREO

Las técnicas de remuestreo son herramientas indispensables en la estadística moderna. Las mismas consisten en tomar repetidas muestras sobre los datos de entrenamiento para ajustar nuevamente los modelos de interés, obteniendo así información adicional sobre los mismos. En el caso de este trabajo, las muestras para entrenar y testear los modelos requirieron de una clasificación manual de las imágenes y por ello resultó dificultoso contar con un volumen de datos considerable para trabajar. Si bien los resultados obtenidos en las imágenes testeadas fueron buenos, se consideró necesario hacer un análisis estadístico más profundo de las performance de los distintos métodos aplicados. Para ello fue necesario plantear una manera de darle variabilidad a los datos. Luego de evaluar varias posibilidades se decidió aplicar técnicas de remuestreo sobre los pixeles de las imágenes de entrenamiento. Todos los clasificadores construidos a partir de los diferentes métodos y entrenados con distintas muestras (del remuestreo) fueron evaluados utilizando los datos test considerados en la Tabla I.

Para el análisis siguiente se supuso que la clasificación del experto fue perfecta. Con el objetivo de poder cuantificar la performance de los diferentes modelos, se utilizó como índice de bondad el valor de la proporción de puntos bien clasificados sobre el total de cada elevación (CSI, por sus siglas en inglés critical success index).

El procedimiento aplicado fue el siguiente: se llama N a la cantidad de pixeles totales de la muestra de entrenamiento, entonces:
1) Se seleccionaron de manera aleatoria y con reposición N puntos de la muestra de entrenamiento.
2) Tomando como muestra para entrenar la resultante del ítem anterior, se ajustaron los cuatro modelos (BN, LDA, QDA, LG).
3) Se clasificaron las 12 imágenes de los casos test.
4) Se aplicó el postprocesamiento.
5) Se calcularon los valores de CSI para cada caso test y para cada modelo.
6) Se volvió al punto 1.
Se computó la secuencia anterior 100 veces (en adelante, replicaciones).

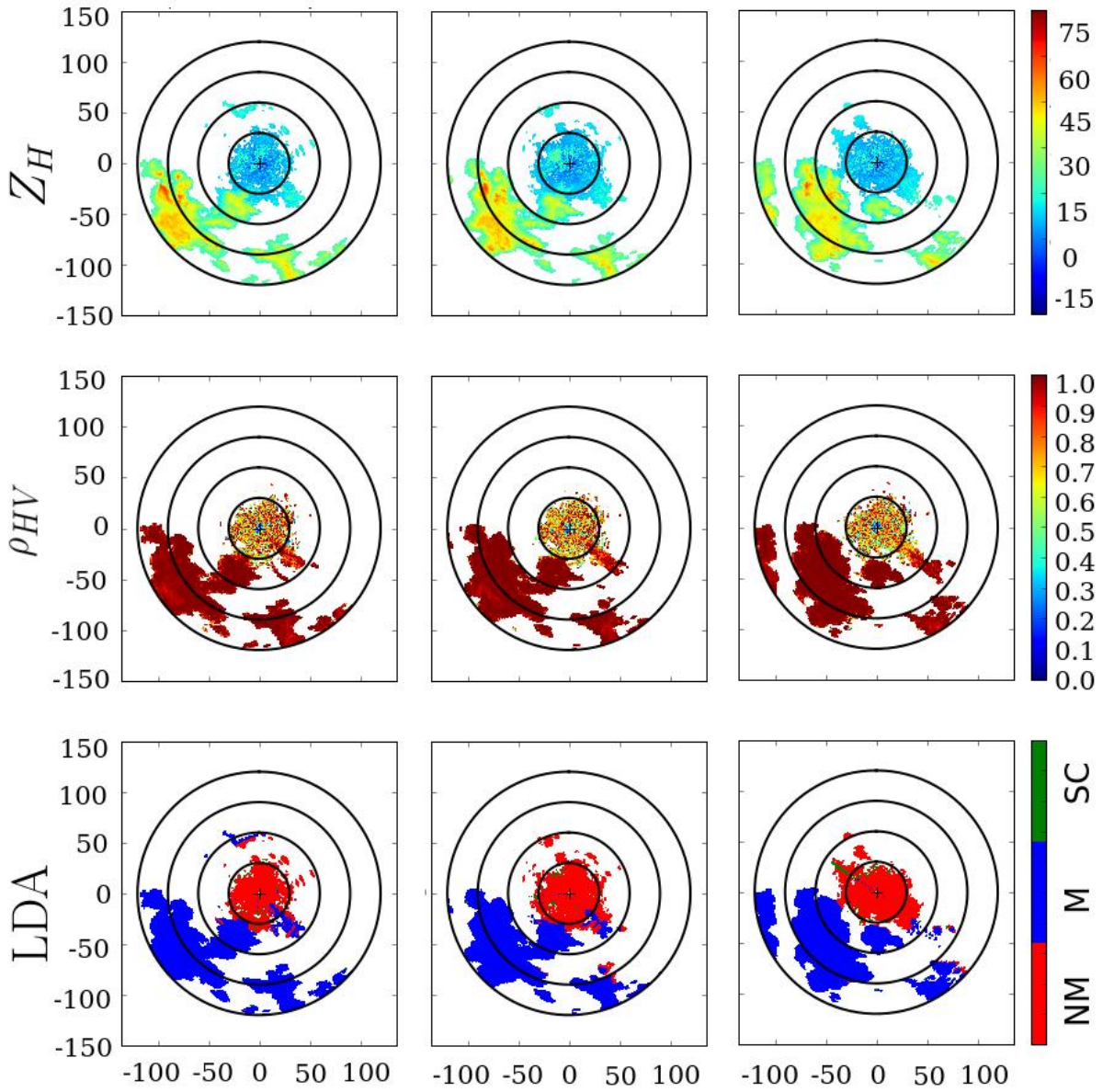

Figura 3: Resultados obtenidos para el 22 de noviembre de 2009 en el clasificador LDA. Por fila y sobre la primera elevación (0.5°): reflectividad horizontal ZH (dBZ), ρ_{HV}, resultado del clasificador: en azul los puntos clasificados como meteorológicos (M), en rojo los clasificados como no meteorológicos (NM), y en verde los considerados sin clase (SC). Los anillos concéntricos poseen una separación de 30km entre sí. Por columnas: tiempos sucesivos correspondientes a 08:53 UTC, 09:03 UTC y 09:23UTC.

