Maestría en Explotación de Datos y Gestión del Conocimiento

Estimación de ocurrencia de granizo en superficie y daño en cultivos,

mediante datos del radar meteorológico utilizando técnicas de **Data Mining.**

Tesis de Posgrado

Yanina Noemí Bellini Saibene

Octubre de 2015

Instituto Nacional de Tecnología Agropecuaria

UNIVERSIDAD AUSTRAL

Tesis presentada para optar por el grado de *Magíster en Explotación de Datos y Gestión del Conocimiento.*
Universidad Austral

Autor:
Lic. Sist. Yanina Noemí **Bellini Saibene**

Director de Tesis:
Msc. Martín **Volpacchio**

Jurado de la Tesis:

Dr. Alejandro de la Torre
Profesor Titular y Director de Investigaciones en Universidad Austral, Facultad de Ingeniería.
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

PhD. María Daniela López De Luise
CI2S lab director
IEEE Argentina CIS
IEEE Argentina Nominations & Appointments
IEEE CIS-WCI subcommittee 2015
LA-CCI Steering Committee
Facultad de Ingeniería. Universidad Austral

Ph.D., Pablo Mercuri
Director Centro de Investigación de Recursos Naturales (CIRN)
Instituto Nacional de Tecnología Agropecuaria (INTA)

Director de la Maestría:
Dr. Juan M. Ale

19 de Octubre de 2015

"El viento me confió cosas
que siempre llevo conmigo,
me dijo que recordaba
un barrilete y tres niños,
que el sauce estaba muy débil,
que en realidad él no quiso,
que fue uno de esos días
que todo es un estropicio…".

Confesiones del viento
(Roberto Yacomuzzi- Juan Falú).

Para **Ana**, quien se llevó mi corazón y para **Juan Bautista** y **Francesco**, que me lo devolvieron.

Agradecimientos

A **Juan Marcelo Caldera**, mi compañero en esta vida, por estar SIEMPRE y al resto de mi familia por todo el apoyo durante este proceso.

Al **INTA** y sus autoridades por la confianza depositada al incluirme en el programa de perfeccionamiento que me permitió realizar esta maestría.

A **Romina Mezher, Santiago Banchero, Héctor Lorda, Julio Fernández, Guillermina Fernández, Mariana Sozzi, Juan Paulini, Araceli Hernández, Andrés Sipowicz, Jonathan Faerman, Javier Breccia, Alejandra Bellini, Laura Belmonte, "Mariela" Fuentes** y **Pablo Lucchetti** por las explicaciones, lecturas, correcciones, aportes y consejos durante la realización de esta tesis.

A las compañías de seguro: **SanCor, La Segunda y La Dulce** por los datos de campo.

A los **productores, contratistas y profesionales** del INTA y de la actividad privada, en especial a **José** y **Marcelo Caldera,** que brindaron información y acceso a sus establecimientos agropecuarios para la recolección de los datos de campo.

A mi director de tesis **Martín Volpacchio** por su tiempo, guía y consejos en el momento oportuno.

A todos los que de alguna manera me asistieron durante la concreción de este proyecto.

Índice

Resumen

El granizo es capaz de infligir cuantiosos daños y el estudio de su frecuencia e impacto económico es de interés para la industria de los seguros y el sector agroindustrial. Con el objetivo de estimar la ocurrencia de granizo en superficie y el posible daño ocasionado a los cultivos se utilizó Gene Expression Programming (GEP) Regresión Logística usando datos del radar polarimétrico de banda C (INTA Anguil, La Pampa) desde Marzo de 2009 a Marzo de 2013. La complejidad en la captura de los datos implicó el desarrollo de un software específico para procesar los datos del radar, gestionar la información de reportes de ocurrencia y daño por granizo y unificar ambos tipos de datos. La ocurrencia de granizo en superficie se modeló como un problema binario, utilizando solo variables derivadas del radar. A pesar de la simplicidad del modelo obtenido, una comparación entre medidas de performance (Probabilidad de Detección (POD), Falsas Alarmas (FAR) y Porcentaje Correcto (PC)) entre 30 modelos internacionales publicados, que usan diferentes técnicas y set de datos, ubicó a nuestro modelo entre los tres primeros. Confirmando que los modelos con variables polarimétricas funcionan mejor que aquellos que usan variables de simple polarización y que la presencia de granizo aumenta con mayores valores de Reflectividad (Z), menores valores del Coeficiente de correlación co-polar (Rho_{HV}) y valores extremos de Reflectividad Diferencial (Z_{DR}). La implementación del modelo estableció problemas con el uso de técnicas tradicionales de manejo de datos, alentando enfoques adicionales como bases de datos no estructuradas y técnicas de procesamiento paralelos para la operación de los mismos. Para determinar el daño se usaron variables de radar y de cultivo para clasificarlo en cuatro problemas binarios de acuerdo a los porcentajes de destrucción: leve (1-25%)/no leve, moderado (25-50%)/no moderado, severo (50-75%)/no severo y grave (75-100%)/no grave. Los modelos obtenidos no son robustos en diferenciar estas clases debido a: la simplificación del problema, la baja disponibilidad de datos y la subespecificación de las variables de alto impacto. A pesar de esto, la correlación encontrada, sugiere que estas herramientas se pueden usar para análisis futuros en un conjunto de datos más grande y completo. Análisis adicionales reduciendo los niveles a tres: sin daño, < 50% y >50% aumentan la correlación, reforzando la idea que estas herramientas son adecuadas para generar modelos sobre el daño en cultivos.

Abstract

Hail is capable of inflicting considerable damage and the study of their frequency and financial impact is useful for the insurance industry and agribusiness. In order to calculate the probability of hail on the ground and possible crops damage Gene Expression Programming (GEP) Logistic Regression was used with data from polarimetric C-band radar (INTA Anguil La Pampa) from March 2009 to March 2013. The complexity of data capture involved the development of a specific software to process radar data, manage information from occurrence reports and hail damage and unify both types of data. The hailfall was modeled as a binary problem, using only variables derived from radar. Despite the simplicity of the model obtained, a comparison between the measurements of performance Probability of Detection (POD), False Alarm Ratio (FAR) and Percent Correct (PC) among 30 reported international models, using different techniques and data sets, showed our analysis within the three models with higher fitting. Confirming that models with polarimetric variables showed higher performance than those with single polarization and the presence of hail increase with higher of Reflectivity (Z) values, lower *Correlation Coefficient* (Rho_{HV}) values and extreme values of *Differential Reflectivity* (Z_{DR}). The implementation of the model using traditional data management techniques established problems, encouraging additional approaches like unstructured databases and parallel processing techniques for radar data processing and model operation. The radar and crop variables were used to classify into four binary problems according the percentages of destruction: slight (1-25%)/no slight, moderate (25-50%)/no moderate, severe (50-75%)/no severe and grave (75-100%)/no grave. Although the models fail to differentiate these four classes due to the simplification of the problem, the lower data availability and a under specification of the high impact variable, a correlation found suggested that these tools can be used for future analysis on larger and more complete dataset. Further analysis reducing the levels to three categories: without damage, <50% and >50% increased the correlation, reinforcing the idea that these tools are useful for generating models that classified crops losses.

Capítulo 1. Conceptos Generales y Antecedentes

1.1. Introducción

El Instituto Nacional de Tecnología Agropecuaria (INTA) cuenta con una red de tres radares meteorológicos cuyo objetivo se resume en mejorar el conocimiento del ambiente climático en el que se desarrollan las plantas y los animales. Las fluctuaciones del clima constituyen uno de los factores que mayor incertidumbre generan dentro de un ecosistema agropecuario, contribuyendo a la complejidad en el proceso de la toma de decisiones y condicionando la eficiencia del manejo empresarial agropecuario [1]. El granizo es uno de los fenómenos meteorológicos capaz de infligir daños cuantiosos al deteriorar seriamente cosechas, edificios, infraestructura y medios de transporte [2]. En la región Pampeana[1] Argentina, la máxima frecuencia de ocurrencia de granizo se encuentra en Córdoba y La Pampa [4] [5] [6] [7]. El estudio del granizo es de utilidad para: 1) los servicios meteorológicos nacionales y la aviación (alertas) [8] [9], 2) la industria de los seguros (estimación de daño, mitigación y transferencia, riesgo forense) [8] [9] [10] y 3) para la comunidad agropecuaria (protección, mitigación y transferencia del riesgo, caracterización de riesgos agropecuarios, riesgo forense, lucha antigranizo) [8] [9] [10]. Debido a la reducida extensión espacial y temporal de las tormentas de granizo, su detección y ocurrencia en superficie es una tarea difícil y costosa [11] [12] [13] [14] [15], por lo que sensores remotos como los radares meteorológicos son una alternativa a las redes terrestres de mediciones, con la ventaja de abarcar una gran superficie y disponer de una única resolución en tiempo y espacio [9].

Desde la década del 50 se han realizado numerosos estudios que exploran la relación de las variables medidas por los radares con el granizo y se han generados diferentes modelos con los objetivos de clasificar hidrometeoros[2], determinar su precipitación, tamaño o daño que ocasionaron; algunos de estos trabajos utilizan técnicas de minería de datos (Data Mining-DM) (ej.:[12],[17],[18],[19],[20],[21],[22],[23],[24]). Estos modelos deben ser ajustados a la región de estudio y a los instrumentos de medición utilizados [11] [14] [25] [26]. La mayor parte de la bibliografía presente para Argentina hace referencia a trabajos para radares de banda S, de simple polarización y localizados

[1] La región Pampeana está compuesta por las provincias de Buenos Aires, este de La Pampa y centro sur de Santa Fe, Córdoba y Entre Ríos [3].

[2] Un hidrometeoro o meteoro acuoso es un fenómeno que tiene lugar en la atmósfera y que consiste en una suspensión, precipitación o deposición de partículas acuosas. Este fenómeno puede ser de naturaleza óptica o eléctrica [16].

en la provincia de Mendoza (ej.: [11],[12],[14],[27],[28],[29]). La técnica de DM más utilizada en estos trabajos es la regresión logística (ej.: [12],[14]). Los estudios que usan radares de banda C, no utilizan técnicas de DM y analizan variables de simple polarización en Mendoza (ej.: [30]) y variables de doble polarización en las provincias de Buenos Aires, Entre Ríos y La Pampa (ej.:[31], [32], [33]).

Uno de los radares de INTA se localiza en la Estación Experimental Agropecuaria (EEA) Anguil, en La Pampa. Realiza mediciones cada diez minutos en un área de 240 km de radio generando un gran volumen de información diaria[3] [34]. El radar cuenta con un software provisto por el fabricante que implementa el algoritmo Z-based Hail Warning (ZHAIL) [34] para la estimación de probabilidad de granizo. ZHAIL se basa en [20] que fue desarrollado para radares de bandas X y S y de simple polarización [20] [34] [35] [36]; además de los datos del radar, necesita información de radio sondeos [34].

Las técnicas de DM permiten analizar grandes volúmenes de datos con el objetivo de encontrar relaciones impensadas y nuevas formas de resumir y presentar los datos que resulten útiles y compresibles para el usuario [37].

En este contexto, este trabajo intenta responder al interrogante: ¿Es posible utilizar las técnicas de DM para la generación de un modelo que relacione la caída de granizo y el daño que ocasiona en los cultivos, con las variables registradas por un radar meteorológico polarimétrico?.

Para ello el objetivo general de este trabajo es el **desarrollo de un modelo de estimación de ocurrencia de granizo en superficie y daño en cultivos ajustado a la región que comprende el este de La Pampa, sur de Córdoba y oeste de Buenos Aires, por medio de técnicas de DM, tomando como base los datos generados por el radar meteorológico de la EEA Anguil**.

Los objetivos específicos incluyen: 1) examinar la utilidad del uso de técnicas de minería de datos para el procesamiento de este tipo de información, 2) aprovechar las características de doble polarización del radar, 3) generar herramientas que permitan minimizar el uso de software propietario para el manejo de los datos, 4) contribuir a la caracterización del comportamiento de las variables de doble polarización de banda C, para la identificación de granizo y daño en cultivos en la región de estudio, 5) generar herramientas que faciliten el acceso a los datos del radar para futuros estudios y 6)

[3] Aproximadamente dos millones de archivos con la configuración actual [34].

contribuir al entendimiento del uso de radares meteorológicos para aplicaciones agropecuarias.

1.2. Conceptos Generales

1.2.1. Minería de datos

DM es la exploración y análisis por medios automáticos o semiautomáticos de grandes cantidades de datos con el fin de descubrir reglas y patrones significativos [38]. Es parte del proceso de descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases - KDD) que se define como la extracción no trivial, desde los datos, de información implícita, previamente desconocida y potencialmente útil [38].

La creciente disponibilidad de información meteorológica y climática (ej.: registros históricos, simulaciones de modelos numéricos, datos de radares meteorológicos y satélites, etc.) hace necesario el uso de nuevas técnicas más eficientes y automáticas que las clásicas herramientas estándar de modelización estadística [39], por lo que DM está emergiendo como un método adecuado para la extracción de patrones a partir de conjuntos amplios de datos heterogéneos relacionados con la predicción de fenómenos meteorológicos [39][40].

Estas herramientas deben funcionar bajo la restricción que los datos disponibles son observacionales (en contraposición a datos experimentales) debido que en la mayoría de los casos se trabaja con datos que han sido recolectados para un propósito diferente al análisis de DM. Esto implica que la estrategia de recolección no tuvo en cuenta el objetivo del análisis, a diferencia del análisis estadístico, donde los datos son recolectados utilizando estrategias eficientes para contestar preguntas específicas [37].

Una segunda restricción es que los conjuntos de datos analizados con DM son grandes e involucran problemas relacionados con su manejo (almacenamiento, mantenimiento, accesibilidad, visualización, etc.) y con su análisis (determinar la representatividad de los datos, analizar los datos en un tiempo razonable, decidir si un patrón encontrado realmente refleja la realidad subyacente, etc.) [37].

El método seguido en el proceso de DM es una mezcla de los métodos matemáticos y científicos, existen diversas caracterizaciones del mismo, una de las más difundida, y la utilizada en este trabajo, es Cross-Industry Standard Process for Data Mining (CRISP-DM) y consta de seis fases concebidas como un proceso cíclico (Figura 1) [41] [42]:

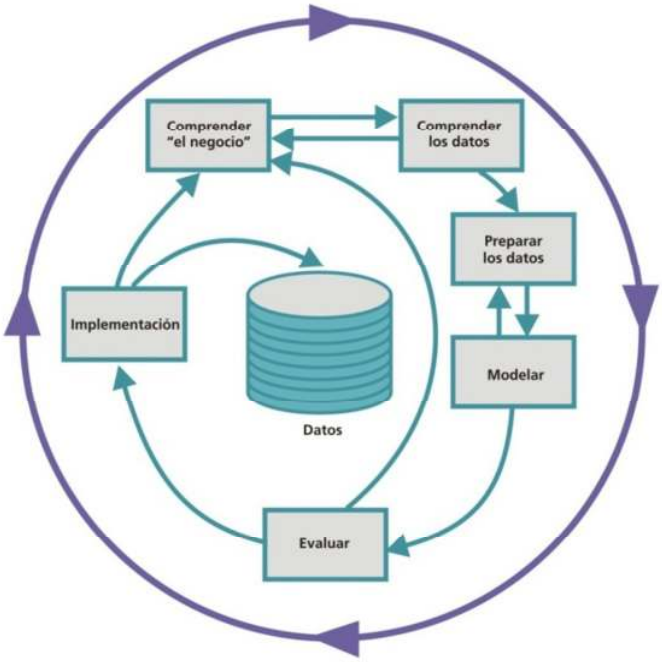


Figura 1. Fases del proceso de CRISP-DM (Adaptado de: [42])

1) *Comprender "el negocio":* el término "negocio" hace referencia al área de conocimiento en la cual se aplicará DM. Esta fase incluye determinar los objetivos, evaluar la situación actual, establecer tareas de DM, y desarrollar un plan de proyecto [40] [41] [42].
2) *Comprender los datos:* este paso incluye la recopilación, descripción y exploración de los datos. Se verifica la calidad de los mismos y se realizan los cálculos de las estadísticas básicas de los principales atributos y sus correlaciones. Se generan visualizaciones [40] [41] [42].
3) *Preparar los datos:* es la fase más compleja y la que más tiempo insume (60% a 70 % del tiempo total del proceso). Identificados los datos disponibles, es necesario seleccionar, limpiar y transformarlos para la fase de modelado. Incluye una exploración más profunda [40] [41] [42].
4) *Modelar:* es el núcleo del proceso de DM. Se selecciona el método más adecuado de acuerdo a los datos disponibles y los objetivos del análisis. Cada método tiene sus requisitos para los datos de entrada, por lo tanto un cambio de método puede conducir a la necesidad de repetir parte de la etapa de preparación de datos [40] [41] [42].
5) *Evaluar:* los resultados del modelo son evaluados en el contexto de los objetivos establecidos en la primera fase y teniendo en cuenta el costo de realizar una

clasificación incorrecta. La evaluación se basa en indicadores cuantitativos de la calidad de los modelos creados midiendo la tasa de error de los casos mal clasificados en un conjunto de datos con el cual el modelo trabaja por primera vez [40] [41] [42] [43].

6) *Implementación:* en esta fase se desarrolla un plan de implementación práctica para el modelo. También se determina la estrategia de supervisión y mantenimiento del mismo para asegurar su confiabilidad con el paso del tiempo [40] [41] [42].

Para el desarrollo de la etapa "Comprender el negocio" se presentan en forma somera los conceptos relacionados al granizo (ocurrencia y generación) y a los radares meteorológicos (funcionamiento).

1.2.2. El Granizo

El granizo es un fenómeno meteorológico que consiste en un tipo de precipitación formado por partículas de hielo (conglomerado policristalinos) de diámetro igual o superior a 5mm [26]. Generalmente son esféricos y su superficie externa puede ser irregular, con lóbulos y protuberancias (figura 2) [2]. Es considerado un riesgo agroclimático[4] y produce importantes pérdidas económicas en cultivos [26] [32] [44] [45] [46] [47] [48]. Más del 90% de los seguros agropecuarios en la Argentina son relativos a coberturas de granizo [49].

Figura 2. Diferentes tipos de granizo (Adaptado de:[50]).

Las tormentas que pueden generar granizo son del tipo convectivo[5] y pueden estar formadas por una o varias "celdas". Las celdas son una unidad dinámica caracterizada por una fuerte corriente ascendente [2]. Durante el transcurso de la tormenta, la celda

[4] Se denomina riesgo agro-climático a la probabilidad de afectación del rendimiento o la calidad de los cultivos por efecto de un fenómeno climático adverso [44].