4. RESULTADOS

Se implementaron los cuatro métodos descriptos en la sección anterior: LDA, QDA, BN y LG. Tal como fue descripto previamente, luego de la clasificación inicial sobre la muestra de testeo según cada una de las metodologías, se realizó un postprocesamiento de forma de incluir la estructura espacial de los datos, siguiendo la regla de la mayoría. Tal como se indicó en la Sección 2 todos los clasificadores fueron testeados en los cuatro eventos de la Tabla 1.

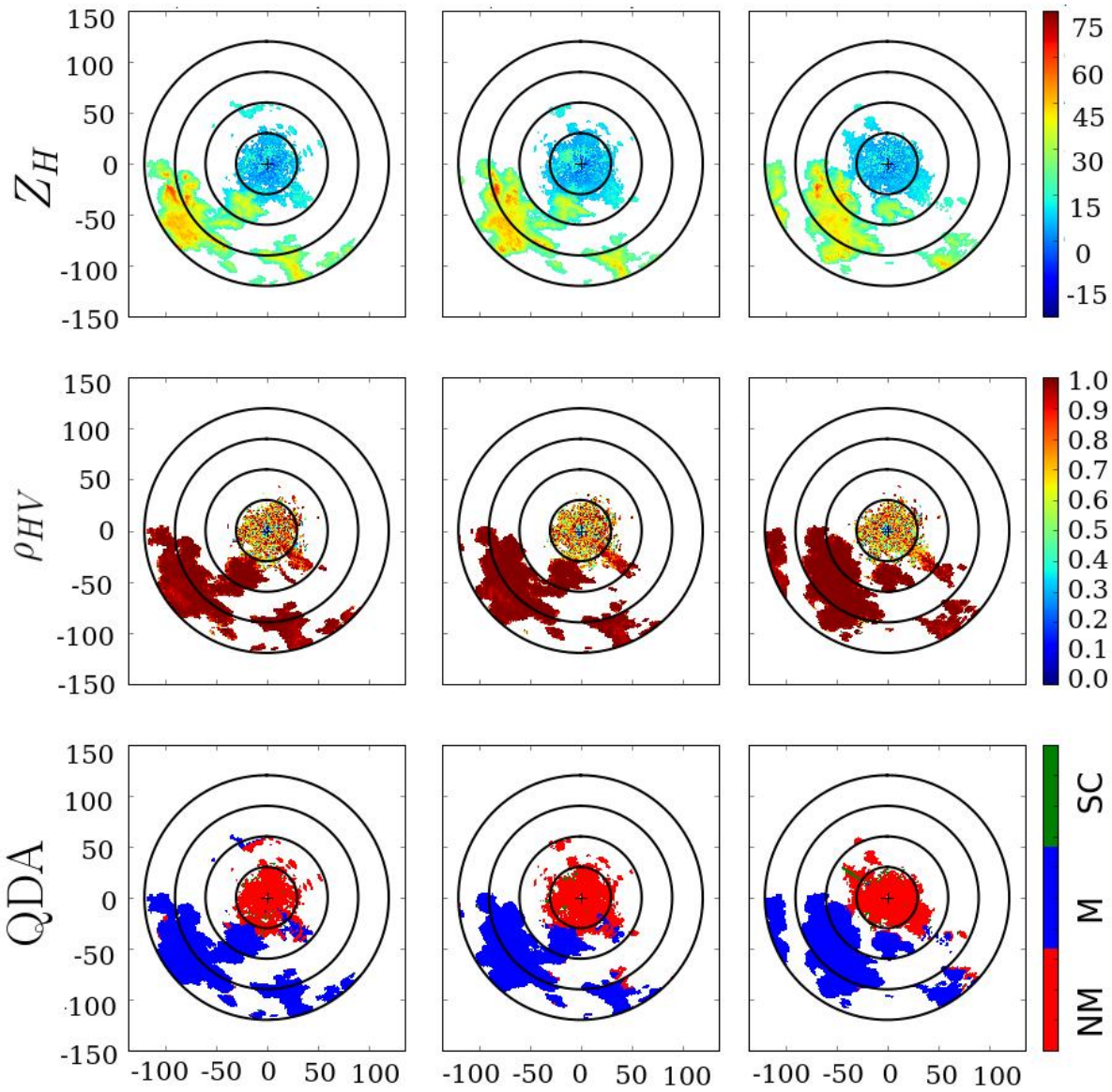

Figura 4: Ídem Figura 3 pero para clasificador QDA.

En las Figuras 3, 4, 5, y 6 se muestran los resultados de los cuatro modelos para el caso del 22 de Noviembre de 2009. Los resultados para el caso del 9 de Febrero de 2010 se encuentran representados en las Figuras 7, 8, 9 y 10. Si bien es posible observar que en ambos casos se dan situaciones en donde las dos clases (M y NM) se encuentran presentes, es claro que en el caso de 2009 los ecos están más mezclados, en comparación con la situación de 2010. Se puede observar que los resultados obtenidos con los diferentes métodos de clasificación coinciden en un alto porcentaje. De todas formas, es posible notar que las diferencias se encuentran en las zonas de frontera entre ambas clases, obteniendo discrepancias mayores en la situación de 2009. Esto último es de esperar ya que los puntos de frontera poseen cantidades similares de vecinos de las dos clases lo que afecta por un lado al postproceso, que considera los valores de estas cantidades, y por otro lado al comportamiento mismo de la variable SZDR en las zonas de frontera. En este caso el efecto se debe a que SZDR se construye a partir de la variabilidad