[5] Este tipo de tormenta tiene su origen en la inestabilidad de una masa de aire más caliente que las circundantes. La masa de aire caliente asciende, se enfría, se condensa y se forma la nubosidad denominada *Cumuliforme* que tienen gran desarrollo vertical. El ascenso de la masa de aire se debe, generalmente, a un mayor calentamiento en superficie [51].

pasa por diferentes estados. En la etapa de maduración, que dura entre 15 y 30 minutos, la nube se denomina *Cúmulonimbus* y se pueden dar condiciones para la formación de granizo ya que al ser una nube de gran desarrollo vertical se produce glaciación en su parte superior [2] [51]. Se caracterizan por presentar una marcada actividad eléctrica y generalmente pueden generar intensas precipitaciones [2] [51].

En el interior de la nube, el aire asciende y se enfría. El vapor de agua contenido, se satura y se forman gotitas de agua que son arrastradas por las corrientes ascendentes. Al continuar el ascenso, en los niveles de temperaturas inferiores a cero grado centígrado, algunas gotas se congelan y forman cristales de hielo. Esta transformación no es abrupta; las gotas pequeñas, de agua pura, pueden permanecer en estado líquido hasta 30ºC bajo cero en un estado metaestable[6]; se habla de gotas de agua sobreenfriadas. Como consecuencia, en las zonas en que la temperatura varía entre -4ºC y -30ºC, la nube está formada por una mezcla de gotas de agua y cristales de hielo [2] (figura 3).

Mientras el aire asciende, algunas gotas de agua chocan con los cristales de hielo y se unen formando cristales y gotas de mayor tamaño. Las gotas más grandes se congelan antes que las de menor tamaño y por ser más pesadas comienzan a caer, llevando consigo a las gotas más pequeñas durante el descenso. Comienza así el proceso de crecimiento del granizo por choque y congelación de gotas sobreenfriadas [2].

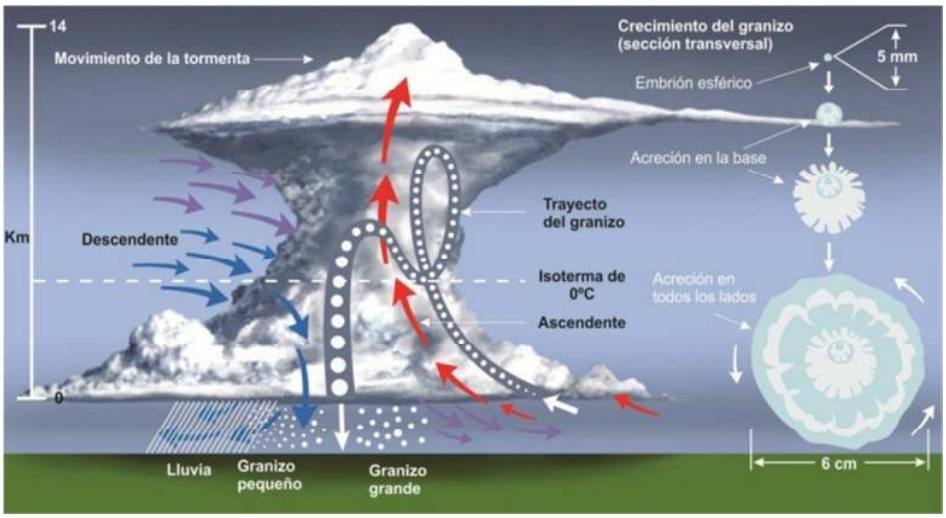


Figura 3. Crecimiento del granizo dentro de una tormenta (Adaptado de [53])

[6] La **metaestabilidad** es la propiedad que un sistema con varios estados de equilibrio, tiene que exhibir, durante un considerable espacio de tiempo. Bajo la acción de perturbaciones externas dichos sistemas exhiben una evolución temporal hacia un estado de equilibrio fuertemente estable. Como ejemplo de metaestabilidad física se presenta el agua en sobrefusión. Las gotas de agua pura en suspensión en un aire también muy puro no se congelan a los 0 °C, sino que siguen en estado líquido hasta alcanzar los - 30 °C. Este estado de sobrefusión cesa bruscamente cuando la gota entra en contacto con un cuerpo externo (como un cristal de hielo) [52].

Después de esta etapa, algunos granizos son suficientemente pesados como para iniciar el descenso, a pesar de la resistencia que les opone la corriente ascendente. Durante el pasaje por la parte inferior de la nube y en el trayecto hasta llegar al suelo, las piedras comienzan a fundirse. Las más pequeñas pueden transformarse totalmente en agua, dando lugar a la formación de grandes gotas de lluvia, mientras que las de mayor tamaño sólo funden su capa externa y llegan al suelo en estado sólido [2] (figura 3).

Las condiciones más favorables para el desarrollo de estas tormentas se producen en horas de la tarde-noche, durante la primavera y el verano y sobre el centro de los continentes y latitudes medias, salvo en el continente africano (figura 4a) [2] [54]. La caída de granizo se produce en regiones puntuales, localizadas y durante períodos de tiempo de corta duración [2] [54].

En la Argentina la mayor frecuencia de granizo ocurre en las regiones montañosas, costeras y comprendidas entre las latitudes de 30° y 40° S [4] [5] (figura 4b). La distribución anual, en las áreas orientales y costeras, presenta la mayor cantidad de eventos en los meses de primavera, pudiendo comenzar en invierno y continuar hasta el comienzo del verano [4] [5]. En el oeste y centro del país, los eventos también predominan en primavera, pero las frecuencias máximas se observan durante los meses de verano. (figura 4b) [4] [5]. Sobre la región Pampeana, la máxima frecuencia de ocurrencia se encuentra en Córdoba (durante la primavera) y La Pampa (en primavera y verano) [4] [5] [6]. Esta frecuencia es mínima en otoño-invierno hasta julio [5].

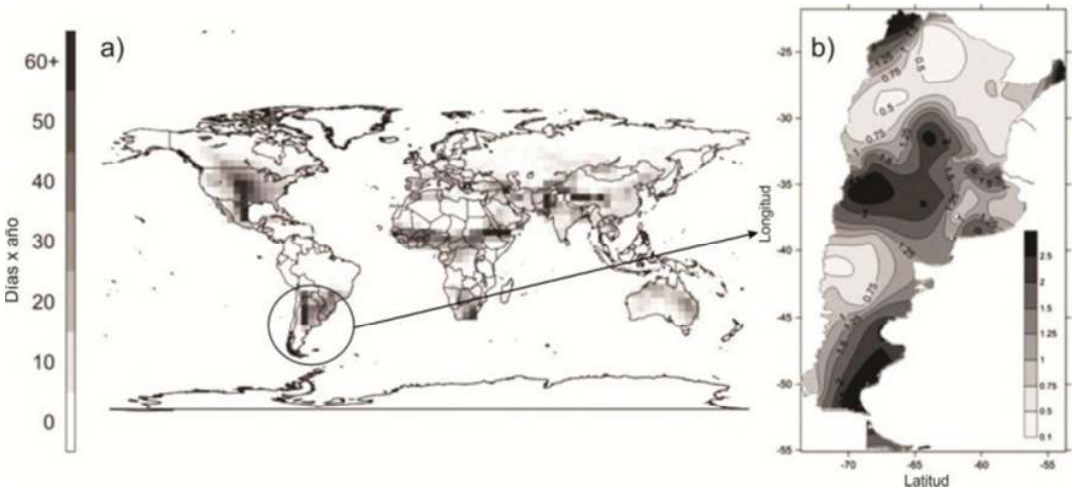


Figura 4a.Días por año con condiciones capaces de generar tormentas fuertes con granizo. (Fuente:[54]), **4b.**Promedio anual de eventos de granizo en Argentina para la serie 1960-2008 (Fuente:[5]).

1.2.3. El Radar Meteorológico

RADAR es el acrónimo de *radio detection and ranging* [55]. En la figura 5a se presenta un diagrama con sus componentes básicos. En [55] y [56] se explican el detalle su funcionamiento y aplicaciones.

Los radares emiten a la atmósfera pulsos de energía electromagnética en el rango de frecuencias de las microondas. La mayoría de los radares meteorológicos operan en las longitudes de onda de 3 cm (banda X), 5 cm (banda C) y 10 cm (banda S) [55]. Cuando dichos pulsos encuentran una partícula, parte de esa energía es absorbida y el resto se dispersa en todas direcciones, devolviendo una fracción hacia la antena del radar que, normalmente, se usa tanto para transmitir como para recibir (figura 5b) [55] [57] [58].

Los pulsos se emiten mientras la antena gira 360 grados en forma horizontal o acimut. Estos giros (revoluciones) comienzan con una elevación cercana a los 0 grados y luego aumenta el ángulo de elevación entre 10 y 12 veces, hasta llegar aproximadamente a 20 grados (figura 5c). Esta exploración, que se ejecuta en aproximadamente 5 minutos cómo mínimo [58], se llama barrido o escaneo del volumen porque explora el volumen completo que rodea al radar adquiriendo información sobre la estructura vertical y distribución de la tormenta [26] (figura 5c).

El factor de reflectividad Z o reflectividad radar se calcula a partir de la medición de la energía del eco reflejado. Es una medida de la capacidad de los blancos de interceptar y devolver la energía del radar, depende de los parámetros físicos del objeto (ej.: forma, composición) y la relación entre la longitud de onda y el tamaño del objeto [55] [58] [59]. Mientras la energía se propaga por el aire o por la tormenta se produce una pérdida (atenuación) de la señal, que es más pronunciada para las longitudes de onda más cortas [55]. La ecuación (1) muestra que Z es la suma de la sexta potencia de los diámetros (D) de todos los blancos que dispersan la energía transmitida en la unidad de volumen muestreada[7] [55] [58] (figura 5d).

$$Z = \sum_{i=1}^{n} D_i^6 \qquad (1)$$

[7] Para partículas esféricas con diámetros significantemente más pequeños que la longitud de onda, como las gotas de lluvia, se asume la dispersión de Rayleigh [55].

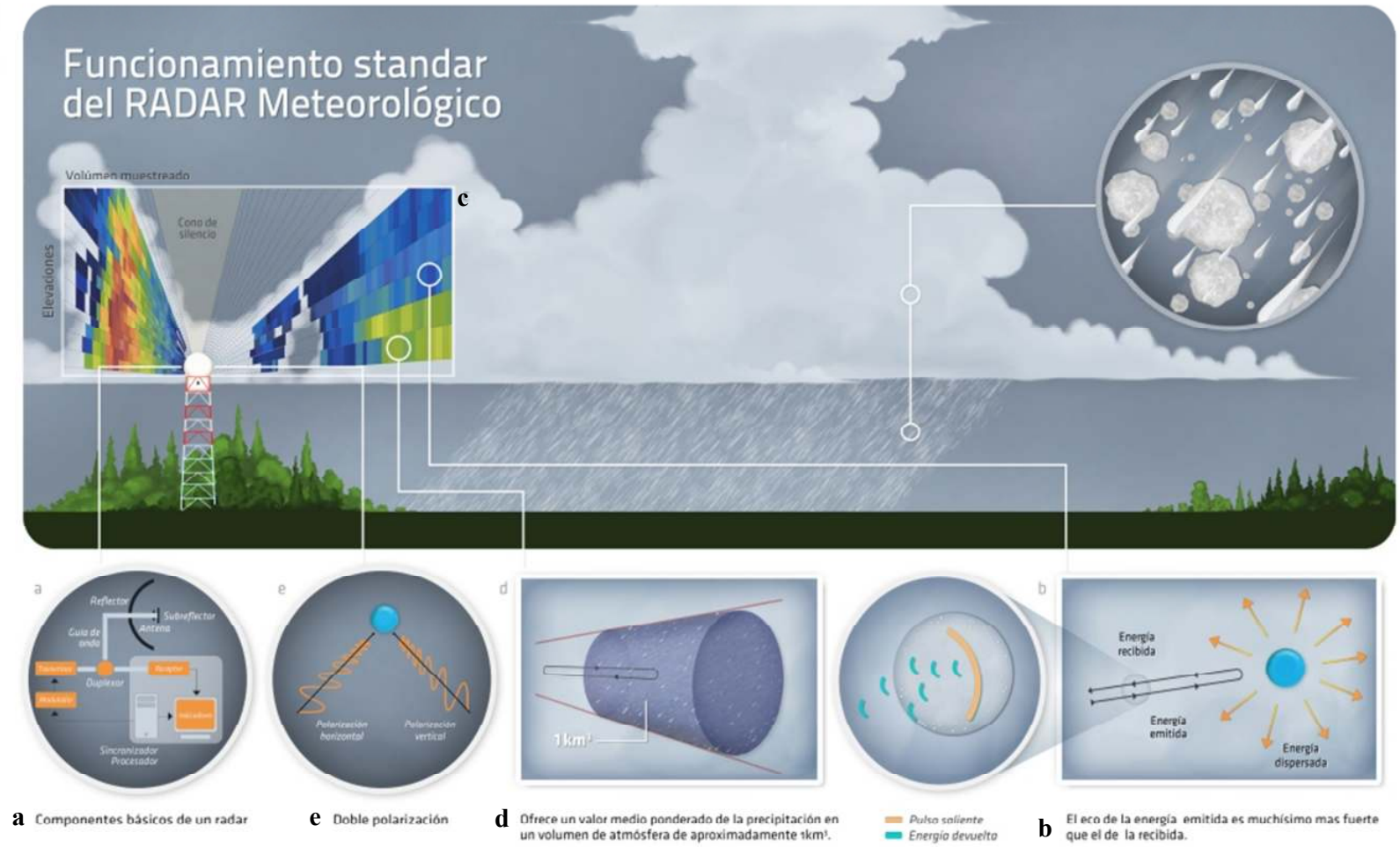


Figura 5. Funcionamiento estándar del RADAR Meteorológico (Adaptado de [55], [60] y [61])

La dependencia de la sexta potencia significa que las partículas grandes predominan en el valor de Z calculado, por lo que la presencia de pocas gotas grandes producen el mismo valor de Z que muchas gotas pequeñas (Figura 6) [55] [58]. Este factor se mide en decibelios (dBZ)[8]. Un alto valor de dBZ significa alta densidad de partículas por unidad de volumen [57].

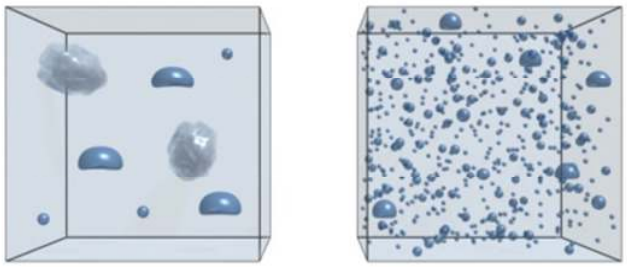


Figura 6. Muestras volumétricas con valores de Z equivalentes (Adaptado de [58]).

La radiación electromagnética de los radares tiene una orientación preferencial porque está polarizada [55]. Los radares de simple polarización utilizan una polarización horizontal [55], en los de doble polarización, el haz electromagnético emitido, tiene orientación vertical y horizontal (figuras 5e) [26]. La comparación de diferentes maneras (proporciones, correlaciones, etc.) de estas dos potencias reflejadas permiten obtener información sobre el tamaño, forma y densidad de las partículas de las nubes [56] y es por esta razón que los radares de doble polarización son más precisos en la detección de hidrometeoros [30] [56] [59] [62] [63] [64] [65] [66]. Las medidas polarimétricas que se utilizan para la detección de granizo y están disponibles en el radar de la EEA Anguil son:

- *Differential Reflectivity - Reflectividad diferencial (Z_{DR}):* valores de -8 a +12 dB (decibelios). Muestra la diferencia en la energía devuelta entre los impulsos horizontales y verticales del radar [55] [56] [59] [60] [61]. Valores de Z_{DR} cercanos a 0 indican que los hidrometeoros en el volumen tienen, en promedio, una forma esférica (su eje vertical es casi igual a su eje horizontal), si los valores son >0, los hidrometeoros en el volumen están orientados horizontalmente (su eje horizontal es más largo que su eje vertical), mientras que si $Z_{DR} < 0$ significa que los

[8] Escala logarítimica que se utiliza para poder disminuir los niveles de magnitud de los valores y hacer más sencilla la interpretación de los mismos [55] [56].

hidrometeoros en el volumen están orientados verticalmente (su eje vertical es más largo que su eje horizontal) (figura 7).

- *Correlation Coefficient - Coeficiente de correlación co-polar (ρHV, Rho_{HV}, CC):* valores entre 0 y 1. Mide la similitud del comportamiento (características de retrodispersión) de los pulsos horizontales y verticales en un volumen [55] [56] [61]. Su precisión disminuye con la distancia y con la mezcla de tipos de hidrometeoros en el volumen. Valores de $Rho_{HV} < 0.80$ evidencia una dispersión compleja: los pulsos horizontales y verticales cambian de diferentes maneras de pulso a pulso. Valores de 0.80 a 0.97 muestran diferencias moderadas entre los pulsos y valores de $Rho_{HV} > 0.97$ implican una dispersión ordenada, por lo tanto hay muy pequeñas diferencias entre los pulsos horizontales y verticales. Para esferas perfectas Rho_{HV} =1 [55] [60] [61] (figura 7).
- *Differential Propagation Phase - Fase de propagación diferencial (ΦDP, Phi_{DP}):* valores de 0º a 360 º (grados). Las ondas electromagnéticas cambian de fase mientras pasan a través de material más denso [55]. Dependiendo de la forma, orientación y concentración de los hidrometeoros, se producen ligeros cambios de fase entre las señales vertical y horizontal, retardándose una con respecto de la otra. El agua líquida es uno de los medios que más resistencia presenta [55] [56] [67]. Por ejemplo, si los hidrometeoros son oblatos (como las gotas de lluvia), Phi_{DP} aumenta porque la fase horizontal se propaga más despacio que la vertical. Esta variable no dice nada por sí misma, pero su cambio en el tiempo y en el espacio si, por eso se calcula Specific Differential Phase.
- *Specific Differential Phase – Fase diferencial específica (K_{DP}):* valores de -18; +36 º/km (grados por kilómetro). Deriva de Phi_{DP} [67]. Es la diferencia, en un rango, entre los impulsos horizontales y verticales del radar a medida que se propagan a través de un medio y son atenuadas [55] [61]. La correspondencia de valores positivos (horizontal) y negativos (vertical) es análoga a Z_{DR}. K_{DP} es dependiente de la concentración de partículas, por lo que muchas gotas pequeñas pueden generar un valor positivo mayor que pocas gotas grandes [55] [61] [67].