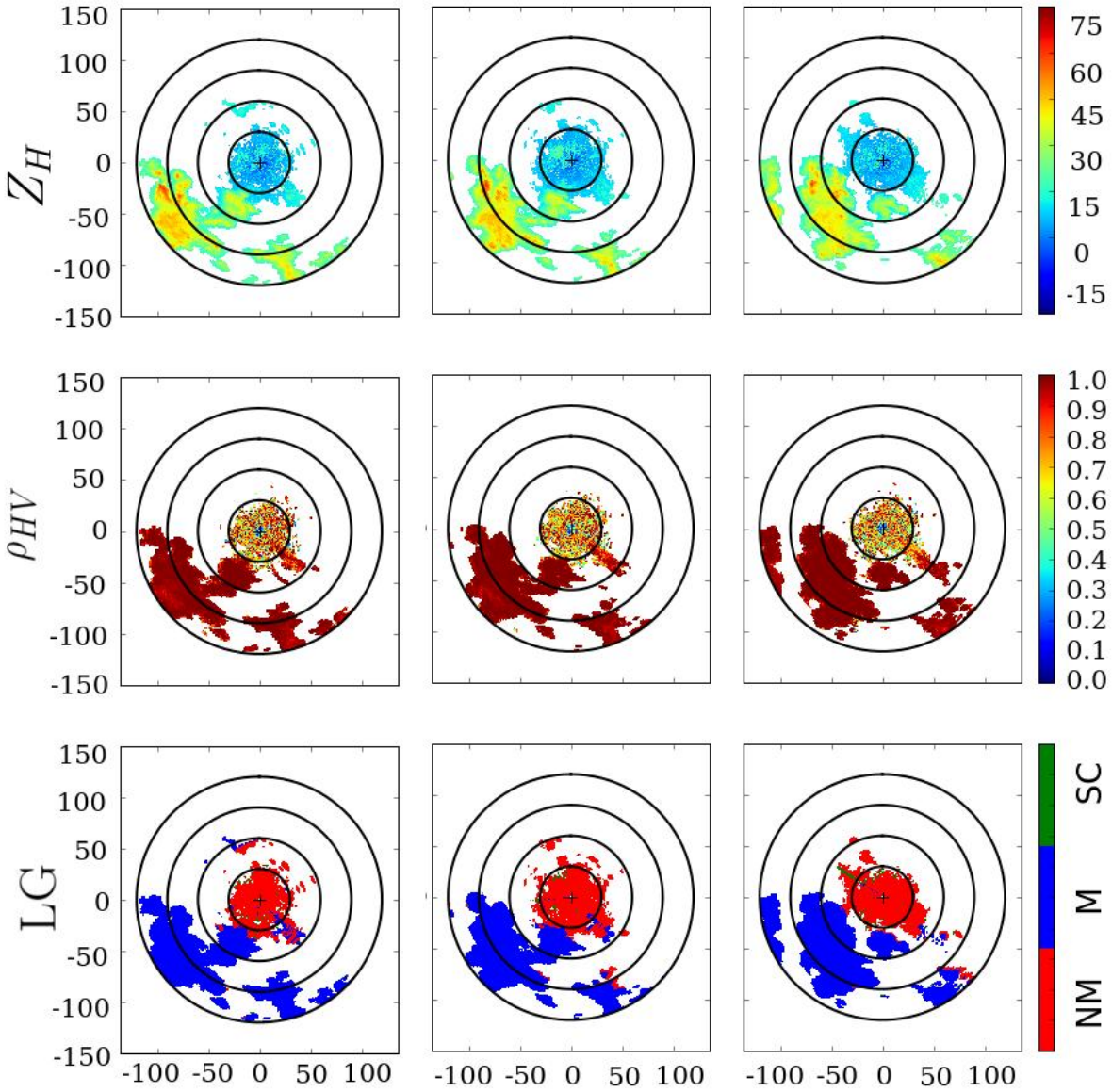

Figura 5: Ídem Figura 3 pero para clasificador LG.

espacial de Z_{DR}. Por consiguiente, sería posible que tanto el postproceso como el valor de SZDR en las zonas de frontera introdujeran errores al momento de la clasificación. De todas formas, la distribución espacial de los datos es sumamente importante al momento de clasificar puntos que no se encuentren en las zonas de frontera, ya que otorga una distinción importante en el comportamiento de las variables en las dos clases (ZDR que suele ser más "ruidosa" en presencia de ecos no meteorológicos) y a la vez permite capturar la continuidad de los ecos (postproceso). Resulta entonces evidente que los puntos de frontera son los que proponen el desafío mayor al momento de clasificar los datos en ecos meteorológicos o no meteorológicos.

Con respecto a la variable Velocidad Doppler, se observó que la misma no presenta demasiada relevancia a la hora de la clasificación. La velocidad permite en términos generales dar una idea de la naturaleza de los objetivos escaneados,

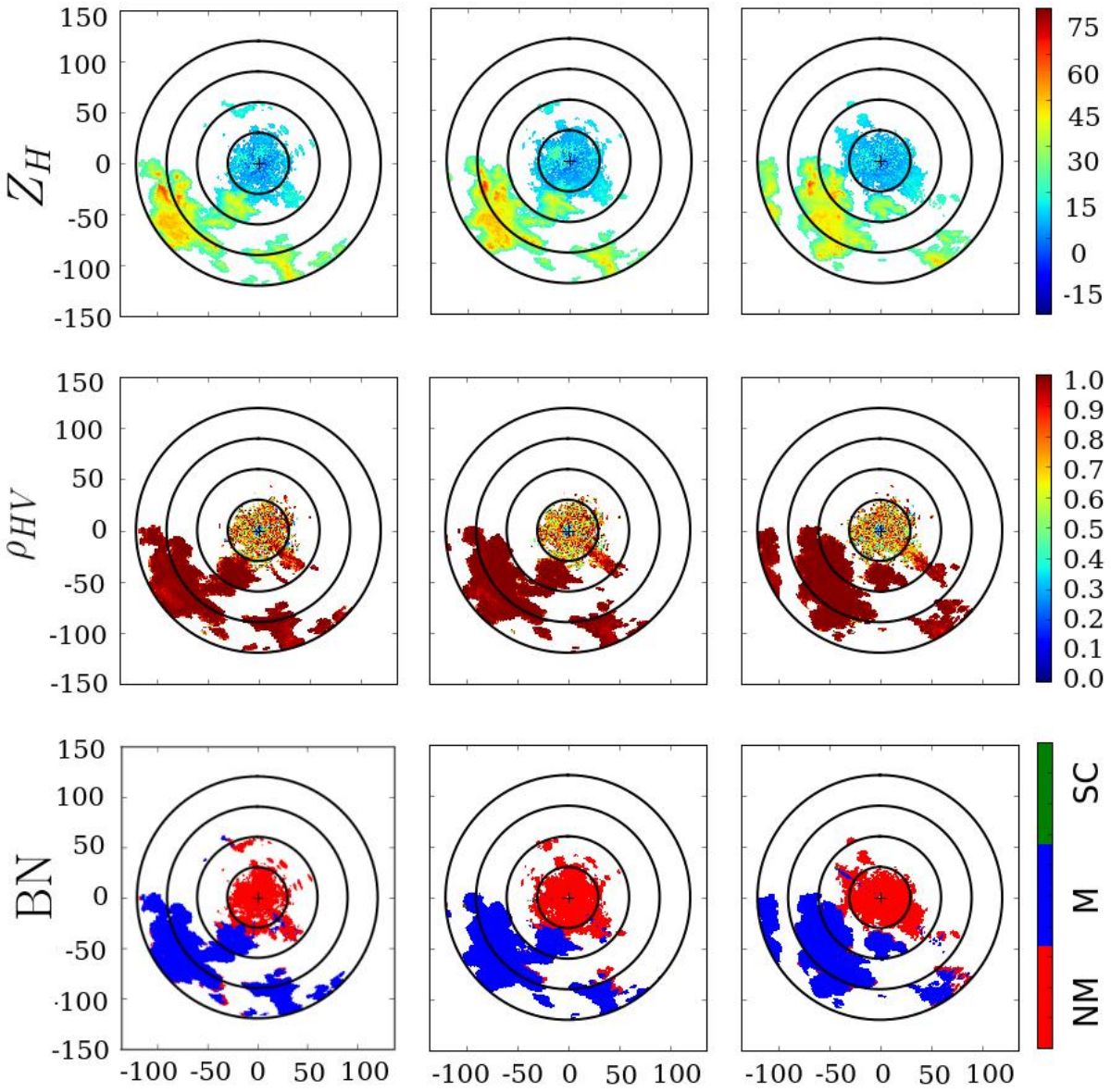

Figura 6: Ídem Figura 3 pero para clasificador BN.