La forma más común de mostrar los datos del radar es por medio de imágenes [55], dos de las más comunes son 1) Plan Position Indicator (PPI- Indicador de posición en un plano): en este escaneo el radar tiene su elevación fija y varía su ángulo de acimut [55] [68] y 2) Constant Altitude PPI (CAPPI – Indicador de posición en un plano de altitud

constante): es un PPI de altura constante. Provee el dato radárico a una altitud prefijada utilizando la información de las diferentes elevaciones para realizar una imagen completa [55].

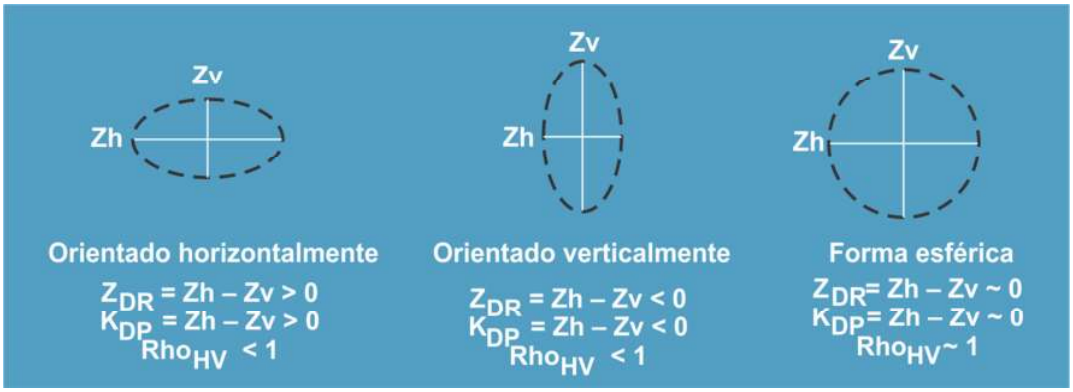


Figura 7. Relación entre los ejes verticales y horizontales de los hidrometeoros y las variables registradas por el radar.

1.3. Antecedentes

1.3.1. Criterios de detección de granizo basados en radares meteorológicos

Existen numerosos estudios sobre la relación de las señales registradas por los radares y el granizo. Los primeros exploraron datos de simple polarización solos o combinados con datos de otros sensores [64]. Con la aparición de los radares de doble polarización, los trabajos se concentraron en estudiar la relación entre ambas reflectividades que permite distinguir entre la forma esférica y rotativa de los granizos versus la forma no esférica de las gotas de lluvias [69]. Finalmente, las investigaciones más recientes, estudian las diferencias de comportamiento de las señales de diferentes bandas (C vs. S) ante el granizo, ya que las longitudes de onda más cortas tienen mayor atenuación, especialmente ante la presencia de lluvia intensa [12] [62] [63] [70] [71] [72] [73].

En [20] y [69] se realizan revisiones muy completas de métodos de simple polarización para detectar ocurrencia de granizo. En la tabla 1 y la figura 8 se resumen los principales métodos y las variables que utilizan; en síntesis todos hacen referencia a la presencia de núcleos de Z por encima de un umbral (36 a 55 dBZ) y en algunos casos la combinan con información de temperatura y variables calculadas a partir de Z, como la energía cinética (E) [20] [26] [30] [34] [66] [69] [74]. Finalmente, en [36] se aplica un criterio de combinación de cuatro métodos y se clasifica cada pixel de acuerdo a la mayor coincidencia de al menos dos de los métodos, logrando mejorar el rendimiento con respecto a cada método de forma individual.

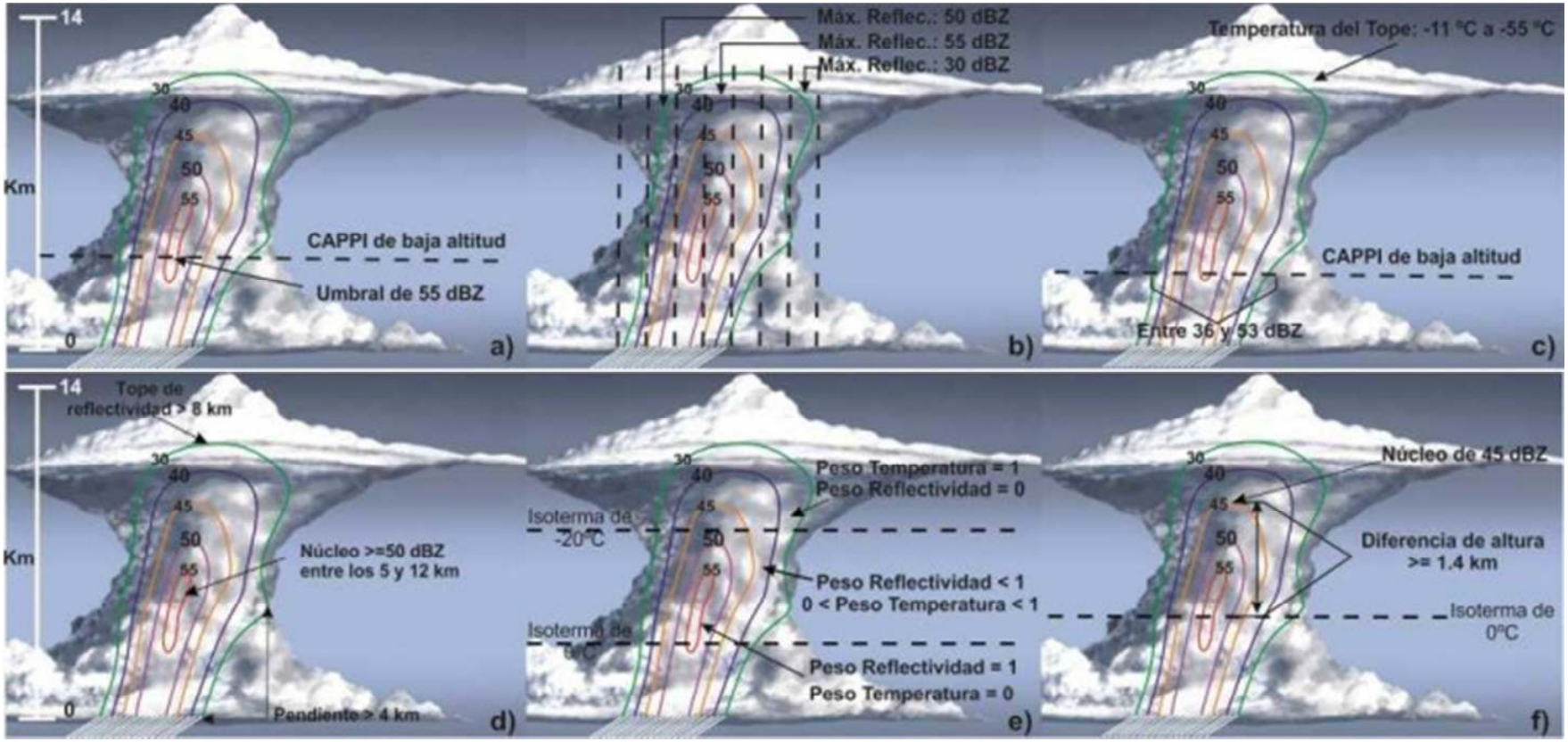


Figura 8. Esquema de variables utilizadas por el método: a) CAPPI (elaboración propia), b) máxima reflectividad (elaboración propia), c) Reflectividad y temperatura de tope de nube (elaboración propia),d) HDA(Adaptado de: [20]), e) SHI(Adaptado de: [20]), f) Waldvogel, Witt, Holleman (Adaptado de: [20]).

Tabla 1. Resumen de variables predictoras de granizo mediante métodos de simple polarización.

Método	Variables predictoras
CAPPI	Valor de Z a una altitud baja y constante. Se sugiere un umbral de 55 dBZ para distinguir entre granizo y lluvia utilizando este método (figura 8a) [69].
Máxima reflectividad y persistencia de las máximas reflectividades.	Máximo valor de Z observado en una columna vertical. Umbral: 45 dBZ (figura 8b) [20] [26].
Vertically Integrated Liquid (VIL)	Total de agua líquida presente en una columna sobre cierta posición. A mayor valor, mayor probabilidad de granizo. Discrepancia entre autores sobre el mejor umbral de VIL [26] [69].
VIL Density	División de VIL por la altura del tope del eco del radar. Se propone como umbral universal para alarma de granizo 3.5 g/m^3. Existen estudios que cuestionan este trabajo [26] [69].
Reflectividad y temperatura de tope de nube	Combinación de Z (entre 36 y 53 dBZ) a baja altitud y temperatura en el tope de la nube (entre -11ºC y -55ºC.) [69] (figura 8c).
Hail Detection Algorithm (HDA)	La primera versión tenía como parámetros: un núcleo de Z >=50 dBZ, ecos mayores a 55 dBZ entre los 5 y 12 km de altitud y presencia de ecos igual o más altos que 8 km (figura 8d) [66] [69]. Las otras versiones corresponden a las de Waldvogel, Witt y Holleman [20].
Severe Hail Index (SHI)	Se calcula por una integración vertical del producto del flujo de E del granizo con una función de temperatura y de Z (figura 8e) [20] [26] [74], (ver 1.3.2).
Waldvogel, Witt, Holleman	Núcleo de Z >= 45 dBZ a 1.4 km o más por encima del nivel de congelamiento. Una diferencia de 1.6 Km corresponde a un 10% de probabilidad de granizo y una de 6 km es 100% de probabilidad

	(figura 8f) [20][26][69]. *ZHAIL* está, basado en estos algoritmos. Una explicación detallada del mismo se presenta en [34].

En el caso de la doble polarización, la tabla 2 resume los valores que toma cada variable polarimétrica en presencia de granizo. En [63], [65] y [75] se detallan las técnicas de doble polarización y en [71] se describen las diferencias de comportamiento encontradas entre las bandas S y C en presencia de granizo. La mayoría de los algoritmos para banda S se basan en la anticorrelación que existe entre altos valores de Z y pequeños valores de Z_{DR} cuando hay granizo. También se ha encontrado que los valores Rho_{HV} son bajos [56] [63]. Diversos autores sugieren que esta anticorrelación se puede aplicar directamente en banda C [63], pero existen estudios con evidencias que refutan esta teoría [13] [31] [63] [71] [75]. En banda C se encontraron valores elevados de Z, valores altos de Z_{DR}, valores bajos y negativos de K_{DP} y una abrupta caída de Rho_{HV} [31] [63] (tabla 2). Algunos autores sugieren que la relación $Z_{DR}<1.0$ dB y $Z>50$ dBZ se debe observar en radares de banda C cuando los granizos son grandes (12 a 15 mm de diámetro) y en mucha cantidad como para dominar Z y aparecer como esferas [63]. Se utiliza el enfoque de lógica difusa como metodología más difundida para la generación de algoritmos de clasificación [76] [77], la cual que se explica en detalle en [25] [73] [78] y [79]. Brevemente, se basa en una combinación de funciones miembro asociada con un tipo de hidrometeoro en particular. Estas funciones tiene dos argumentos: Z y una de las variables polarimétricas. La clasificación se hace tomando el valor más alto de la combinación de funciones para cada píxel de la imagen [73] [77]. El uso de esta técnica se justifica en el hecho que en una tormenta, el granizo se superpone con otros hidrometeoros, presentando límites pocos claros, por lo que se necesita buscar una solución sin límites rígidos [73] [77]. Las funciones obtenidas muestran un alto grado de flexibilidad y pueden ser empíricamente adaptadas a evidencias experimentales [76].

Tabla 2. Listado de variables polarimétricas y sus valores ante la presencia de granizo para **banda S** ([13] [56] [63] [64] [65] [80]) y **banda C** ([13] [63] [77] [81] [82]).

Variable	Valores para Banda S	Valores para Banda C
Z*	> 45 dBZ > 50 dBZ	Alto > 50 dBZ Granizo pequeño (< 2 cm): 50 a 60 dBZ Granizo grande (> 2 cm): 55 a 65 dBZ Granizo con lluvia: 45 a 80 dBZ
Rho_{HV}	$0.90 < Rho_{HV} < 0.98$	< 0.94, < 0.95 Granizo pequeño (< 2 cm): 0.92 a 0.95 Granizo grande (> 2 cm): 0.90 a 0.92 Granizo con lluvia: > 0.90
Z_{DR}*	< 1 dB granizo solo 1 dB <= Z_{DR} <= 4 dB granizo rodeado de lluvia.	Alto (3 – 8 dB) Granizo pequeño (< 2 cm): -0.5 a 0.5 dB Granizo grande (> 2 cm): -1 a 0.5 dB Granizo con lluvia: -1 a 6 dB > 4 dB, > 3 dB
K_{DP}		Alto > 4º/Km
* Estas dos variables se presentan como las más discriminantes de acuerdo a [73] y [76].		

1.3.2. Criterios de estimación de daño en cultivos por granizo basados en radares meteorológicos

Los primeros trabajos en relacionar el granizo con el daño en cultivos[9] utilizaron redes de granizómetros junto a información de compañías de seguros y redes de observadores a campo, determinando las variables que permiten predecir el nivel de daño en cultivos por granizo: a) E [12] [45] [83] [84] , b) impulso [45], c) masa [45], d) número de granizos con diámetro mayor a un determinado tamaño (ej:6.4 mm [83], 12.7 mm [45]) [84], e) tamaño del granizo [12] [48] f) viento [45] [84] g) cultivo [83] [84] y h) estado fenológico del cultivo [45] [83] [84]. Como las redes de granizómetros son muy costosas, se realizaron estudios para relacionar las variables

[9] Daño realizado por granizo no simulado. Existen trabajos donde se simula el daño ocasionado por el granizo, pero estudios posteriores cuestionan la exactitud de dicha simulación [45].

que caracterizan la caída del granizo por medio de la reflectividad medida por un radar meteorológico [85] [86]. Los resultados para banda S y C determinaron una relación empírica [85], de la forma:

$$E = a \times 10^{-6} \times 10^{b \times Z} \quad (2)$$

Para la cual se han definido diversos valores de a y b de acuerdo a la región de estudio (ej.: [9],[29],[30],[85],[47]), el más utilizado es a=5 y b=0,084 determinado por [85]. Para utilizar la relación (2), se generaron diversas técnicas para decidir en qué zonas tiene sentido calcular E[10] [29]: a) *"Cutting method":* se fijan umbrales de Z (ej.: de 49 a 65 dBZ) [24] [29]) y de la altura de H_{45}[11] (las cuales dependen de la región de estudio [29][85][47]), b) *Función de peso de Z:* donde a (2) se le agrega un multiplicando W(Z), el cual se define como [26] [74]:

$$W(Z) = \begin{cases} 0 & para\ Z \leq Zmin \\ \frac{Z - Zmin}{Zmin - Zmax} & para\ Zmin < Z < Zmax \\ 1 & para\ Z \geq Zmax \end{cases} \quad (3)$$

Los valores Zmin y Zmax también dependen de la región de estudio, algunos de los valores utilizados son: 40-50, 45-55, 50-60 [26] [36] [74] [88]. Finalmente, para muchas aplicaciones, el valor de E se debe integrar en el tiempo, durante la duración de la tormenta y para cada pixel del radar [89]. La forma de relación más común es la definición de una regresión logística de la variable E, para cada cultivo y estado fenológico del mismo [45] [47] [87].

En nuestro país se estudió la relación de la probabilidad de granizo generada por el algoritmo ZHAIL de los radares de INTA Paraná e INTA Pergamino con el porcentaje de daño en cultivos y no se encontró una relación clara [32], generando muchas falsas alarmas [32][33].

En cuanto a las variables polarimétricas de banda C, [33] analizó la relación entre el parámetro Hail Differential Reflectivity (H_{DR}), calculado con datos del radar de Paraná y el daño en cultivos, encontrando una buena relación, donde el mayor daño

[10] Para aplicar este método es necesario distinguir entre granizo y lluvia dentro de la tormenta, porque solo se debe calcular E en las zonas con granizo [87].

[11] H_{45} hace referencia a la altura sobre la isoterma de cero grado de los ecos iguales o mayores a 45 dBZ, los cuales se asocian a granizo.

en cultivo coincide con el núcleo de 50 H_{DR}, que puede indicar granizo de tamaño medio [33].

El parámetro H_{DR} fue desarrollado por [59] para distinguir granizo de lluvia utilizando Z y Z_{DR}, por medio de la siguiente relación [33]:

$$H_{DR} = Z - f(Z_{DR}) \text{ donde:}$$

$$\begin{aligned} f(Z_{DR}) &= 21 \text{ dB} && \text{if } Z_{DR} \leq 0 \text{ dB} \\ &= 19\, Z_{DR} + 27 \text{ dB} && \text{if } 0 < Z_{DR} \leq 1.74 \text{ dB} \\ &= 60 \text{ dB} && \text{if } Z_{DR} > 1.74 \text{ dB} \end{aligned} \quad (4)$$

H_{DR} es un buen indicador de presencia de granizo, para las elevaciones más bajas y que se encuentran por debajo del nivel de congelamiento. Mayor valor de H_{DR} indica mayor probabilidad que las medidas de Z no son debido a gotas de lluvia [33].

1.3.3. Algoritmos de minería de datos aplicados a datos de radares meteorológicos.

Existen diversos estudios que utilizan técnicas de DM en el análisis de datos de radares meteorológicos (ej.: [12], [14], [18], [19], [22], [40], [90], [91], [92], [93], [94], [95]). Estos trabajos, en general, presentan cuatro tipos de objetivos: a) clasificar hidrometeoros donde una de las clases es "granizo" solo o combinado (ej.: [14], [17], [20], [21], [96]), b) clasificar tormentas donde uno de los tipos indica granizo (ej.: [18], [19], [22], [23], [57], [90], [94], [97], [98], [99]), c) determinar el tamaño del granizo (ej.: [17], [20], [21], [96]) y d) determinar el daño producido por el granizo (ej.: [9], [12], [45], [47], [83], [89], [87]).

Las regresiones, y en particular la regresión logística, aparecen como el método más usado (ej.: [9], [12], [17], [45], [47], [83], [87], [89], [97]), especialmente para determinar el daño en cultivos. Otros métodos mencionados en la bibliografía son: a) las redes neuronales (ej.: [18], [21], [22], [90], [91]), b) los vectores soporte (ej.:[22], [92], [93], [98], c) los algoritmos genéticos (ej.:[19], [90]), d) los árboles de decisión (ej.: [23], [90], [96]), e) técnicas de clustering (ej.: [23], [90], [94], [98], [99]) f) clasificador Bayes (ej.: [96], [98]) y g) métodos rough set (ej.: [57], [92], [95]).