pero puede llevar a conclusiones erróneas si se la observa por sí sola. Es decir, ecos producidos por el fenómeno de la propagación anómala poseen velocidades cercanas a cero, pero objetivos como ecos biológicos pueden poseer velocidades similares a las de un eco de precipitación. A su vez, puede suceder que objetivos meteorológicos posean velocidades bajas. Quizás, si el objetivo de la clasificación fuese lograr una discriminación más fina dentro de los ecos no meteorológicos (por ejemplo distinguir entre ecos biológicos, ecos de terreno y propagación anómala) la variable velocidad podría tener mayor relevancia para el análisis.

En la Tabla II se presentan los resultados obtenidos para la media, mediana, máximo, mínimo, primer cuartil, tercer cuartil y desvío del índice CSI correspondiente a cada uno de los cuatro métodos utilizados a lo largo de las 100 replicaciones del remuestreo. Es posible notar como los cuatro métodos contienen valores

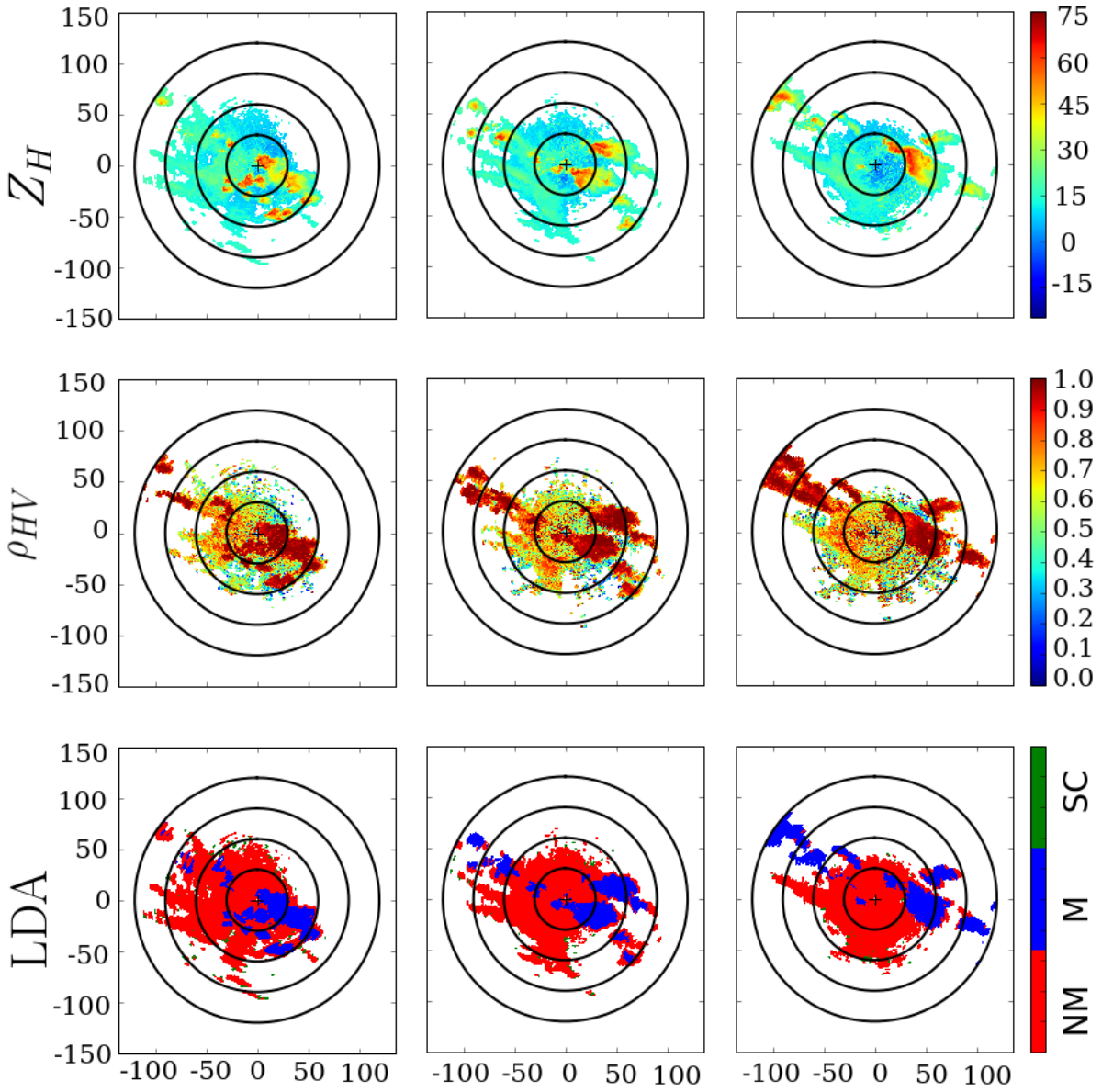

Figura 7: Resultados obtenidos para para el 9 de febrero de 2010 en el clasificador LDA. Por fila y sobre la primera elevación (0.5º): reflectividad horizontal ZH (dBZ), ρ_{HV}, resultado del clasificador: en azul los puntos clasificados como meteorológicos (M), en rojo los clasificados como no meteorológicos (NM), y en verde los considerados sin clase (SC). Los anillos concéntricos poseen una separación de 30km entre sí. Por columnas: tiempos sucesivos correspondientes a 21:03 UTC, 21:43 UTC y 22:33 UTC.

similares de mediana, pero no así de la media, lo que indica de alguna manera que hay diferencias en la dispersión de los datos.

En las Figuras 11 y 12 se muestran los histogramas y los boxplots correspondientes a los resultados de los valores de los CSI en cada uno de los cuatro métodos, basados en las 100 replicaciones. Como primera observación es posible notar que si bien los valores de

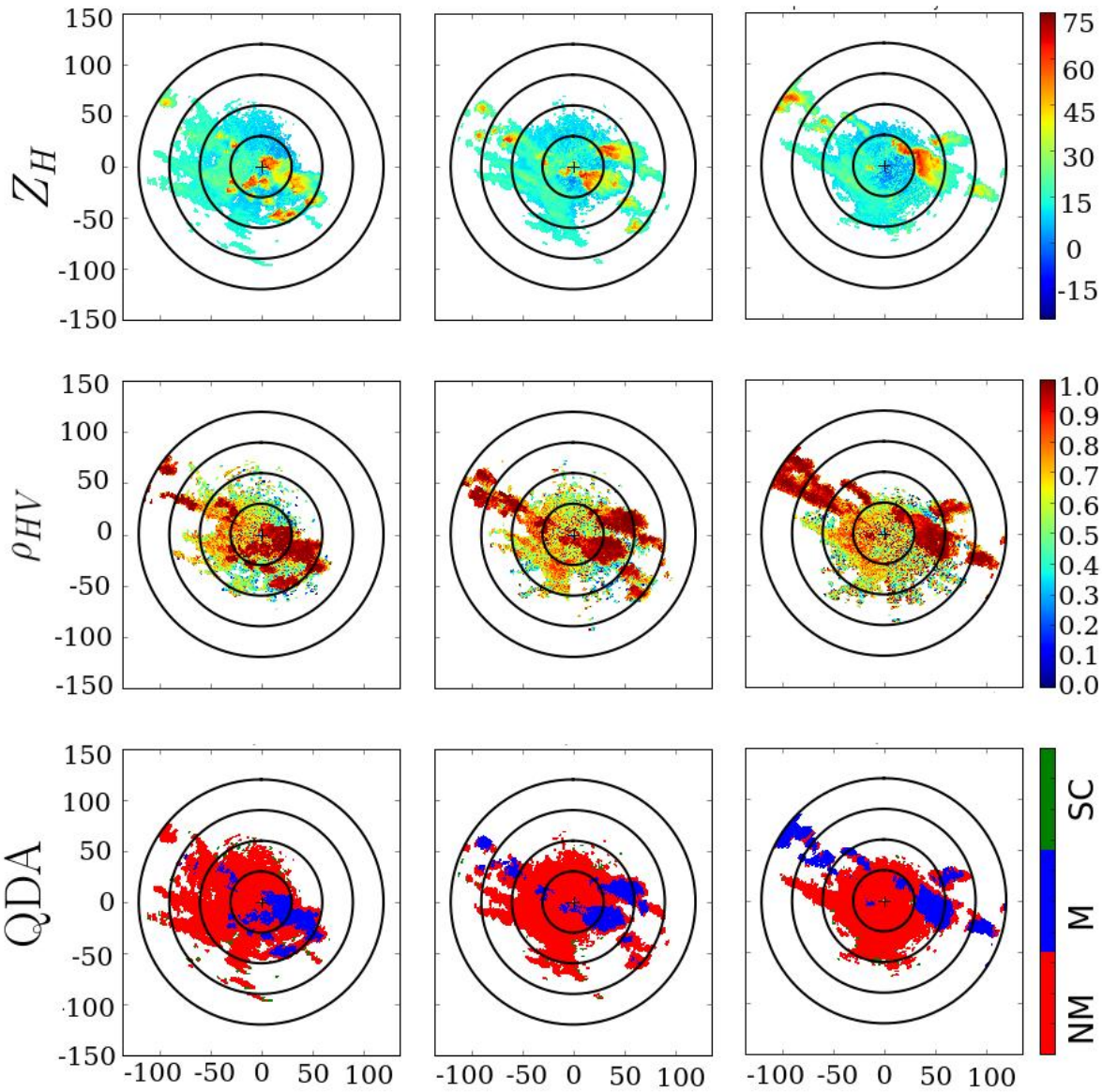

Figura 8: Ídem Figura 7 pero para el clasificado QDA

las medianas en los cuatro modelos son muy semejantes entre sí, el modelo con menor varianza es el modelo lineal. Además el modelo de Bayes Naive, si bien tiene mayor nivel de varianza, es el que logra valores más altos de CSI. Teniendo en cuenta la relación sesgo-varianza presente en estos modelos sería lógico esperar que el modelo más simple (LDA), es decir el modelo con menos parámetros que ajustar sea el que posea menor varianza y mayor sesgo. Es interesante notar cómo en este caso sucede lo primero pero no lo segundo, o por lo menos no es evidente. El modelo Bayes Naive no hace suposiciones sobre la forma de las funciones de densidad, otorgándole al método mayor flexibilidad. Si bien esto último en principio es algo positivo ya que permite obtener un modelo menos restrictivo, puede suceder que si el número de observaciones no es suficiente, se obtenga una sobre-estimación que lleve a errores