Los trabajos analizados utilizan diversas variables de entrada donde se combinan datos del radar con información de otros sensores (ej.: radio sondeos, satélites, etc.)

[26]. En diversos trabajos (ej.:[18] [19] [22] [57] [90] [92] [95]) se cuenta con un banco de datos de tormentas con variables previamente calculadas de las mismas.

Cuando se trabaja solo con datos de radar, las variables de los modelos son: a) variables que caracterizan la tormenta (ej.: tipo de tormenta, duración, etc.), b) diversas variables derivadas a partir de Z (ej.: VIL, VIL Density, E, SHI, etc.) y c) variables derivadas de las variables polarimétricas (ej.: $H_{DR,}$ Linear Depolarization Radio (L_{DR}), etc.)

Capítulo 2. Materiales y Métodos

Este capítulo se ocupa de continuar con la descripción de los trabajos realizados para la etapa de "Comprender los datos" y detallar las actividades de las etapas "Preparar los datos" y "Modelar" del proceso CRISP-DM.

2.1. Área de estudio y período de tiempo

El área de estudio corresponde a la zona de influencia del radar situado en la EEA Anguil en el rango de los 240 km de radio con centro en el radar (figura 9). El período temporal se fijó de Marzo de 2009 (puesta en funcionamiento del radar [100]) a Marzo de 2013 donde se registraron 111 fechas con tormentas, de ellas, 93 corresponden al período primavero-estival, en el cual se centró el análisis por ser el más prolífico en tormentas convectivas en la zona bajo estudio (ver sección 2.1.2).

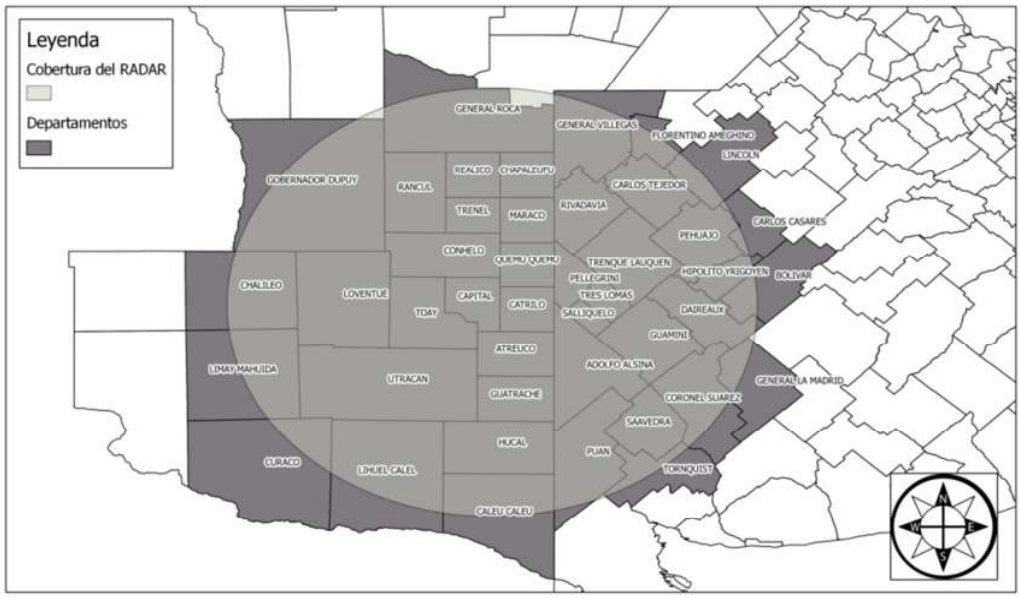


Figura 9. Mapa del área de influencia de 240 km de radio del radar de la EEA Anguil (Fuente: elaboración propia)

2.2. Datos

Se tomaron dos tipos de datos: 1) los datos de campo que detallan la ocurrencia de granizo y daño en cultivos y 2) los registrados por el radar.

2.2.1. Datos de campo

No existe a nivel nacional un sistema oficial que concentre y gestione el registro de la información de ocurrencia e intensidad de eventos de granizo de forma exhaustiva; por lo que esta información de campo se tuvo que relevar por medio de diferentes fuentes, con el objetivo de obtener la mayor cantidad de puntos posibles para las fechas registradas.

2.2.1.1. Variables

A partir del análisis de antecedentes se determinó que las variables de terreno necesarias para el armado de los modelos son: 1) *fecha del evento* (formato: dd-mm-aaaa) y *hora del evento* (hora de inicio y fin en formato hh:mm), 2) *ocurrencia de granizo* (si, no), 3) *posición geográfica* (en coordenadas geográficas, latitud y longitud), 4) *cultivo* (en caso que corresponda a una localización no urbana se registra el uso del suelo, si es un cultivo, se registra su nombre vulgar, por ejemplo: Girasol, Soja, Maíz, Trigo, etc.), 5) *estado fenológico* (dato mínimo: Vegetativo, Reproductivo, Madurez, de ser posible nombre correspondientes de acuerdo a escalas fenológicas definidas para cada cultivo: Trigo: escala de Zadoks, Chang y Konzak, 1974; Girasol: escala de Schneiter y Miller, 1974; Soja: escala de Fehr y Caviness, 1977; Maíz: escala de Ritchie y Hanway, 1982), 6) *daño* (Sin Daño, Leve, Moderado, Severo y Grave, para detalle ver sección 3.2.2.1) y 7) *área afectada*. La figura 10 resume el esquema del trabajo realizado.

2.2.1.2. Fuentes de Datos

Las compañías de seguros agrícolas proveen reportes que se utilizan ampliamente como información para la validación de métodos de estimación de ocurrencia y tamaño de granizo ([20], [31], [32], [33], [69], [83], [101], [102]). Se contactaron vía correo electrónico 29[12] empresas. Las empresas que enviaron información fueron Sancor Seguros, La Segunda y La Dulce Cooperativa de Seguros; si bien los registros recibidos tienen diferente formato (.xls[13], .kml[14] y .txt[15]) y diferentes datos,

[12] Esta cantidad de empresas se obtuvo de [49]. Los datos de contacto se buscaron en los sitios web de cada empresa.
[13] Extensión por defecto del formato Excel en versiones anteriores o iguales a Excel 2003 [103].

todos comparten dos atributos: 1) georeferenciación del establecimiento donde ocurrió el evento y 2) una calificación del daño ocasionado por el granizo en el cultivo. La metodología utilizada para dicha calificación difiere entre compañías, pero en todos los casos es realizada por un perito, que hace una evaluación en terreno.

Figura 10. Esquema de relevamiento de datos de campo (Fuente: elaboración propia)

Como segunda fuente mencionada en otros estudios de granizo (ej.: [59], [72], [97], [101], [106], [107], [108]) aparecen los medios de comunicación y redes sociales. Los eventos severos son un tema sensible capturado por los diarios y radios locales en el área de estudio. Las notas sobre las tormentas y sus consecuencias son

[14] Extensión de los archivos escritos en el lenguaje de marcado basado en XML para representar datos geográficos en tres dimensiones [104]. Utilizado por google.
[15] Extensión más difundida de los archivos de texto plano; compuestos únicamente por texto sin formato, sólo caracteres [105].

acompañadas por evidencia fotográfica o fílmica, además de testimonios de entidades oficiales como la policía, los bomberos, cooperativas, asociaciones de productores o gobiernos municipales; sumado a los contenidos de fotos y videos que comparten los usuarios en la web 2.0, permitió obtener 223 reportes en 63 medios de comunicación y redes sociales que se presentan en formato digital (diarios, radios, televisión, blogs, videos y sitios oficiales). Del análisis de estos reportes se extrajo información diversa como: localidades afectadas, superficie del daño, horario, duración, tipo de evento (lluvia, viento, granizo o sus combinaciones) y efecto causado en terreno por las tormentas. Para la base de datos de verdad de campo se registró: la fecha del evento, la georeferenciación de las localidades afectadas y el horario de inicio y fin del evento, en caso que figurara. Si las localidades mencionadas en los reportes, cuentan con un medidor en alguna red de observación meteorológica a la cual se tenía acceso, esa localidad solo se agregaba si en la red se informaba precipitación.

Numerosos autores como [5], [59], [68], [90], [101], [106] y [109], utilizan diferentes tipos de redes de información como fuente de datos. En este trabajo se utilizaron datos de: 1) Red de informantes del Servicio Meteorológico Nacional (SMN), los cuales reportan eventos extremos a este organismo que se publican en http://www.smn.gov.ar/?mod=voluntarios&id=1, 2) Red de pluviómetros de la policía de La Pampa que informa en http://www.policia.lapampa.gov.ar/lluvias.php, 3) Red termo pluviométrica de la Red de Información Agropecuaria Nacional (RIAN), que releva información diaria de precipitaciones en aproximadamente 100 puntos en el área de estudio (http://rian.inta.gov.ar/agua/informes.aspx) [110] y 4) Sistema Integrado de Información Agropecuaria (SIIA) que genera informes semanales de estimaciones de superficie sembrada por delegación provincial, indicando para cada cultivo el estado general, avances de labores, estado fenológico y principales adversidades (http://www.siia.gob.ar/informes).

También se recurre a los informantes calificados para obtener datos de campo (ej.: [14], [17], [19], [26], [63], [90], [101], [106], [111], [112]). En este trabajo se contactaron a extensionistas de INTA, productores y contratistas locales. Con Google Earth se ubicaron y digitalizaron los lotes informados y se cargó la información de ocurrencia de granizo (si, no), cultivo afectado y porcentaje de daño (figura 11).

Figura 11. Primer tramo de la recorrida de campo correspondiente al evento 10-12-2012 (Fuente: elaboración propia).

Otra herramienta utilizada en trabajos previos, son las recorridas de campo posteriores a una tormenta ([59], [64], [111] y [113]). Se aprovecharon recorridas mensuales de la RIAN donde se registran datos referentes al estado y evolución de los cultivos y se incluyen las principales adversidades de origen ambiental (granizo, heladas, sequías, etc.) [110]. También se realizaron dos salidas posteriores a las tormentas del 10-12-2012 y el 24-12-2012. Ambas recorridas se realizaron por rutas y caminos vecinales, acompañados por productores y contratistas locales, los cuales guiaron en el recorrido y permitieron el acceso a los lotes con daño en sus establecimientos agropecuarios. Durante el recorrido se obtuvo información de georeferenciación (latitud y longitud con un GPS Garmin eTrex Legend HCx), cultivo dañado, porcentaje de daño y tareas a realizar en los lotes con daño por granizo (ej.: resiembra, picado, etc.). También se registraron las coordenadas geográficas de lotes sin daño de granizo presentes en el área de la tormenta. En la figura 11 se muestra un mapa con las recorridas realizadas para el relevamiento correspondiente al evento del 10-12-2012 y en la figura 12 se presentan fotos de lotes dañados por granizo.

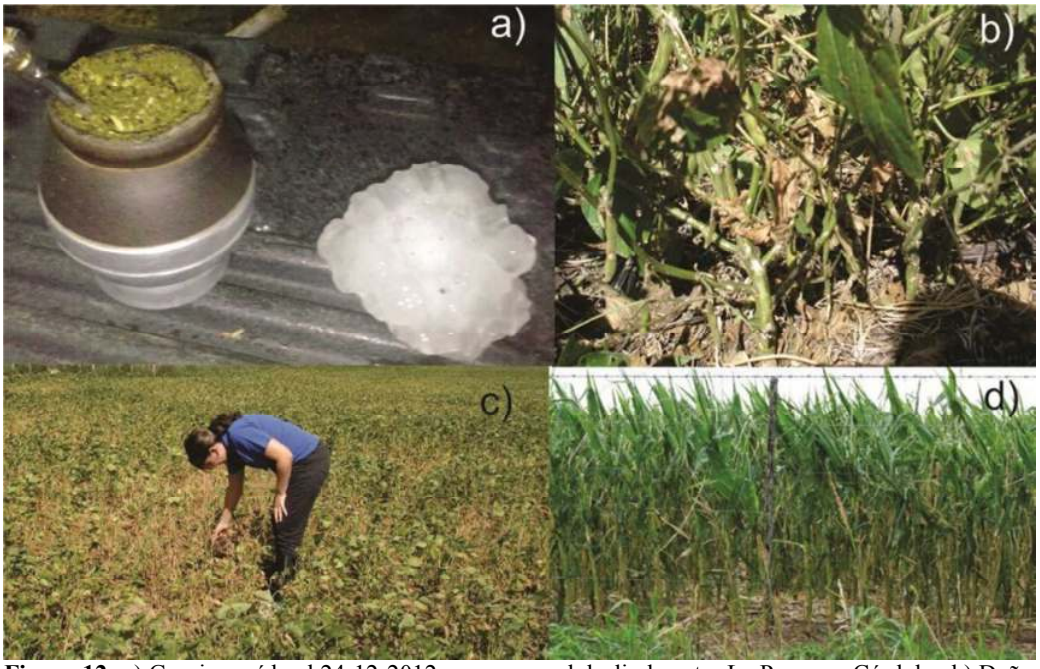


Figura 12. a) Granizo caído el 24-12-2012 en zona rural deslinde entre La Pampa y Córdoba. b) Daño en lote de soja tormenta del 10-12-2012. c) Lote de soja con daño de granizo tormenta del 01-03-2013, d) Lote de maíz con daño tormenta del 10-12-2012 (Fuente: elaboración propia).

Para almacenar y procesar toda la información relevada sobre una fecha, se diseñó y compiló una base de datos que se realizó en SQL Server 2008 Express con extensión para el tratamiento de datos geográficos. La misma almacena para cada reporte, las localizaciones geográficas del evento, sus fuentes correspondientes y datos adicionales (duración del evento, cultivo, daño en cultivo, fotografías, videos, etc.). Se administra por medio de un sistema de información que se desarrolló en plataforma web ASP.NET, utilizando Visual Basic .NET como lenguaje de programación. En el anexo 2 se presentan los detalles del sistema desarrollado y de la información relevada.

Los datos sistematizados se analizaron para determinar su inclusión en el estudio. Se obtuvieron 3.758 puntos geo referenciados de ocurrencia de granizo (1.771, 47% positivos y 1.987, 53% negativos). Más del 80% tienen información sobre intensidad y porcentaje de daño, el 30% tiene información de fenología, el 34% cuenta con el detalle del cultivo y del tipo de cosecha (fina o gruesa). El horario de la tormenta no se pudo determinar en la mayoría de los casos debido a la ocurrencia de más de una tormenta en el mismo día, a diversas celdas de una misma tormenta en diferentes localizaciones o a datos de inicio y fin diferentes debido al movimiento propio de la tormenta. Primero se filtraron las localizaciones que estuvieran contenidas dentro

del área de cobertura del radar. Los datos fueron exportados a archivos separados por tabulaciones e importados al software Quantum GIS (QGIS), en el cual se tenían cargados los departamentos de las provincias del área de estudio y el área de influencia del radar (Figura 13). Para aquellos puntos que no entraban en el área de cobertura del radar se colocaba el atributo *"Dentro"* con valor 0 (cero), el cual posteriormente se actualizaba en la base de datos con sentencias transact-SQL (T-SQL)[16].

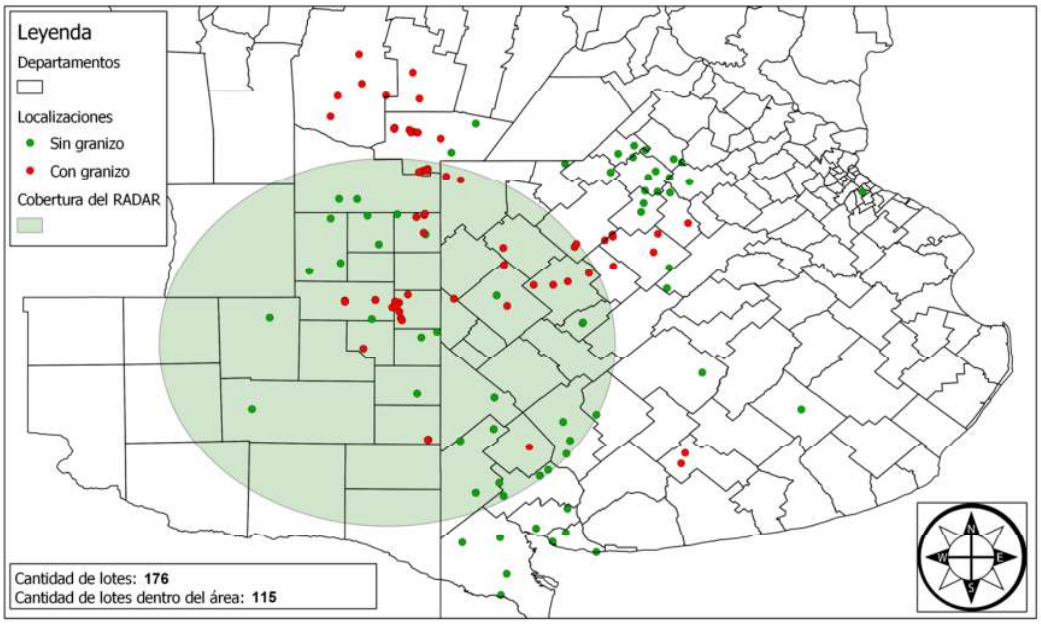


Figura 13. Localizaciones relevadas para el 08-11-2011. Sesenta y un lotes se descartan para el estudio por caer fuera del área de cobertura (Fuente: elaboración propia).

Con este filtro, la cantidad de localizaciones por fecha disminuyó. De las 111 fechas, 77 presentaban al menos un lote dentro del área de cobertura del radar. Se decidió incluir en este trabajo las fechas con veinte o más datos de campo dentro del área de cobertura (tabla 3). Este corte se realizó buscando un equilibrio entre esfuerzo de procesamiento por cada fecha y la cantidad de casos que esa fecha aportaba, ya que el tiempo necesario para procesar una fecha es de 7 días aproximadamente (ver sección 3.3. para más detalle). Las fechas descartadas presentaban 8 lotes o menos (23 tenían 1 lote, 7 tenían 2 lotes, 13 tenían 3 lotes, 5 tenían 4 y 5 lotes, 2 tenían 6 lotes, 3 tenían 7 lotes y 1 tenía 8 lotes), descartando 52 fechas y 162 lotes en total, de

[16] T-SQL es una extensión al SQL de Microsoft y Sybase. SQL es un lenguaje de búsquedas estructurado estandarizado para realizar búsquedas, alterar y definir bases de datos relacionales utilizando sentencias declarativas [114].

los cuales el 49% la fuente son los medios de comunicación, el 44% las aseguradoras y el 7% las Redes de Información.