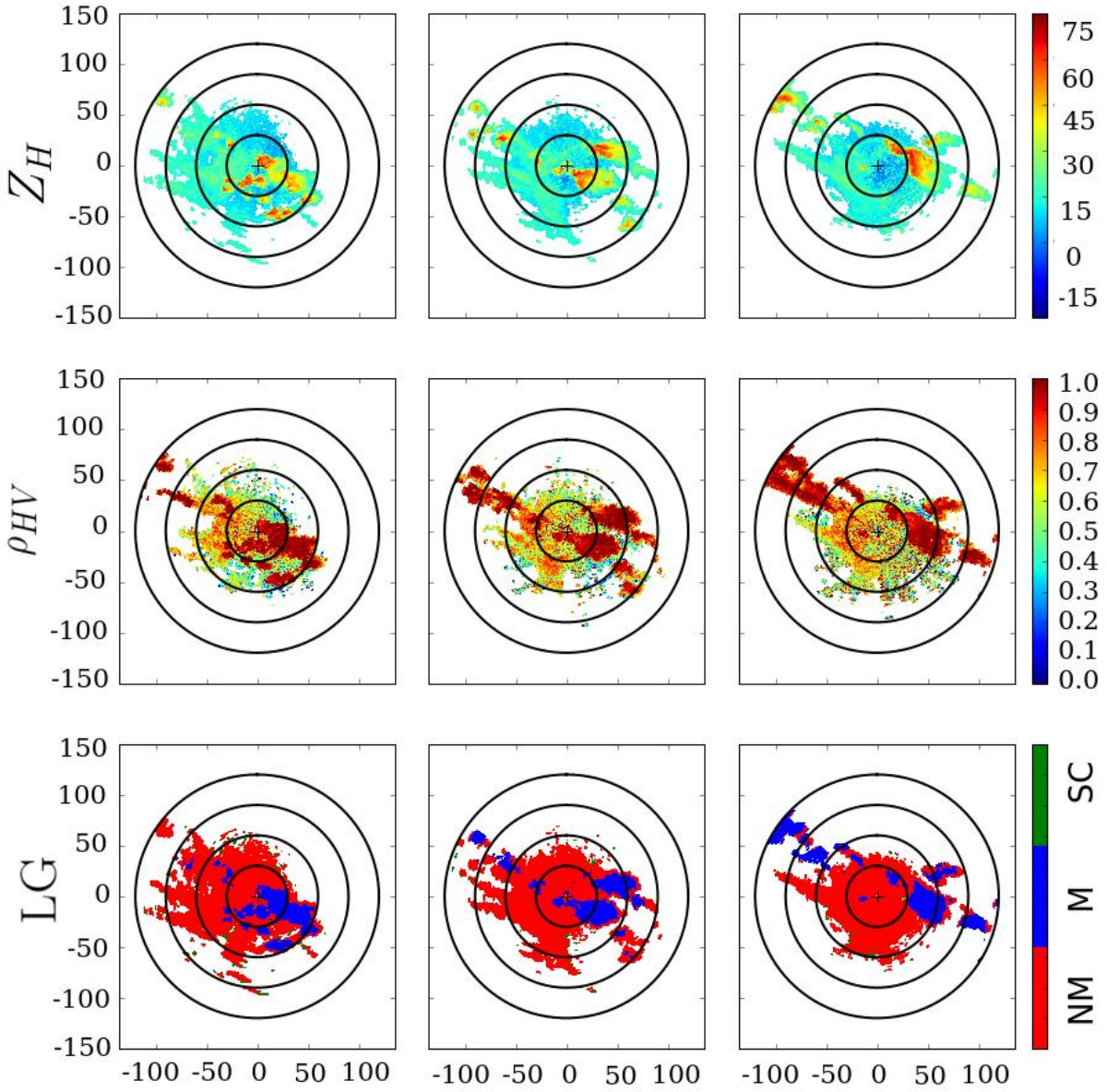

Figura 9: Ídem Figura 7 pero para el clasificado LG.

en las predicciones. Es posible entonces que la estimación dada por este modelo se encuentre bajo este escenario.

En general, elegir el nivel adecuado de flexibilidad para un modelo, no es una tarea sencilla, depende en parte del problema y de los datos con los que se cuente. Por consiguiente, a juzgar por los resultados obtenidos y en el caso de ser necesario elegir únicamente alguno de los cuatro modelos, en principio parecería que el lineal es el que mejor ajusta.

5. CONCLUSIONES

En este trabajo se presentaron cuatro técnicas de clasificación supervisada las cuales buscan dar una solución al problema de discriminación entre ecos meteorológicos y no meteorológicos en imágenes de radares meteorológicos muy importante al momento de realizar un control de calidad de la información para su uso en múltiples

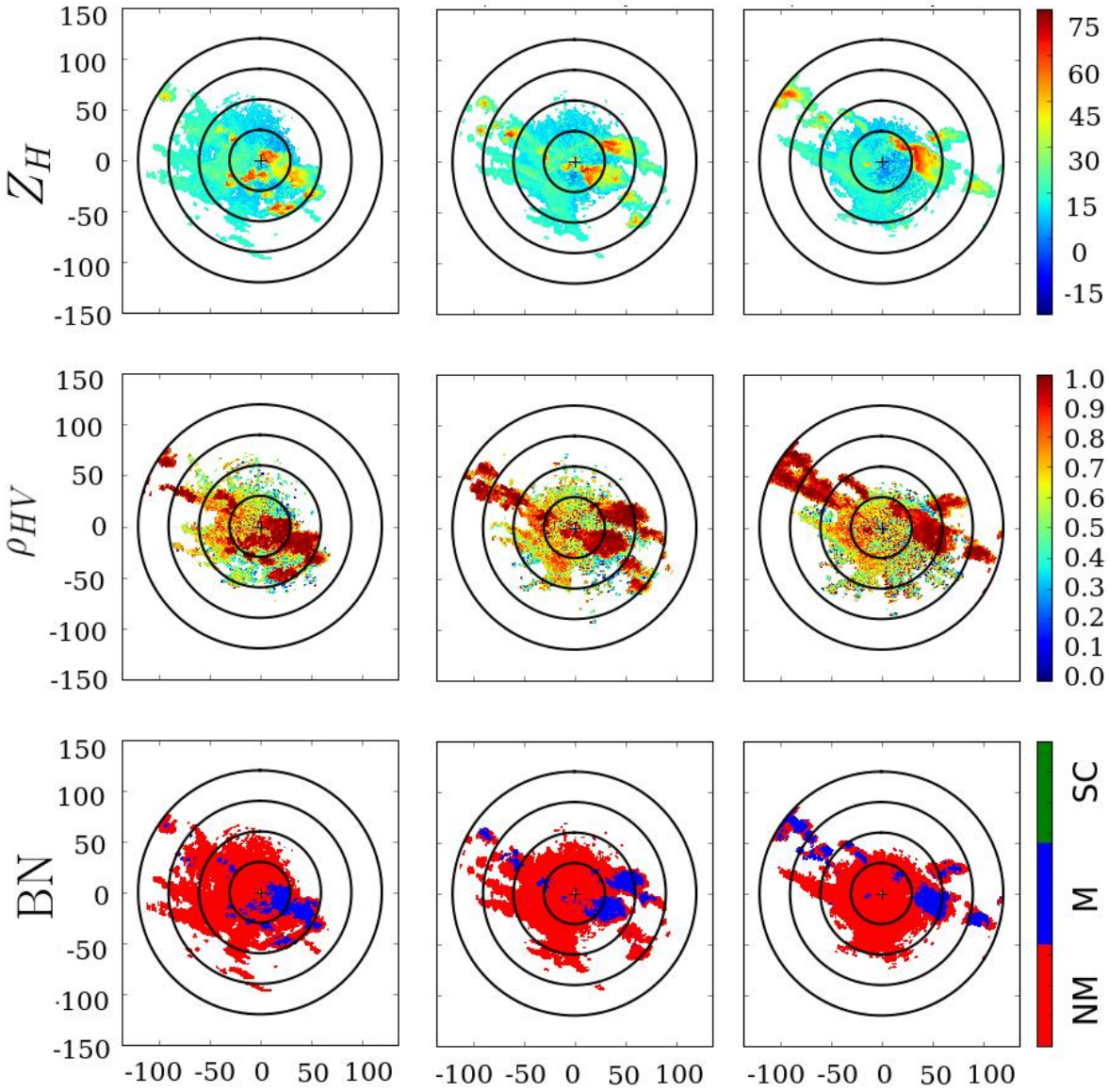

Figura 10: Ídem Figura 7 pero para el clasificado BN.

aplicaciones hidrometeorológicas. A partir de datos provenientes del radar Doppler banda C de doble polarización instalado en la Estación Experimental Agropecuaria INTA Anguil, se testearon cuatro modelos basados en las cuatro técnicas de clasificación presentadas con distintos niveles de flexibilidad. Mediante técnicas de remuestreo se estudiaron los desempeños de los mismos y se analizaron los resultados.

Las clasificaciones finales de los casos testeados en los cuatro métodos fueron coincidentes y correctas en un alto porcentaje (siendo las zonas de frontera las menos precisas). De todas formas, se vio que aplicar una metodología más flexible (BN) no incorporaba precisión en las predicciones, obteniendo mejores clasificaciones con los métodos más rígidos. Una posible razón puede ser que al no contar con un gran número de observaciones los ajustes más flexibles sobreestimen los datos, resultando entonces el modelo con mejor desempeño el Modelo Lineal.

No obstante, se podría suponer que si se contase

Modelo	Primer Cuartil	Mediana	Tercer Cuartil	Media	Desvío	Máximo	Mínimo
LDA	0,8912	0,9428	0,9606	0,9227	0,0577	0,9920	0,8080
QDA	0,8733	0,9490	0,9701	0,9177	0,0738	0,9932	0,7705
LG	0,8303	0,9494	0,9755	0,9134	0,0772	0,9924	0,7815
BN	0,8360	0,9470	0,9881	0,9040	0.0977	0,9907	0,7151

Tabla II: Resultados obtenidos a partir de las replicaciones. Los valores de la tabla son referidos al índice de éxito crítico (CSI) que se muestra en la Figura 11.

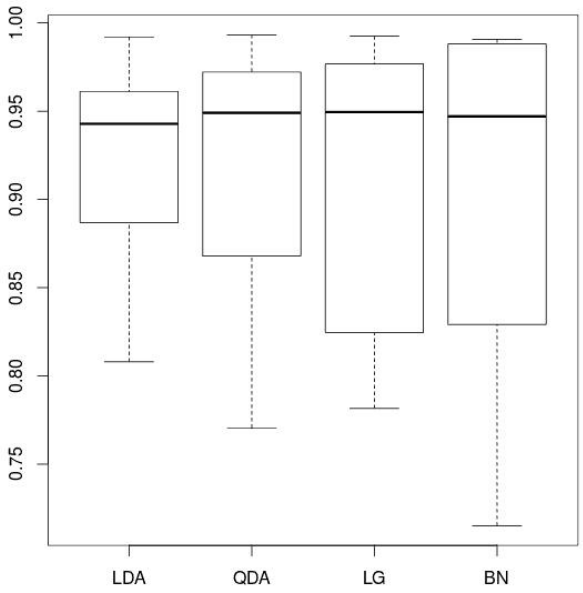

Figura 11: Boxplot de los valores de CSI obtenidos a partir de las 100 replicaciones en los cuatro modelos. De izquierda a derecha: LDA, QDA, LG y BN. La línea negra oscura marca el valor de la mediana en cada uno de los modelos. Los límites de las cajas están dados por el primer cuartil (25 %, límite inferior) y tercer cuartil (75 %, límite superior). Por último los bigotes de los extremos dan los valores del máximo (bigote superior) y del mínimo (bigote inferior).

con un set más amplio de datos para entrenar los modelos, la clasificación resultante para el caso del modelo BN podría resultar mejor. Obtener una muestra de entrenamiento mayor resultaría costoso, ya que la misma fue realizada de forma manual por el experto meteorólogo, lo que implica una mayor demanda de tiempo para poder incluir un número superior de casos. De todas formas es importante aclarar que por la manera en que está planteado este modelo, para cada punto a clasificar es necesario computar el estimador núcleo de la densidad con cada uno de los puntos de la muestra de entrenamiento (Ecuación 6), lo que implica que a medida que aumenta la muestra de entrenamiento, aumenta el tiempo de cómputo. Por esta razón, habría que analizar la factibilidad de éste método si se aumentase considerablemente la muestra de entrenamiento.

Si bien los resultados encontrados son muy alentadores, se cree necesario seguir avanzando en esta línea de investigación a fin de incorporar una mayor cantidad de casos y tener una mayor significancia de los resultados. Como primera medida, se cree que incorporar nuevas variables derivadas del radar al análisis ayudaría a obtener mejores resultados. Para ello será necesario trabajar en la calidad y calibración de las mismas sobre los radares que se encuentran en funcionamiento. En segunda instancia tal como se consideró la variable SZDR para incorporar nociones de la estructura espacial en sentido horizontal, se estima que incluir también variables que den nociones de la estructura espacial en el sentido vertical sería positivo. Por otra parte, como en este trabajo se utilizaron técnicas clásicas de clasificación se cree necesario explorar y evaluar algunas técnicas más avanzadas, como redes neuronales o técnicas de clusterización. Por último se espera poder extender estas técnicas a otros radares de doble polarización existentes en el país.

Agradecimientos: La realización del presente trabajo fue financiada por los proyectos PIDDEF N°5 2014-2017, PICT 2013-1299 y UBACyT 20020130100618BA.