Tabla 3. Fechas con más de veinte localizaciones dentro del área de cobertura.

Fecha	Total	Dentro del área	Fecha	Total	Dentro del área
10/12/2012*	604	546 (90%)	29/11/2011	230	119 (51%)
15/01/2011*	297	289 (97%)	08/11/2011	176	115 (65%)
25/11/2010*	276	261 (94%)	01/03/2013	237	115 (48%)
29/12/2011	227	227 (100%)	13/01/2011*	105	102 (97%)
25/12/2012	212	212 (100%)	07/11/2011	99	99 (100%)
16/01/2012	212	212 (100%)	24/12/2012	108	82 (76%)
03/12/2011	212	212 (100%)	20/11/2011	91	53 (58%)
15/01/2012	171	171 (100%)	03/11/2011	55	47 (85%)
23/11/2011	131	131 (100%)	24/11/2011	35	25 (71%)
Total lotes volúmenes completos				2.196	1.820 (83%)
Total lotes volúmenes incompletos (fechas marcadas con *)				1.282	1.198 (93%)
Total lotes disponibles				3.478	3.018 (87%)

En segunda instancia se trabajaron las variables target, definiendo dos tipos:

a) *Granizo (0,1):* variable binaria con valor uno en caso de ocurrencia de granizo y valor cero en caso de no ocurrencia. Generada en el relevamiento de los datos.
b) *Daño (Sin Daño, Leve, Moderado, Severo, Grave):* variable categórica con diferentes valores de acuerdo a la intensidad del daño ocasionado por el evento de granizo sobre cultivos extensivos. Se generó procesando los datos contenidos en atributos de texto que contenían una calificación del daño o un porcentaje de daño (ejemplos en tabla 4), utilizando sentencias T-SQL, con el criterio de [33]: Sin Daño: 0%, Leve: 1% a 24%, Moderado: 25% a 49%, Severo: 50% a 74%, Grave: 75% a 100%.

Tabla 4. Ejemplo de datos de porcentaje de daño presente en campos de texto.

Atributo de texto Nombre	**Atributo de texto Observación**
Lote 16.Con Daño. Soja, 100% perdida, resembrado.	Fecha Relevamiento: Diciembre de 2012 […]
	Daño Estimado: 18.3%, Cultivo: TRIGO […]
	Nivel de Daño: LEVE - % Daño: 0,0291588785046729 […]

2.2.2. Datos del RADAR

Uno de los objetivos de este trabajo es aprovechar la característica de doble polarización del radar y contribuir a la caracterización del comportamiento de las variables polarimétricas. Es de interés que el modelo generado no contenga variables exógenas al radar, para no depender de otro sensor.

Se accedió a los archivos de las variables polarimétricas en su formato original, pero no se contó con el software propietario del radar para el manejo de los datos; por lo que se desarrolló un software que leyera y procesara los archivos de datos. Esto también contribuye a los objetivos específicos de minimizar el uso de software propietario y generar herramientas que faciliten el acceso a los datos del radar. Estos archivos presentan dos formatos diferentes, uno para las variables Z, Rho_{HV}, Z_{DR} y otro para Phi_{DP} y K_{DP}, por lo que es necesaria una programación diferenciada para cada conjunto de variables. Los antecedentes indican que Z, $Z_{DR,}$ Rho_{HV} y $K_{DP,}$ son las variables de doble polarización discriminantes de granizo, siendo las dos primeras las más importantes. Además, contar con Z y Z_{DR} permite el cálculo de otras variables tanto polarimétricas como de simple polarización. Ante esta situación, se programó la lectura de Z, Rho_{HV} y Z_{DR} porque permitió contar con mayor cantidad de datos del radar. A partir de las variables y herramientas disponibles, se programa el cálculo de las siguientes variables derivadas de Z y mencionadas en los antecedentes: a) $H_{DR,}$ b) valor máximo de Z vertical, c) Z media vertical, d) elevación más alta con Z diferente de 0, e) valor de Z de la elevación más alta donde Z es diferente de 0, f) número de elevaciones de Z igual o superior a los umbrales 45, 50, 55 y 60 dBZ, g) Número de elevaciones con Z diferente de 0 y h) E. También se agregaron otros cálculos como por ejemplo mínimos, máximos, promedios, totales por elevación y de todas las elevaciones, etc. (ver detalle en la sección 2.2.3.)

Se utilizan los datos provistos por el radar del INTA que opera en la EEA Anguil, localizado en latitud -36,539684, longitud -63,990067 a 200 metros sobre el nivel del mar. La tabla 5 presenta sus características técnicas.

Los datos son recolectados por escaneos del volumen que rodea al radar a 120, 240 y 480 kilómetros con giros de la antena de 360 grados en forma horizontal, iniciando con una elevación de 0,5 grados y aumentando el ángulo de elevación 11 veces más: 0,9º; 1,3º; 1,9º; 2,3º; 3,0º; 3,5º; 5,0º; 6,9º; 9,1º; 11,8º y 15,1º [34] [115]. Los escaneos están configurados para ocurrir cada 10 minutos, toman datos cuya unidad de muestreo es de 1 km^2 y 1º [34] [115] y se almacenan en archivos separados llamados volúmenes [34] [115] [116].

Tabla 5. Características técnicas específicas del radar meteorológico de la EEA Anguil (Adaptado de [115])

Modelo:	Meteor 600C	*Marca:*	Gematronik
Banda:	C	*Longitud de onda*	5,4 cm
Frecuencia	5,6 Ghz	*Distancia típica de operación:*	200 km
Soft. Mantenimiento:	Ravis®	*Distancia no ambigua máxima:*	125 – 500 km
Soft. Usuario Meteorológico:	Rainbow®5	*Opción de segundo trip:*	250 – 1000 km
Salida de datos polarización simple (SP)		*Salida de datos polarización dual (DP)*	
Z, velocidad radial, anchura de espectro.		Z_{DR}, Phi_{DP}, K_{DP}, Rho_{HV}, simultáneamente. L_{DR} a petición.	

Estos volúmenes contienen: el valor de la variable medida (Z, Z_{DR}, Rho_{HV}, Phi_{DP} y K_{DP} [34]) para cada elevación y la posición de cada unidad muestreada (1 km^3,figura 5d) con respecto del radar en coordenadas polares [116]. Son almacenados en un servidor in situ del radar, al cual solo se puede acceder por vía remota. La figura 14 presenta un listado de ejemplo de volúmenes y el detalle del formato del nombre que lo identifica de forma única.

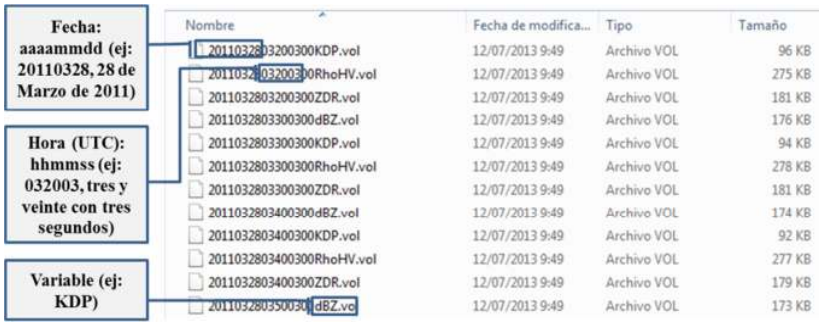


Figura 14. Ejemplos de volúmenes del radar de la EEA Anguil (Fuente: elaboración propia).

El formato de los volúmenes fue definido por el fabricante del radar y cuenta con una sección XML (eXtensible Markup Language) que almacena datos del contexto de la toma de datos y una sección en formato binario con compresión que contiene el dato plano y también presenta una estructura al estilo XML [116][117]. Esta situación impone una restricción al uso de herramientas estándares para manipular XML [116], por lo que el acceso a estos datos del radar se dificulta, teniendo que utilizar el software nativo provisto por el fabricante el cual es propietario y de un alto costo y no estaba disponible al inicio de este trabajo. Como se mencionó anteriormente se desarrollaron un conjunto de programas, basados en [118], que permiten descargar los archivos, transformarlos a dos formatos estándar (ASCII[17] y GeoTIFF[18]) y procesarlos (Figura 15).

Para desarrollar este software se utilizó Python 2.7 y las librerías externas PyQt4, Numpy, lxml, Gdal y Pyodbc. Como IDE[19] de desarrollo de software se utilizó Geany y la visualización de las imágenes se realizó con QGIS. El software se puede ejecutar en plataforma Linux o Windows.

El primer paso fue obtener los archivos del radar, para esto se usó el script de Python *ftpRADAR.py* indicando el radar a utilizar (Anguil), la serie de fechas (ver tabla 3) y el rango en km (240). El script se conecta con el servidor del radar especificado y descarga los archivos de todas las variables disponibles para el rango y fechas indicados.

[17] **A**merican **S**tandard **C**ode for **I**nformation **I**nterchange. Sistema de codificación de caracteres alfanuméricos que asigna un número del 0 al 127 a cada letra, número o carácter especial recogidos; el ASCII extendido permite hasta 256 caracteres distintos [119].

[18] Es un formato de intercambio basado en Tagged Image File Format (TIFF) para imágenes raster georeferenciada. Es un estándar de metadatos y es ampliamente soportado por aplicaciones SIG y de manipulación de imágenes [120] [121].

[19] IDE: entorno de desarrollo integrado.

El segundo paso fue convertir estos archivos al formato estándar ASCII: con el proceso por lotes *Proc-vol*[20] se recorre un directorio con archivos los volúmenes del radar y se pasan como parámetro al proceso *Vol.py*. Este script descomprime los datos binarios, realiza el cálculo del dato "crudo" de las variables Z, Z_{DR} y Rho_{HV} (figura 16a), transforma las coordenadas polares a geográficas y finalmente genera un archivo de texto por cada elevación con el formato: latitud, longitud y valor de la variable (figura 16b).

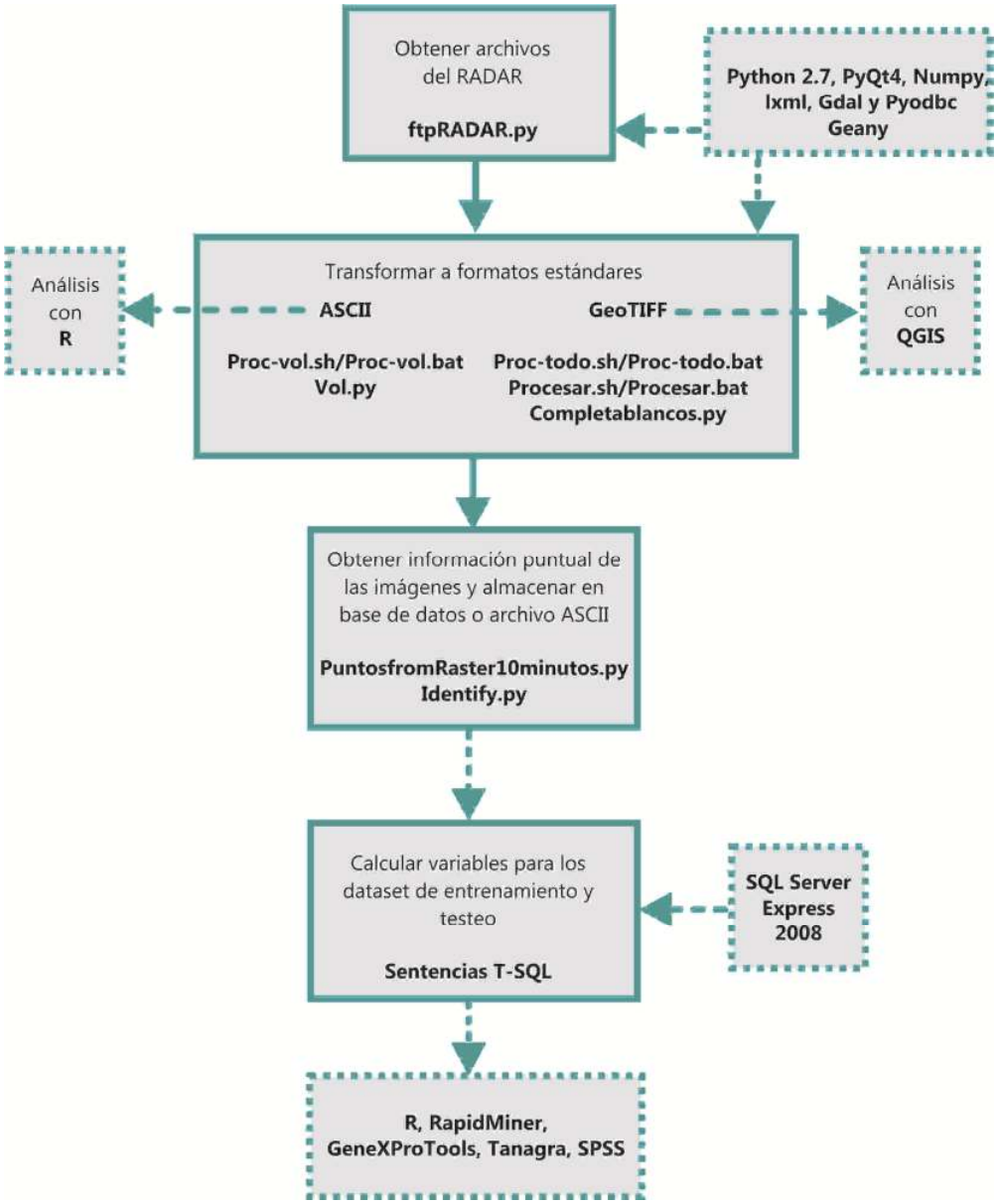


Figura. 15. Esquema de programas para el procesamiento de los datos del radar (Disponible en: https://github.com/INTA-Radar)

[20] Los procesos por lotes tienen extensión .sh para la plataforma Linux y .bat para la plataforma Windows.

Si la toma de datos no tuvo inconvenientes, en un día completo (00:00 h a 23:50 h), se generan 8.640 archivos solo para el rango de 240 kilómetros. Para este trabajo se procesaron 7.902[21] volúmenes, generando 94.026 archivos ASCII que ocupan 459 Gb y consumiendo 270 horas[22] de cómputo.

El programa *Vol.py* controla una serie de problemas que se presentaron en el manejo de los volúmenes: a) fechas incompletas (faltan tomas de datos, en ocasiones, corresponden al horario de mayor actividad de la tormenta), b) archivos vacíos, c) archivos que cambian el formato XML determinado.

$$\frac{depth - dmi}{dma - dmi} x\ ((db\_\max - db\_\min) + db\_\min) \quad (5)$$

```
def get_depth_as_rho(depth):
    db_min = 0.0
    db_max = 1.0
    dmi = 1.0
    dma = 255.0
    return round(((float(depth) - dmi)/(dma - dmi))*(db_max - (db_min)) + db_min,3)
```

Figura 16. a) Ejemplo de implementación de (5): función que realiza el cálculo del valor real de Rho_{HV}. Los valores de las variables db_min, db_max, dmi y dma se obtuvieron de [117].

```
1  lon lat dbz
2  -63.990067 -36.539683 49.500000
3  -63.992167 -36.543851 10.000000
4  -63.994267 -36.548019 17.500000
5  -63.996367 -36.552188 1.500000
6  -63.998467 -36.556356 0.000000
7  -64.000568 -36.560524 6.500000
8  -64.002668 -36.564692 12.500000
9  -64.004769 -36.568860 12.500000
```

Figura 16. b) Ejemplo del contenido de un archivo ASCII generado desde un volumen de la variable Z.

Estos archivos ASCII se leyeron y analizaron con el software R para verificar que los valores de las variables registradas por el radar estén dentro de los rangos esperados. Para el análisis se revisó la primera elevación en las 24 horas ya que al ser la más cercana a la superficie es la que mejor representa lo que puede precipitar a nivel del suelo [9] [31] [33] [47] [59] [66]. Las variables presentan datos dentro de los rangos válidos indicados en [117].

El tercer paso, es convertir los archivos de texto al estándar raster[23] GeoTIFF, el cual permite trabajar los datos en Sistemas de Información Geográfico (SIG) y contar con

[21] Estos volúmenes corresponden a las variables: Z, Z_{DR} y Rho_{HV}.
[22] Procesado con una notebook con procesador Intel Core i7, 8 Mb de RAM y Windows 7 de 64 bits.
[23] Un archivo raster consta de una matriz de celdas (o píxeles) organizados en filas y columnas (rejilla) en la que cada celda contiene un valor que representa la información de fenómenos del mundo real [122].

una representación similar a las imágenes generadas por el software provisto por el fabricante de los radares. El formato raster representa adecuadamente los datos del radar con imágenes de 1 km^2 de resolución, en coordenadas geográficas (latitud y longitud) y Datum WGS84[24]. Este formato facilita la extracción de información de diferentes áreas del radar. Para generar las imágenes se utiliza *Proc-todo* que recorre el directorio de los archivos ASCII generados en el paso anterior y los pasa como parámetro al proceso por lotes *Procesar*. Este algoritmo recibe cada archivo de texto y genera una imagen de 487x505 píxeles por medio de la utilidad *gdal_rasterize* de la librería GDAL (versión 1.11.0), que convierte un archivo vectorial, en este caso de puntos, a una imagen raster. Se generó una plantilla GeoTIFF con el sistema de coordenadas que coincida con el de los archivos ASCII, para realizar una correcta proyección de los datos [124]. Como el haz emitido por el radar se va ensanchando a medida que se aleja, la toma de los datos no es completamente uniforme (las muestras más alejadas abarcan un volumen mayor a 1 km^3 y las más cercanas un volumen menor [67]), por lo que en la transformación de puntos a raster pueden quedar pixeles sin datos. Para completar toda la matriz correctamente, se ejecuta el script *Completablancos.py* el cual rellena todos los pixeles del área del radar interpolando el dato máximo correspondiente [25] (figuras 17 y 18).