REFERENCIAS

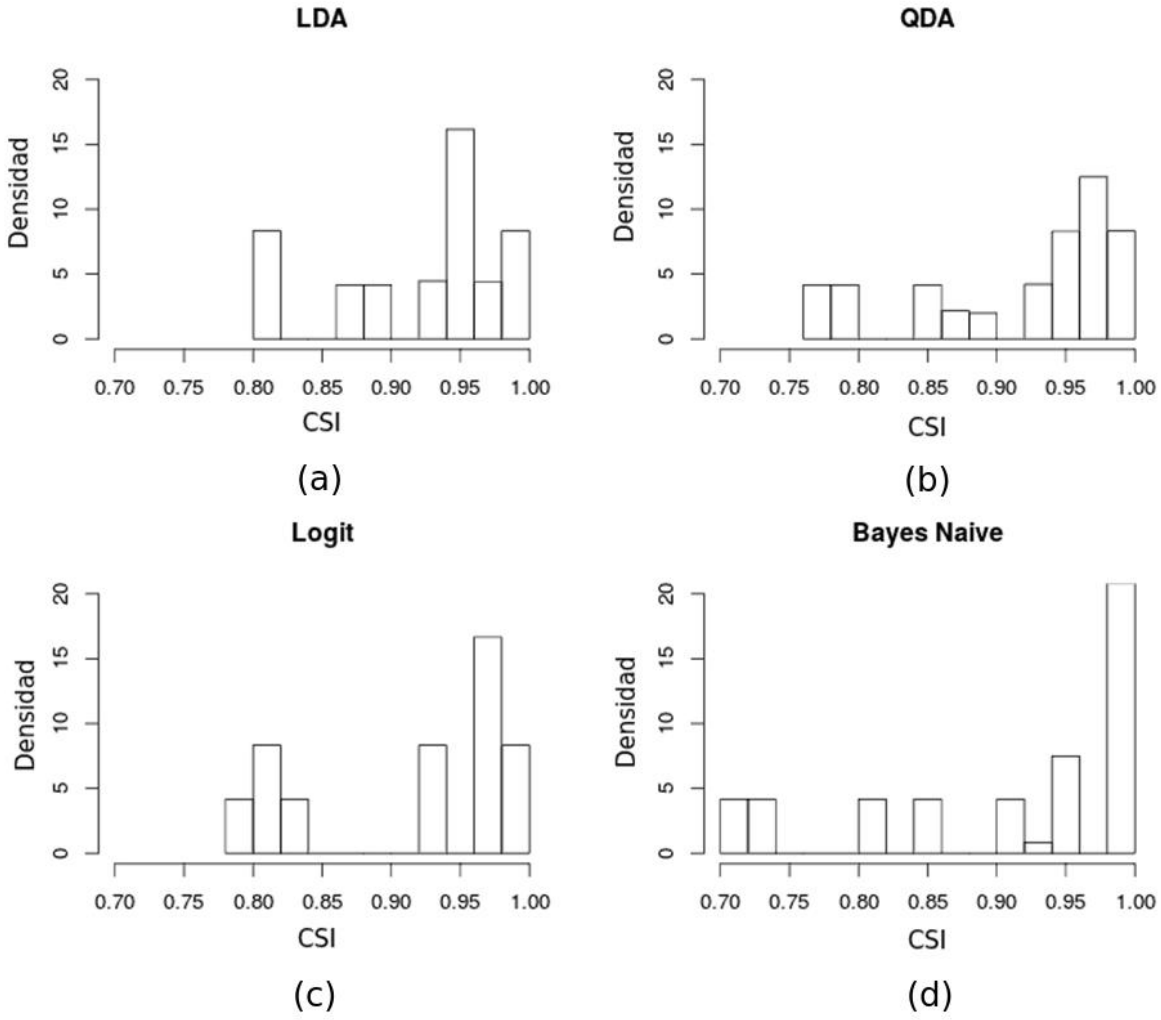

Figura 12: Histogramas de los valores de CSI obtenidos a partir de las 100 replicaciones en los cuatro modelos: (a) LDA, (b) QDA, (c) LG y (d) BN.

Battan L.J. 1973: Radar observation of the atmosphere. Q.J.R. Meteorol. Soc., 99: 793.

Berenguer, M., Corral C., Sanchez-Diezma R. y Sempere-Torres D., 2006: A fuzzy logic technique for identifying non precipitating echoes in radar scans. J. of Atmos. and Oceanic Tech., 23, 1157-1180.

Bo Young Y., Gyu Won L. y Hong-Mok P., 2015: Identification and Removal of Non-meteorological Echoes in Dual-polarization Radar Data Based on a Fuzzy Logic Algorithm. Advances in atmospheric sciences, 32, 1217–1230

ChoY. H., Lee. G, Kim K. E. y Zawadski I., 2006: Identification and removal of ground echoes and anomalous propagation using the characteristics of radar echoes. J. of Atmos. And Oceanic Tech., 23, 1206-1222.

Delicado P., 2008: Curso de Modelos no Paramétricos. Departamento de estadística e investigación operativa, Universidad de Cataluña. 192 págs.

Greku M. y Krajewski W. F., 2000: An Efficient Methodology for Detection of Anomalous Propagation Echoes in Radar Reflectivity Data Using Neural Networks. J. of Atmos. and Oceanic Tech., 17, 121-129.

Gourley J., Chatelet J. P. y Tabary P., 2006: A

Fuzzy Logic Algorithm for the Separation of Precipitating from Nonprecipitanting Echoes Using Polarimetric Radar Obsservations, J. of Atmos. and Oceanic Tech, 24, 1439-1451.

Hastie L., Friedman J. y Tibshirani R., 2009: The Elements of statical learning: Data Mining, Inference, and Prediction. Springer Text in Statistics. 747 págs.

Hubbert J. C., Dixon M. y Ellis S.M., 2009: Weather Radar Ground Clutter. Part II: Real-Time Identification and Filtering. J. of Atmos. and Oceanic Tech., 26, 1181-1197.

Kosko B., 1994: Neural Networks and fuzzy systems: a dynamical systems approach to machine intelligence. Pretince-Hall International Editions. 449 págs.

Lakshmananetal V., Zhang J., y Howard K., 2010: A Technique to Censor Biological Echoes in Radar Reflectivity Data. J. of applied meteorology and climatology, 49, 453-462.

Moszkowicz S., Cian G. J. y Krajewskiw F., 1993: Statistical Detection of Anamalous Propagation in Radar Reflectivity Patterns. J. of Atmos. and Oceanic Tech, 11, 1026-1034.

Peña D., 2002: Análisis de datos Multivariantes. S.A. Mcgraw-hill / Interamericanade España. 515 págs.

Rico-Ramirez M. A. y Cluckie I. D., 2008: Classification of ground clutter and anomalous propagation using dual-polarization weather radar. IEEE Transactions on Geosciences and Remote Sensing, 46, 7, 1892-1904.

Ryzhkov A. y Zrnic D. S., 1998: Discrimination between Rain and Snow with a Polarimetric Radar. J. of applied meteorology. 37, 1228-1240.

Schuur T., Heinselman P. y Ryzhkov A., 2003: Observations and classification of echoes with the polarimetric WSR-88D radar. National Severe Storms Laboratory (NOAA) and Cooperative Institute for Mesoscale Meteorological Studies (University of Oklahoma).

Silverman B. W., 1986: Density estimation for statistics and data analysis. Chapman and Hall/CRC.176 págs.

Siggia A. D. y Passarelli R. E., 2004: Gaussian model adaptive processing (GMAP) for improved ground clutter cancellation and moment calculation. Third European Conference on Radar Meteorology (ERAD) 67–73.

Steiner M. y Smith J. A., 2001: Use of Three-Dimensional Reflectivity Structure for Automated Detection and Removal of Nonprecipitating Echoes in Radar Data. J. of Atmos. and Oceanic Tech., 19, 673-685.

Zawadzki, I., 1984: Factors affecting the precision of radar measurement of rain, in 22 Conference on radar meteorology, edited by AMS, pp. 251-256, Zurich, Switzerland.