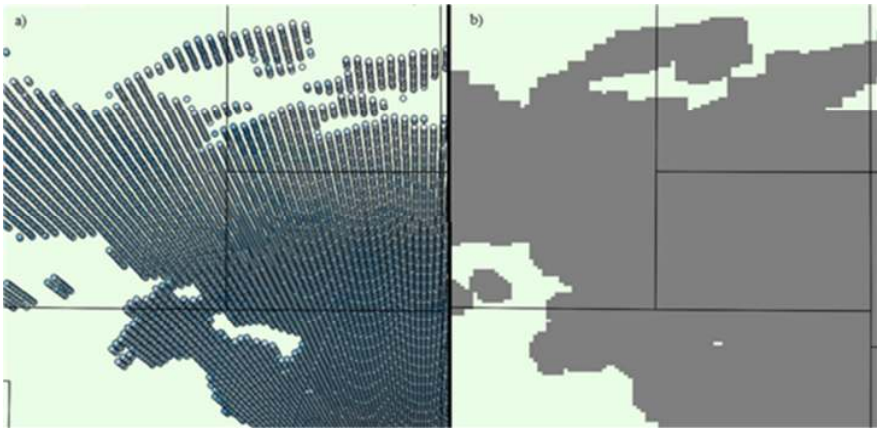


Figura 17. a) Datos del radar en formato punto como están contenido en el archivo ASCII. b) Los mismos datos convertidos a formato raster donde cada dato ocupa un pixel de 1km^2

[24] Sistema de coordenadas geográficas mundial que permite localizar cualquier punto de la Tierra (sin necesitar otro de referencia) por medio de tres unidades dadas [123].
[25] El script permite seleccionar si se interpola con el dato máximo o mínimo.

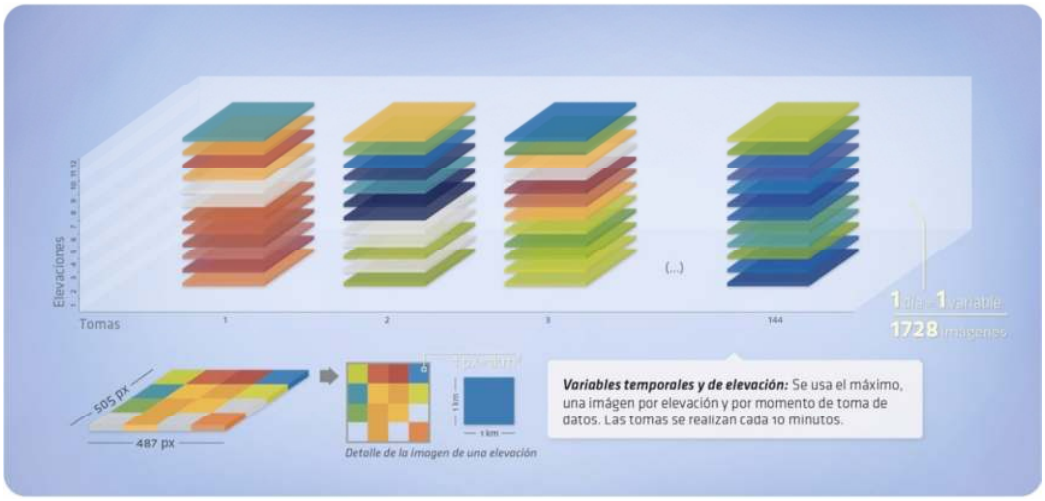


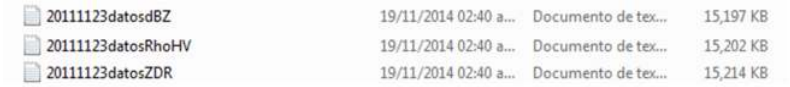
Figura 18. Detalle de la generación de las imágenes correspondientes a cada elevación, cantidad de archivos y resolución espacial de las imágenes.

Para este trabajo se generaron 94.026 archivos GeoTIFF, en 10.000 horas de procesamiento[26] aproximadamente y que contienen 23.124.284.310 datos de radar ocupando 182 Gb aproximadamente.

El cuarto paso consiste en obtener los valores de las tres variables procesadas para la serie de puntos de verdad de campo recolectados sobre cada tormenta. Para esta tarea se generó el script *PuntosfromRaster10minutos.py* que obtiene los valores del pixel correspondiente a cada par de coordenadas geográficas pasadas como parámetros, desde las imágenes de las tomas cada 10 minutos. La información recuperada se puede grabar en una base de datos o en un archivo ASCII. Utiliza el script *Identify.py* el cual convierte las coordenadas geográficas a coordenadas en pixel de la imagen raster a procesar y devuelve el valor correspondiente al pixel determinado (figura 19). Para el presente trabajo se recuperaron 14.544.332 datos correspondientes a 3.018 lotes. Todos los programas realizados, su código fuente y una explicación de cómo instalarlos y usarlos, se pueden descargar desde https://github.com/INTA-Radar.

[26] Realizado con 4 máquinas con 8 MB de RAM, Procesador Intel Core i7 y Windows 7 de 64 bits. En promedio, cada imagen se procesa en 7 minutos, con un máximo de 21 minutos y un mínimo de 3.

a)



20111123datosdBZ	19/11/2014 02:40 a...	Documento de tex...	15,197 KB
20111123datosRhoHV	19/11/2014 02:40 a...	Documento de tex...	15,202 KB
20111123datosZDR	19/11/2014 02:40 a...	Documento de tex...	15,214 KB

b)

	Fecha	IdEvento	IdLocXEvento	lat	lon	granizo	danio	pordanio	Elevacion	Horario	dBZ
2	20111123	70	192	-35.912863	-64.288731	True	NULL	0	1.0000	-99.0	
3	20111123	70	192	-35.912863	-64.288731	True	NULL	0	11.0000	-99.0	
4	20111123	70	192	-35.912863	-64.288731	True	NULL	0	13.0000	-99.0	
5	20111123	70	192	-35.912863	-64.288731	True	NULL	0	15.0000	-99.0	
6	20111123	70	192	-35.912863	-64.288731	True	NULL	0	17.0000	-99.0	
7	20111123	70	192	-35.912863	-64.288731	True	NULL	0	19.0000	-99.0	
8	20111123	70	192	-35.912863	-64.288731	True	NULL	0	21.0000	-99.0	
9	20111123	70	192	-35.912863	-64.288731	True	NULL	0	23.0000	-99.0	
10	20111123	70	192	-35.912863	-64.288731	True	NULL	0	3.0000	-99.0	
11	20111123	70	192	-35.912863	-64.288731	True	NULL	0	5.0000	-99.0	
12	20111123	70	192	-35.912863	-64.288731	True	NULL	0	7.0000	-99.0	

Figura 19.a) Ejemplo de los archivos ASCII generados por el script PuntosfromRaster10minutos.py. b) Ejemplo del contenido de los archivos ASCII con los datos de cada variable del RADAR por localización.

2.2.3. Generación de los DataSet

Se generaron datasets diferentes de acuerdo a los objetivos de análisis. Se agruparon las variables y seleccionaron los casos de acuerdo al tipo, a la compleción de los datos en cada fecha y al target buscado.

2.2.3.1. Cálculo de variables

Los datos de campo tienen una escala temporal diaria, por lo tanto las variables del radar fueron resumidas en valores de 24 horas por medio de diferentes cálculos[27]. Estos resúmenes permiten utilizar juntos los datos de campo y los de radar al llevarlos a la misma escala temporal [69] [97]. Los cálculos realizados por medio de consultas TSQL son: máximo (MAX) mínimo (MIN), promedio (AVG) y total (SUM) de Z, Rho_{HV} y Z_{DR}. Estos cálculos se hacen sobre los valores de todas las elevaciones y sobre los valores de la primera elevación solamente. Esta configuración de los datos permitirá analizar si es factible obtener una buena predicción del target utilizando una sola elevación, haciendo que el modelo sea más rápido y sencillo al necesitar menos datos. Se selecciona la primera elevación porque al ser la más cercana a la superficie es la que mejor representa lo que puede precipitar a nivel del suelo [9] [31] [33] [47] [59] [66].

En el caso de las variables H_{DR} y E estos cálculos se aplicaron al resultado de usar la fórmula 4 y 6 respectivamente.

[27] Las variables relacionadas con Phi_{DP} y K_{DP} no se presentan por no utilizarse en este trabajo, sin embargo la base de datos y los script generados prevén la incorporación de las mismas cuando se realice la programación de su lectura.

$$E = \left(\sum_{i=base}^{i=tope} 5 \times 10^{-6} \times 10^{0,084 \times Z_i} \times W(Z_i) \right) . t \quad (6)$$

Dónde:

i= 1 a 12 elevaciones

t= 600 segundos[28]

$$W(Z) = \begin{cases} 0 & para\ Z \leq 40 \\ \dfrac{Z - 40}{40 - 55} & para\ 40 < Z < 55 \\ 1 & para\ Z \geq 55 \end{cases}$$

También se generaron una serie de variables que intentan capturar la presencia y altura de ecos fuertes. Se calcularon utilizando las sentencia cuenta (COUNT) y CASE de T-SQL sobre la variable Z.

Debido a la forma de la toma de datos y de las tormentas, la aparición del valor *missing* (perdido) en las variables del radar, ocurre cuando no hay tormenta. Esta particularidad hace imposible reemplazar ese valor por algún otro, ya que en ese lugar y momento, no había que medir. Para representar esta particularidad se representó el valor perdido con el número "-99.00"[29]. Los cálculos de máximo, mínimo, promedio y total de Z, Z_{DR} y Rho_{HV} se calcularon con todos los casos por un lado y descartando los casos que tuvieran valores perdidos por otro. De esta manera ninguna variable presenta valores perdidos y por ende se puede usar con las técnicas seleccionadas. Aquellas calculadas sin los "-99.00" también permiten comparar su comportamiento con la bibliografía aportando al objetivo de contribuir con la caracterización de las variables de doble polarización de banda C, para la identificación de granizo y daño en cultivos en la región de estudio. Las variables calculadas con el "-99.00" permite reflejar en los datos el momento y lugar donde la tormenta no registró valores y estudiar cómo influye esta particularidad en el rendimiento de los modelos. Las tablas 7 a 11 muestran el detalle de todas las variables creadas.

[28] Es la cantidad de segundos contenidos en los 10 minutos que transcurren entre cada toma de datos del radar.

[29] El -99.00 es un valor estándar para representar el "Sin Dato" o el "NULL" en los sistemas de información agrometerológica de INTA. Cumple con el requisito de estar muy alejado de los valores válidos de las diferentes medidas agroclimáticas, incluyendo los datos de radar.

Tabla 7. Variables calculadas a partir de Z

Nº	Nombre	Descripción	Cálculo	Target
1	MxDbz1	*Máximo* de Z ocurrido durante las *24 horas* en la *primera* elevación.	MAX(Z) where Elevacion=1 and Z <>-99.0	Granizo Daño
2	MnDbz1	*Mínimo* de Z ocurrido durante las *24 horas* en la *primera* elevación.	MIN(Z) where Elevacion=1 and Z <>-99.0	Granizo Daño
3	AvDbz1	*Promedio* de Z ocurrido durante las *24 horas* en la *primera* elevación.	AVG(Z) where Elevacion=1 and Z <>-99.0	Granizo Daño
4	TotDbz1	*Total* de Z ocurrido durante las *24 horas* en la *primera* elevación.	SUM(Z) where Elevacion=1 and Z <>-99.0	Granizo Daño
5	MxDbzT	*Máximo* de Z ocurrido durante las *24 horas* en *todas* las elevaciones.	MAX(Z) where Z <>-99.0	Granizo Daño
6	MnDbzT	*Mínimo* de Z ocurrido durante las *24 horas* en *todas* las elevaciones.	MIN(Z) where Z <>-99.0	Granizo Daño
7	AvDbzT	*Promedio* de Z ocurrido durante las *24 horas* en *todas* las elevaciones.	AVG(Z) where Z <>-99.0	Granizo Daño
8	TotDbzT	*Total* de Z ocurrido durante las *24 horas* en *todas* las elevaciones.	SUM(Z) where Z <>-99.0	Granizo Daño
9	MxEWt1	*Máximo* de E ocurrido durante las *24 horas* en la *primera* elevación.	MAX(E) where Elevacion=1 and E <>-99.0	Daño
10	MnEWt1	*Mínimo* de E ocurrido durante las *24 horas* en la *primera* elevación.	MIN(E) where Elevacion=1 and E <>-99.0	Daño
11	AvEWt1	*Promedio* de E ocurrido durante las *24 horas* en la *primera* elevación.	AVG(E) where Elevacion=1 and E <>-99.0	Daño
12	TotEWt1	*Total* de E ocurrido durante las *24 horas* en la *primera* elevación.	SUM(E) where Elevacion=1 and E <>-99.0	Daño
13	MxEWtT	*Máximo* de E ocurrido durante las *24 horas* en *todas* las elevaciones.	MAX(E) where E <>-99.0	Daño
14	MnEWtT	*Mínimo* de E ocurrido durante las *24 horas* en la *todas* las elevaciones.	MIN(E) where E <>-99.0	Daño
15	AvEWtT	*Promedio* de E ocurrido durante las *24 horas* en *todas* las elevaciones.	AVG(E) where E <>-99.0	Daño
16	TotEWtT	*Total* de E ocurrido durante las *24 horas* en *todas* las elevaciones.	SUM(E) where E <>-99.0	Daño
17	Dbz01	Indica la presencia de ecos mayor a *0 dBZ* en la *primera* elevación.	CASE WHEN Z = 0 THEN 1 ELSE 0 where Elevacion=1	Daño
18	Dbz451	Indica la presencia de ecos mayor o igual a *45 dBZ* en la *primera* elevación.	CASE WHEN Z > 44 THEN 1 ELSE 0 where Elevacion=1	Daño
19	Dbz501	Indica la presencia de ecos mayor o igual a *50 dBZ* en la *primera* elevación.	CASE WHEN Z > 49 THEN 1 ELSE 0 where Elevacion=1	Daño

20	Dbz551	Indica la presencia de ecos mayor o igual a *55 dBZ* en la *primera* elevación.	CASE WHEN Z > 54 THEN 1 ELSE 0 where Elevacion=1	Daño
21	Dbz601	Indica la presencia de ecos mayor o igual a *60 dBZ* en la *primera* elevación.	CASE WHEN Z > 59 THEN 1 ELSE 0 where Elevacion=1	Daño
22	Dbz0T	Indica la presencia de ecos mayor a *0 dBZ* en *todas* las elevaciones.	If Z > 0 = 1 else = 0	Daño
23	Dbz45T	Indica la presencia de ecos mayor o igual a *45 dBZ* en *todas* las elevaciones.	If Z >= 45 = 1 else = 0	Daño
24	Dbz50T	Indica la presencia de ecos mayor o igual a *50 dBZ* en *todas* las elevaciones.	If Z >= 50 = 1 else = 0	Daño
25	Dbz55T	Indica la presencia de ecos mayor o igual a *55 dBZ* en *todas* las elevaciones.	If Z >= 55 = 1 else = 0	Daño
26	Dbz60T	Indica la presencia de ecos mayor o igual a *60 dBZ* en *todas* las elevaciones.	If Z >= 60 = 1 else = 0	Daño
27	Dbz0c	Indica la cantidad de elevaciones que tienen Z diferente de 0 dBZ.	COUNT (Z) where Z >0	Daño
28	Dbz45c	Indica la cantidad de elevaciones que tienen Z mayor o igual a 45 dBZ.	COUNT (Z) where Z >=45	Daño
29	Dbz50c	Indica la cantidad de elevaciones que tienen Z mayor o igual a 50 dBZ.	COUNT (Z) where Z >=50	Daño
30	Dbz55c	Indica la cantidad de elevaciones que tienen Z mayor o igual a 55 dBZ.	COUNT (Z) where Z >=55	Daño
31	Dbz60c	Indica la cantidad de elevaciones que tienen Z mayor o igual a 60 dBZ.	COUNT (Z) where Z >=60	Daño
32	MxDbz1C99	*Máximo* de Z ocurrido durante las *24 horas* en la *primera* elevación.	MAX(Z) where Elevacion=1	Granizo
33	MnDbz1C99	*Mínimo* de Z ocurrido durante las *24 horas* en la *primera* elevación.	MIN(Z) where Elevacion=1	Granizo
34	AvDbz1C99	*Promedio* de Z ocurrido durante las *24 horas* en la *primera* elevación.	AVG(Z) where Elevacion=1	Granizo
35	TotDbz1C99	*Total* de Z ocurrido durante las *24 horas* en la *primera* elevación.	SUM(Z) where Elevacion=1	Granizo
36	MxDbzTC99	*Máximo* de Z ocurrido durante las *24 horas* en *todas* las elevaciones.	MAX(Z)	Granizo
37	MnDbzTC99	*Mínimo* de Z ocurrido durante las *24 horas* en *todas* las elevaciones.	MIN(Z)	Granizo
38	AvDbzTC99	*Promedio* de Z ocurrido durante las *24 horas* en *todas* las elevaciones.	AVG(Z)	Granizo
39	TotDbzTC99	*Total* de Z ocurrido durante las *24 horas* en *todas* las elevaciones.	SUM(Z)	Granizo

Tabla 8. Variables calculadas a partir de Z_{DR} y Z

40	MxHDR1	*Máximo* de H_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	MAX(HDR) where Elevacion=1 and Z <> -99.0	Daño
41	MnHDR1	*Mínimo* de H_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	MIN(HDR) where Elevacion=1 and Z <> -99.0	Daño
42	AvHDR1	*Promedio* de H_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	AVG(HDR) where Elevacion=1 and Z <> -99.0	Daño
43	TotHDR1	*Total* de H_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	SUM(HDR) where Elevacion=1 and Z <> -99.0	Daño
44	MxHDRT	*Máximo* de H_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	MAX(HDR)	Daño
45	MnHDRT	*Mínimo* de H_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	MIN(HDR) where Z <> -99.0	Daño
46	AvHDRT	*Promedio* de H_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	AVG(HDR) where Z <> -99.0	Daño
47	TotHDRT	*Total* de H_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	SUM(HDR) where Z <> -99.0	Daño

Tabla 9. Variables calculadas a partir de Z_{DR}.

Nº	**Nombre**	**Descripción**	**Cálculo**	**Target**
48	MxZDR1	*Máximo* de Z_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	MAX(ZDR) where Elevacion=1 and ZDR <> -99.0	Granizo Daño
49	MnZDR1	*Mínimo* de Z_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	MIN(ZDR) where Elevacion=1 and ZDR <> -99.0	Granizo Daño
50	AvZDR1	*Promedio* de Z_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	AVG(ZDR) where Elevacion=1 and ZDR <> -99.0	Granizo Daño
51	TotZDR1	*Total* de Z_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	SUM(ZDR) where Elevacion=1 and ZDR <> -99.0	Granizo Daño
52	MxZDRT	*Máximo* de Z_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	MAX(ZDR) where ZDR <> -99.0	Granizo Daño
53	MnZDRT	*Mínimo* de Z_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	MIN(ZDR) where ZDR <> -99.0	Granizo Daño
54	AvZDRT	*Promedio* de Z_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	AVG(ZDR) where ZDR <> -99.0	Granizo Daño
55	TotZDRT	*Total* de Z_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	SUM(ZDR) where ZDR <> -99.0	Granizo Daño
56	MxZDR1C99	*Máximo* de Z_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	MAX(ZDR) where Elevacion=1	Granizo

57	MnZDR1C99	*Mínimo* de Z_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	MIN(ZDR) where Elevacion=1	Granizo
58	AvZDR1C99	*Promedio* de Z_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	AVG(ZDR) where Elevacion=1	Granizo
59	TotZDR1C99	*Total* de Z_{DR} ocurrido durante las *24 horas* en la *primera* elevación.	SUM(ZDR) where Elevacion=1	Granizo
60	MxZDRTC99	*Máximo* de Z_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	MAX(ZDR)	Granizo
61	MnZDRTC99	*Mínimo* de Z_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	MIN(ZDR)	Granizo
62	AvZDRTC99	*Promedio* de Z_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	AVG(ZDR)	Granizo
63	TotZDRTC99	*Total* de Z_{DR} ocurrido durante las *24 horas* en *todas* las elevaciones.	SUM(ZDR)	Granizo

Tabla 10. Variables calculadas a partir de Rho_{HV}.

Nº	**Nombre**	**Descripción**	**Cálculo**	**Target**
64	MxRho1	*Máximo* de Rho_{HV} ocurrido durante las *24 horas* en la *primera* elevación.	MAX(RhoHV) where Elevacion=1 and Rho <> -99.0	Granizo Daño
65	MnRho1	*Mínimo* de Rho_{HV} ocurrido durante las *24 horas* en la *primera* elevación.	MIN(RhoHV) where Elevacion=1 and Rho <> -99.0	Granizo Daño
66	AvRho1	*Promedio* de Rho_{HV} ocurrido durante las *24 horas* en la *primera* elevación.	AVG(RhoHV) where Elevacion=1 and Rho <> -99.0	Granizo Daño
67	TotRho1	*Total* de Rho_{HV} ocurrido durante las *24 horas* en la *primera* elevación.	SUM(RhoHV) where Elevacion=1 and Rho <> -99.0	Granizo Daño
68	MxRhoT	*Máximo* de Rho_{HV} ocurrido durante las *24 horas* en *todas* las elevaciones.	MAX(RhoHV) where Rho <> -99.0	Granizo Daño
69	MnRhoT	*Mínimo* de Rho_{HV} ocurrido durante las *24 horas* en *todas* las elevaciones.	MIN(RhoHV) where Rho <> -99.0	Granizo Daño
70	AvRhoT	*Promedio* de Rho_{HV} ocurrido durante las *24 horas* en *todas* las elevaciones.	AVG(RhoHV) where Rho <> -99.0	Granizo Daño
71	TotRhoT	*Total* de Rho_{HV} ocurrido durante las *24 horas* en *todas* las elevaciones.	SUM(RhoHV) where Rho <> -99.0	Granizo Daño
72	MxRho1C99	*Máximo* de Rho_{HV} ocurrido durante las *24 horas* en la *primera* elevación.	MAX(RhoHV) where Elevacion=1	Granizo
73	MnRho1C99	*Mínimo* de Rho_{HV} ocurrido durante las *24 horas* en la *primera*	MIN(RhoHV) where Elevacion=1	Granizo

		elevación.		
74	AvRho1C99	*Promedio* de Rho_{HV} ocurrido durante las *24 horas* en la *primera* elevación.	AVG(RhoHV) where Elevacion=1	Granizo
75	TotRho1C99	*Total* de Rho_{HV} ocurrido durante las *24 horas* en la *primera* elevación.	SUM(RhoHV) where Elevacion=1	Granizo
76	MxRhoT	*Máximo* de Rho_{HV} ocurrido durante las *24 horas* en *todas* las elevaciones.	MAX(RhoHV)	Granizo
77	MnRhoTC99	*Mínimo* de Rho_{HV} ocurrido durante las *24 horas* en *todas* las elevaciones.	MIN(RhoHV)	Granizo
78	AvRhoTC99	*Promedio* de Rho_{HV} ocurrido durante las *24 horas* en *todas* las elevaciones.	AVG(RhoHV)	Granizo
79	TotRhoTC99	*Total* de Rho_{HV} ocurrido durante las *24 horas* en *todas* las elevaciones.	SUM(RhoHV)	Granizo

Tabla 11. Variables que caracterizan el cultivo.

Nº	**Nombre**	**Descripción**	**Valores**	**Target**
80	Cultivo	Contiene el nombre del cultivo	Gr (Girasol), So (Soja), Sg (Sorgo), Mt (Monte), Mz (Maíz), Av (Avena), Tr (Trigo), Ce (Centeno), CN (Campo Natural), NULL	Daño
81	Fenologia	Contiene el estado fenológico en que se encontraba el cultivo a la fecha de la tormenta	R (reproductivo), M (madurez), V (vegetativo), SD (sin dato)	Daño
82	TipoCultivo	Cosecha fina o gruesa, determinada según el cultivo	Fina, Gruesa, SD (sin dato)	Daño

2.2.3.2. DataSet Target Granizo

2.2.3.2.1. Agrupamiento de variables

El principal interés es analizar el poder clasificatorio y el comportamiento de las variables de doble polarización ante la presencia de granizo. Para esto se generan dos datasets:

1) *dsVariablesPolarimétricas1Ele:* variables de resumen de Z, Z_{DR} y Rho_{HV} para la primera elevación y sin valores perdidos (variables 1 a 4, 48 a 51 y 64 a 67 en las tablas 7 a 10).

2) *dsVariablesPolarimétricasTEle:* variables de resumen de Z, Z_{DR} y Rho_{HV} para todas las elevaciones y sin valores perdidos (variables 5 a 8, 52 a 55 y 68 a 71 en las tablas 7 a 10).

También interesa analizar si se puede generar un modelo donde solo se necesite Z como variable de entrada. Este modelo se podría extender a otros radares de banda C, sean o no, de doble polarización. Se generan dos datasets:

3) *dsVariablesDerivadasdeZ1Ele:* variables derivadas de Z, para la primera elevación y sin valores perdidos (variables 1 a 4, 9 a 12 y 17 a 21 en las tablas 7 a 10).
4) *dsVariablaesDerivadasdeZTEle:* variables derivadas de Z, para todas las elevaciones y sin valores perdidos (variables 5 a 8, 13 a 16 y 22 a 31 en las tablas 7 a 10).

Finalmente, importa conocer cómo influye en la clasificación representar en los datos la falta de tormenta en un momento y lugar específico. Para este fin, se generan dos dataset:

5) *dsVariablePolarimetricasTodo1Ele:* variables de resumen de Z, Z_{DR} y Rho_{HV} para la primera elevación y con valores perdidos representados con -99.00 (variables 32 a 35, 56 a 59 y 72 a 75 en las tablas 7 a 10).
6) *dsVariablePolarimetricasTodoTEle:* variables de resumen de Z, Z_{DR} y Rho_{HV} para todas las elevaciones y con valores perdidos representados con -99.00 (variables 36 a 39, 60 a 63 y 76 a 79 en las tablas 7 a 10).

2.2.3.2.2. Filtrado de casos

De los seis conjuntos de datos, en los cuatro primeros se descartaron los casos que tuvieran valores perdidos y los casos correspondientes a las fechas que les faltaban volúmenes (ver tabla 3). La figura 20 y tabla 12 presentan el detalle de la distribución de los casos.

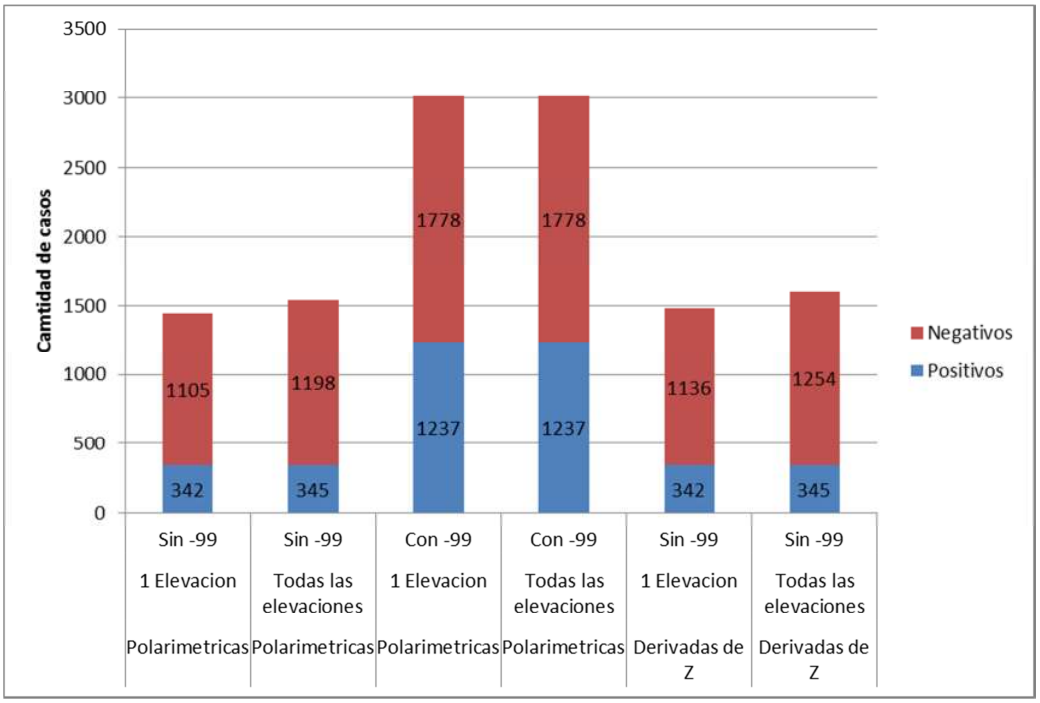


Figura 20. Distribución de los casos en los cuatro dataset para modelar el target Granizo

Tabla 12. Características de los datos contenidos en los dataset para modelar el target Granizo.

Target: Granizo							
Variables Polarimétricas							
1 Elevación				Todas las Elevaciones			
Nombre: *dsVariablesPolarimetricas1Ele*				Nombre: *dsVariablesPolarimétricasTEle*			
Variables	Casos Positivos	Casos Negativos	Total	Variables	Casos Positivos	Casos Negativos	Total
12	342 (24.1%)	1.105 (77.87%)	1.447	12	345 (22.36%)	1.198 (77.64%)	1.543
Nombre: *dsVariablePolarimetricasTodo1Ele*				Nombre: *dsVariablePolarimetricasTodoTEle*			
Variables	Casos Positivos	Casos Negativos	Total	Variables	Casos Positivos	Casos Negativos	Total
12	1.237 (41.03%)	1.778 (58.97%)	3.015	12	1.237 (41.03%)	1.778 (58.97%)	3.015
Variables derivadas de Z							
1 Elevación				Todas las Elevaciones			
Nombre: *dsVariablesDerivadasZ1Ele*				Nombre*: dsVariablesDerivadasZ1Ele*			
Variables	Casos Positivos	Casos Negativos	Total	Variables	Casos Positivos	Casos Negativos	Total
13	342 (23.14%)	1.136 (76.86%)	1.478	18	345 (21.58%)	1.254 (78.42%)	1.599
Ninguno de los dataset presenta valores perdidos (o porque se descartan o porque se reemplazan con -99.0).							

2.2.3.3. DataSet Target Daño.

2.2.3.3.1. Agrupamiento de variables

Para analizar el poder de clasificación y el comportamiento de las variables de doble polarización para clasificar el daño realizado por el granizo en los cultivos se generaron cuatro datasets:

7) *dsVarPol1EleVolCompSinDatosCultivos:* variables de resumen de Z, Z_{DR} y Rho_{HV} para la primera elevación y sin valores perdidos (variables 1 a 4, 48 a 51 y 64 a 67 en las tablas 7 a 10).
8) *dsVarPolTEleVolCompSinDatosCultivos:* variables de resumen de Z, Z_{DR} y Rho_{HV} para todas las elevaciones y sin valores perdidos (variables 5 a 8, 52 a 55 y 68 a 71 en las tablas 7 a 10).
9) *dsVarPol1EleVolCompSinDatosCultivosHDR:* variables de resumen de Z, Z_{DR}, Rho_{HV} y H_{DR} para la primera elevación y sin valores perdidos/ (variables 1 a 4, 40 a 43, 48 a 51 y 64 a 67 en las tablas 7 a 10).
10) *dsVarPolTEleVolCompSinDatosCultivosHDR:* variables de resumen de Z, Z_{DR}, Rho_{HV} y H_{DR} para la primera elevación y sin valores perdidos (variables 5 a 8, 44 a 47, 52 a 55 y 68 a 71 en las tablas 7 a 10).

Importa conocer el aporte de algunas variables relacionadas con los cultivos mencionadas en los antecedentes, para esto se generaron dos data sets:

11) *dsVarPol1EleVolCompConDatosCultivos:* variables de resumen de Z, Z_{DR}, Rho_{HV} y E para la primera elevación, variables de cultivos y sin valores perdidos (variables 1 a 4, 9 a 12, 48 a 51, 64 a 67 y 80 a 82 en las tablas 7 a 11).
12) *dsVarPolTEleVolCompConDatosCultivos:* variables de resumen de Z, Z_{DR}, Rho_{HV} y E para todas las elevaciones, variables de cultivos y sin valores perdidos (variables 5 a 8, 13 a 16, 52 a 55, 68 a 71 y 80 a 82 en las tablas 7 a 11).

Se realizó un reagrupamiento de las clases del target teniendo en cuenta que en los análisis estadísticos básicos y de ANOVA de las variables creadas, las clases Leve y Moderado (1% al 50%) no presentaban diferencias significativas. La clase Severo (51% al 74%) era diferenciada por algunas variables, mientras que si presentaban

diferencias significativas en la mayoría de las variables las clases Sin Daño (0%) y Grave (75% a 100%). Teniendo en cuenta esta evidencia y la cantidad de casos disponibles, el target quedó de la siguiente manera: Sin Daño (0% de daño), Menos50 (1% a 49% de daño) y Mas50 (50% a 100% de daño). Los análisis estadísticos se presentan en las secciones 3.1.1. y 3.2.1. Para generar estos modelos se diseñó un dataset:

13) *dsVarPolTEleVolCompHDRSinDatosCultivos:* variables de resumen de Z, $Z_{DR,}$ Rho_{HV} y H_{DR}, para todas las elevaciones y sin valores perdidos (variables 5 a 8, 44 a 47, 52 a 55 y 68 a 71 en las tablas 7 a 10).

2.2.3.3.2. Filtrado de casos

Cuando se agregan las variables relacionadas con el cultivo la cantidad de casos disminuye considerablemente, por lo que a los dataset 7, 8, 11, 12 y 13 se le agregaron los casos de las fechas que no tenían todos los volúmenes completos, para aumentar la cantidad de casos a procesar:

14) *dsVarPol1EleVolTodosSinDatosCultivos*: ídem dataset 7.

15) *dsVarPolTEleVolTodosSinDatosCultivos*: ídem dataset 8.

16) *dsVarPol1EleVolTodosConDatosCultivos*: ídem dataset 11.

17) *dsVarPolTEleVolTodosConDatosCultivos*: ídem dataset 12.

18) *dsVarPolTEleVolTodosHDRSinDatosCultivos:* ídem dataset 13.

Como las herramientas seleccionadas para modelar tienen como restricción el uso de casos sin valores perdidos, a los dataset 7 a 18 se le sacaron los casos donde alguna variable presentaba valores perdidos. Las tablas 13 y 14 y las figuras 21 y 22 se ve como queda configurada la muestra de casos con respecto del target en todos los data set.

Tabla 13. Características de los datos contenidos en los dataset para modelar el target Daño (tres clases).

				Target: Daño		
				Mas50	**Menos50**	**Sin Daño**
12 Elevac.	Sin Cultivo	Volúmenes Completos	H_{DR}	21 (1,43%)	245 (16,73%)	1.198 (81.8%)
12 Elevac.	Sin Cultivo	Volúmenes Completos + Incompletos	H_{DR}	97 (4.6%)	512 (24.27%)	1.501 (71.1%)

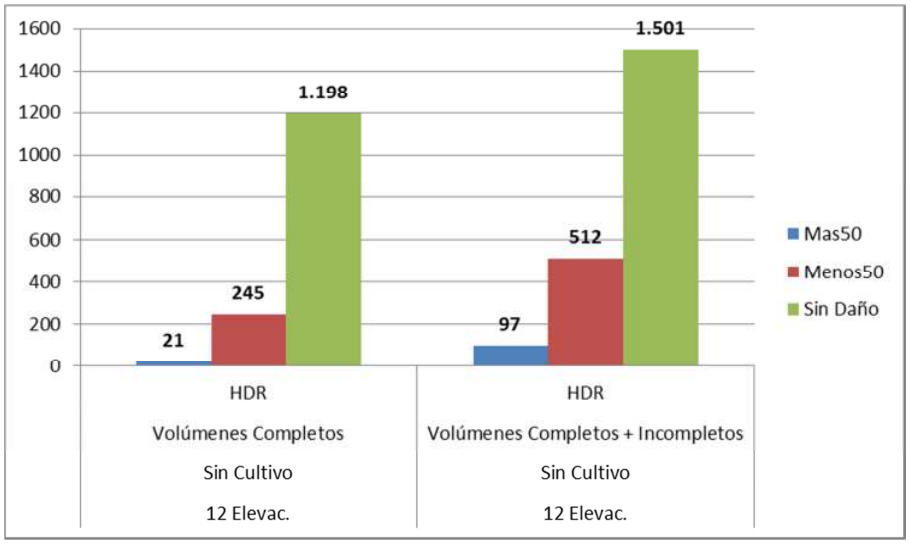


Figura 21. Distribución de los casos en los dos dataset para modelar el target Daño de tres clases.

Tabla 14. Características de los datos contenidos en los dataset para modelar el target Daño (cinco clases).

				Target: Daño				
				Leve	**Mode rado**	**Severo**	**Grave**	**Sin Daño**
1 Elevac.	Con Cultivo	Volúmenes Completos		209 (74.1%)	25 (8.9%)	13 (4.6%)	8 (2.8%)	27 (9.6%)
1 Elevac.	Sin Cultivo	Volúmenes Completos		214 (15.9%)	31 (2.3%)	13 (1%)	8 (0.6%)	1.077 (80.2%)
1 Elevac.	Sin Cultivo	Volúmenes Completos	H_{DR}	214 (15.9%)	31 (2.3%)	13 (1%)	8 (0.6%)	1.077 (80.2%)
12 Elevac.	Con Cultivo	Volúmenes Completos		209 (74.1%)	25 (8.9%)	13 (4.6%)	8 (2.8%)	27 (9.6%)
12 Elevac.	Sin Cultivo	Volúmenes Completos		214 (14.6%)	31 (2.1%)	13 (0.9%)	8 (0.5%)	1.198 (81.8%)
12 Elevac.	Sin Cultivo	Volúmenes Completos	H_{DR}	214 (14.6%)	31 (2.1%)	13 (0.9%)	8 (0.5%)	1.198 (81.8%)
1 Elevac.	Con Cultivo	Volúmenes Completos + Incompletos		425 (63.2%)	60 (8.9%)	45 (6.7%)	49 (7.3%)	94 (14.0%)
1 Elevac.	Sin Cultivo	Volúmenes Completos + Incompletos		431 (21.9%)	67 (3.4%)	45 (2.3%)	51 (2.6%)	1.377 (69.9%)
12 Elevac.	Con Cultivo	Volúmenes Completos + Incompletos		436 (63.1%)	63 (9.1%)	46 (6.7%)	49 (7.1%)	97 (14.0%)
12 Elevac.	Sin Cultivo	Volúmenes Completos + Incompletos		442 (20.9%)	70 (3.3%)	46 (2.2%)	51 (2.4%)	1.502 (71.2%)

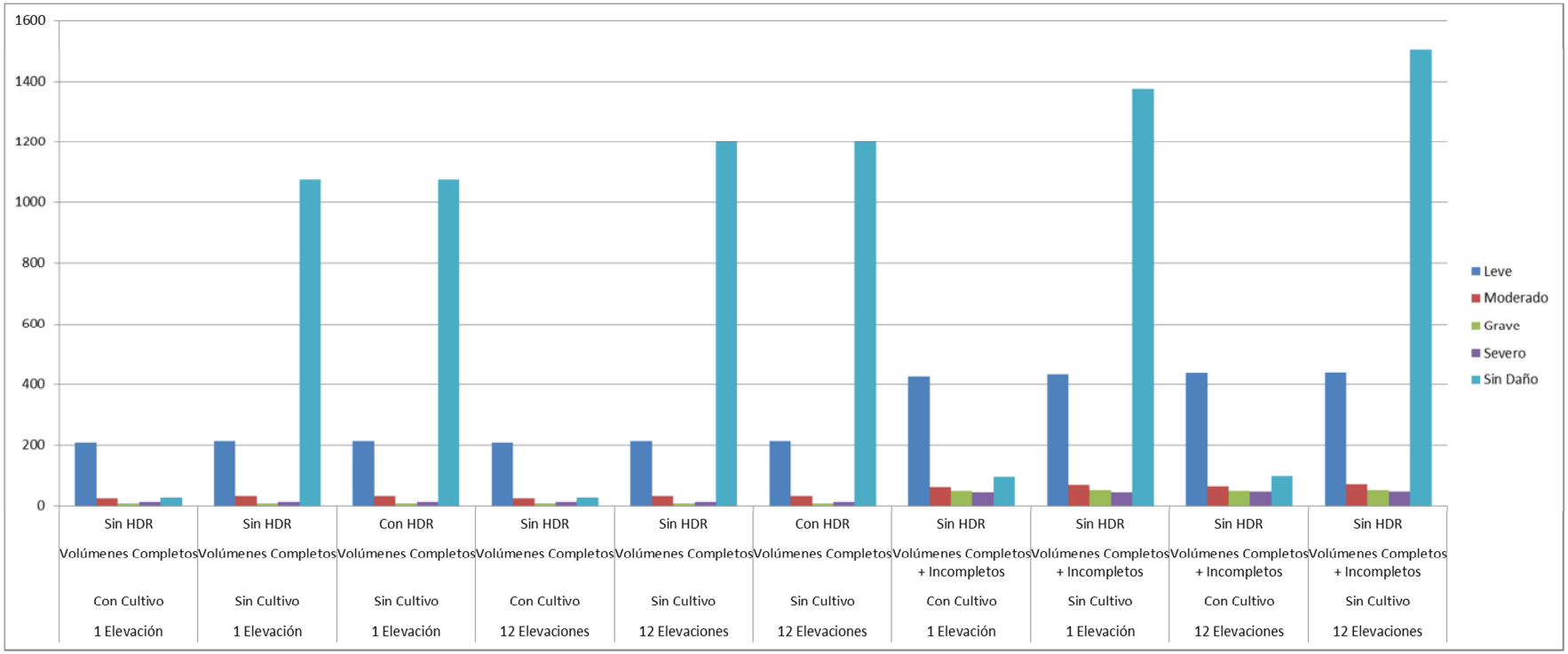


Figura 22. Distribución de los casos en los diez dataset para modelar el target Daño de cinco clases.

2.3. Modelar

2.3.1. Herramientas de Data Mining

Para realizar la tarea de modelar ambos target se trataron de forma binaria. Se modelaron utilizando Gene Expression Programming, seleccionando los métodos de regresión logística y clasificación con las funciones de rendimiento *Máxima Verosimilitud* y *ROC Measure* respectivamente. A continuación se introduce brevemente los conceptos del método GEP y de Regresión Logística.

2.3.1.1. Gene Expression Programming

Es un algoritmo evolutivo que genera automáticamente programas de computadora para modelar las relaciones entre las variables analizadas [125] [126] [127]. Se puede utilizar para construir diferentes tipos de modelos (Ej.: regresiones logísticas, arboles de decisión y redes neuronales) [125] [128]. Como todo algoritmo evolutivo se rige por el principio de la selección natural: los individuos (en este caso modelos o soluciones) mejor adaptados, tienen mayores posibilidades de sobrevivir y reproducirse en las próximas generaciones.

Tiene dos componentes principales: los cromosomas (genotipo) que son cadenas de caracteres de longitud fija para representar a los programas y los árboles de expresión (fenotipo) de diferentes formas y tamaños, por medio de los cuales se expresan los cromosomas (figura 23). Los cromosomas son los que están sujetos a variaciones genéticas utilizando uno o más operadores genéticos, permitiendo la evolución [125] [126] [127].

El algoritmo GEP inicia con una población aleatoria de cromosomas que representan posibles soluciones al problema en estudio. Estos cromosomas se asignan a los árboles de expresión, la aptitud de cada individuo se evalúa en base a una función predefinida y los mejores individuos se seleccionan para su reproducción y modificación genética a través de operadores genéticos de recombinación, mutación y reproducción. Esta nueva generación se somete al mismo proceso hasta que se encuentra una solución o se llega a un número definido de generaciones. El individuo más apto, de acuerdo a una o varias medidas de rendimiento, es la solución final [125] [126]. Para más detalles sobre este algoritmo ver [129].

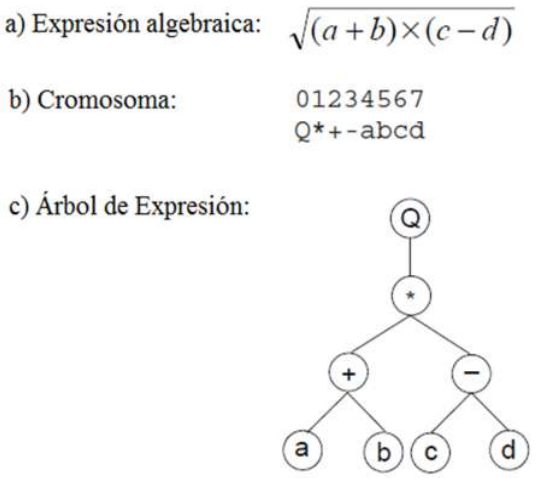


Figura 23. Ejemplo de cromosoma de una función algebraica y su representación por medio de un árbol de expresión utilizado en GEP (Adaptado de: [125])

Este método creado por Ferreyra, C., cuenta con una implementación en la herramienta GeneXProTools[30], utilizada en este trabajo. Este software tiene plataformas para trabajar con clasificación y regresión logística que cuentan con plantillas, con una configuración previa para el modelado, de acuerdo a la cantidad de variables independientes presentes en los datos [130] [131]; en ambos casos los modelos se generan de forma evolutiva [130] [131], armando automáticamente, nuevas variables a partir de las originales y relaciones entre las mismas, por medio de constantes, cálculos y funciones. Los algoritmos de reproducción, cruza y selección están debidamente explicados en [129] y este trabajo se concentró en la aplicación práctica del mismo y los resultados obtenidos, asumiendo como correcta su implementación en el software utilizado.

2.3.1.2. Regresión logística

Se utiliza para modelar problemas de clasificación cuya variable objetivo toma la forma $y \in \{0,1\}$, donde el cero corresponde a la clase negativa y el uno a la clase positiva. En el modelo de regresión logística se trata de calcular la probabilidad en la que una de las opciones de la variable y ocurra a partir de los valores que tomen una serie de variables independientes x. La hipótesis se representa por (7) [132] [133].

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n \quad (7)$$

[30] http://www.gepsoft.com. Empresa también creada por Ferreyra C.

Para que la función cumpla con la restricción de $0 \leq h_\theta(x) \leq 1$ se utiliza la función sigmoidal, por lo que la función de hipótesis queda como (8) [132] [133].

$$\frac{1}{1 + e^{-\theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n}} \quad (8)$$

Para obtener los valores de los coeficientes (θ) de los predictores que mejor los relacionan a la variable objetivo, la regresión logística utiliza la estimación de la máxima verosimilitud. La función de costo a minimizar queda dada por (9) [132] [133].

$$J(\theta) = -\frac{1}{n}[\sum_{i=1}^{n}(y_i \log h_\theta(x_i) + (1 - y_i)\log(1 - h_\theta(x_i))] \quad (9)$$

Una vez obtenidos los valores de θ, el resultado que arroja la función $h_\theta(x)$ determina la probabilidad que tiene el caso estudiado de pertenecer a la clase positiva, en general se predice y=1 si $h_\theta(x) \geq 0.5$ e y=0 si $h_\theta(x) < 0.5$.

2.3.1.4. Medidas de rendimiento de los modelos

Para presentar el rendimiento de los modelos obtenidos y compararlos se utilizan una serie de indicadores presentes en la literatura, basados en la matriz de confusión (tabla 15) de los datos de entrenamiento y testing utilizados en diferentes técnicas de DM [18] [19] [21] [22] [23] [57] [90] [92] [96] [95] y en el ámbito de la predicción de fenómenos climáticos [17] [21] [59] [66] [69] [75] [94] [97] [98] [134]. Algunos de los indicadores se comparten entre ambas temáticas; los más utilizados en clima, se concentran en el comportamiento del modelo con respecto de los verdaderos positivos, mientras que en DM también importan los verdaderos negativos.

Tabla 15. Matriz de confusión y medidas de rendimiento disponibles (Adaptado de: [132], [134], [135]).

Matriz de Confusión (Contingencia)			
		Valores Observados	
		Si	**No**
Valores Predichos	**Si**	**VP (Verdaderos Positivos) Identificados Correctamente**	**FP (Falso Positivo) Identificados incorrectamente**
	No	**FN (Falso Negativo) Incorrectamente rechazados**	**VN (Verdadero Negativo) Correctamente rechazados**

Medidas de Rendimiento			
Nombre	**Cálculo**		**Descripción**
FAR (False Alarm Ratio)	**FP/(VP+FP)**	(10)	Número de falsas alarmas positivas dividido por el número total de positivos predichos. Para un modelo perfecto FAR=0.
CSI (Critical Success Index), TS (Threat Score)	**VP/(VP+FP+FN)**	(11)	Total de verdaderos positivos dividido por el número total de las valores predichos SIN los VN por lo que no se ve afectado por el número de rechazos correctos.
GS (Gilbert Skill Score)	**(VP-CH)/(VP+FP+FN-CH) donde CH= (VP+FP)(VP+FN)/n**	(12)	Se comporta similar al CSI, pero tiene en cuenta los verdaderos positivos obtenidos debido a la casualidad (CH).
HSS (Heidke Skill Score)	**(VP+VN-E)/(VP+FP+FN+VN-E) donde E= [(VP+FP)(VP+FN)+ (FP+VN)(FN+VN)]/n**	(13)	Rango (-1,1). Pronóstico al azar, HSS=0; pronóstico perfecto, HSS=1; pronóstico imperfecto HSS = -1.Valores negativos del HSS indican que los desaciertos dominan el análisis. Para un modelo aceptable debe ser positivo.
POD (Probability Of Detection), Sensibilidad, Recall, Exhaustividad	**VP/(VP+FN)**	(14)	Capacidad para identificar resultados positivos correctamente. Modelo perfecto POD=1.
PC (Percent Correct), Accuracy	**(VP+VN)/(VP+VN+FP+FN)**	(15)	Proporción de los resultados reales (VP y VN) en la población.
Especificidad	**VN/(VN+FP)**	(16)	Capacidad para identificar los resultados negativos.
Error rate	**(FP+FN)/(VP+FP+FN+VN)**	(17)	Proporción de casos identificados incorrectamente contra todos los casos.
Precision, PPV(Positive Predictive Value), SR (Success Ratio)	**VP/(VP+FP) o 1-FAR**	(18)	Proporción de VP contra todos los resultados positivos (VP Y FP). Lo contrario de FAR.
NPV (Negative Predictive Value)	**VN/(VN+FN)**	(19)	Proporción de VN contra todos los resultados negativos (VN Y FN).
AUC (Area Under the Curve)	**Sensibilidad vs 1-especificidad**	(20)	Representación gráfica de la sensibilidad frente a (1 – especificidad) para un clasificador binario según se varía el punto de corte.
Medida F1	**2 . [(precision . recall)/precisión+recall]**	(21)	Promedio ponderado de precision y recall. El mejor valor de F1 es 1 y el peor es 0 .

2.3.2. Análisis Estadísticos

Se realizaron análisis estadísticos comparando las fechas que presentaban volúmenes completos, con las que no, con respecto a las variables polarimétricas. Este análisis permite visualizar si existen diferencias entre los datos de las tormentas con todos los volúmenes y los datos de las tormentas que le faltan volúmenes. Además se realizaron análisis de estadísticas univariadas y gráficos de caja de cada dataset. En el caso de los dataset 5 y 6, los valores se ven distorsionados por la presencia del "-

99.00" como valor perdido, pero estos análisis permiten visualizar si existe una relación entre las variables y el target buscado. También se realizaron análisis ANOVA de un factor para determinar la importancia de cada variable en la separación de las clases del target Daño y determinar si las diferencias de las variables del radar entre las fechas completas de las incompletas es significativa. Para ejecutar estas tareas se usó el software R, su entorno RStudio y el paquete RCommander.

2.3.3. Ejecución de las herramientas de DataMining

La figura 24 presenta el esquema utilizado para generar los modelos para el target granizo. Los modelos se corrieron con los valores por defecto de la plantilla para regresión logística de GEP[31], cambiando solamente la función fitness a "Máxima Verosimilitud". En el paso "Corrida Modelo" de la figura 24 se resumen las características de esta plantilla. La condición de parada del algoritmo es: llegar al valor 1000 en el "Max Fitness" o llegar a la generación 30.000, lo que ocurra primero.

En el caso del target Daño con cinco clases, se generó un modelo por cada clase de forma binaria (Severo/No Severo, Grave/No Grave, Moderado/No Moderado, Leve/No Leve y Sin Daño/Con Daño). La figura 25 muestra el esquema de ejecución. Los modelos se corrieron con los valores por defecto de la plantilla para clasificación de GEP[32], para este caso como función de fitness se utilizó "ROC Measure" y "Máxima Verosimilitud". En el paso "Corrida Modelo" de la figura 25 se resumen las características principales de esta plantilla.

El target con tres clases (Sin Daño, <50% y >50%), también se trató de forma binaria y con la plantilla correspondiente. La función de fitness utilizada fue "Máxima Verosimilitud". La figura 26 detalla el esquema de ejecución.

Para ambas opciones de clases del target Daño, se utilizó la misma condición de parada que para el target Granizo.

[31] Para los detalles completos de la plantilla ver Capítulo "13.1.14. Run Templates" en [136]
[32] Para los detalles completos de la plantilla ver Capítulo 13.1.14. Run Templates en [136]

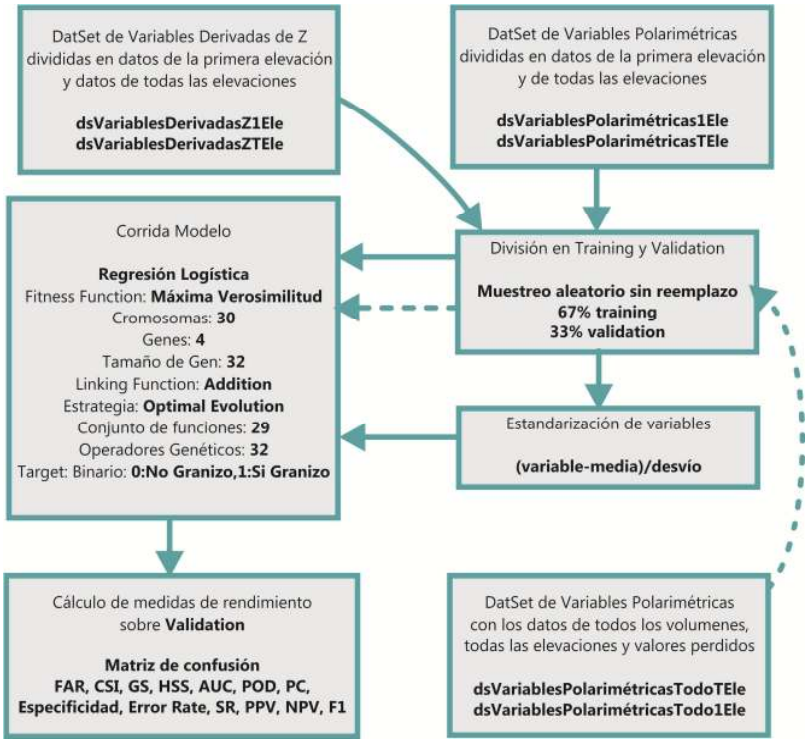


Figura 24. Esquema de tareas realizadas en la etapa de modelado del target Granizo

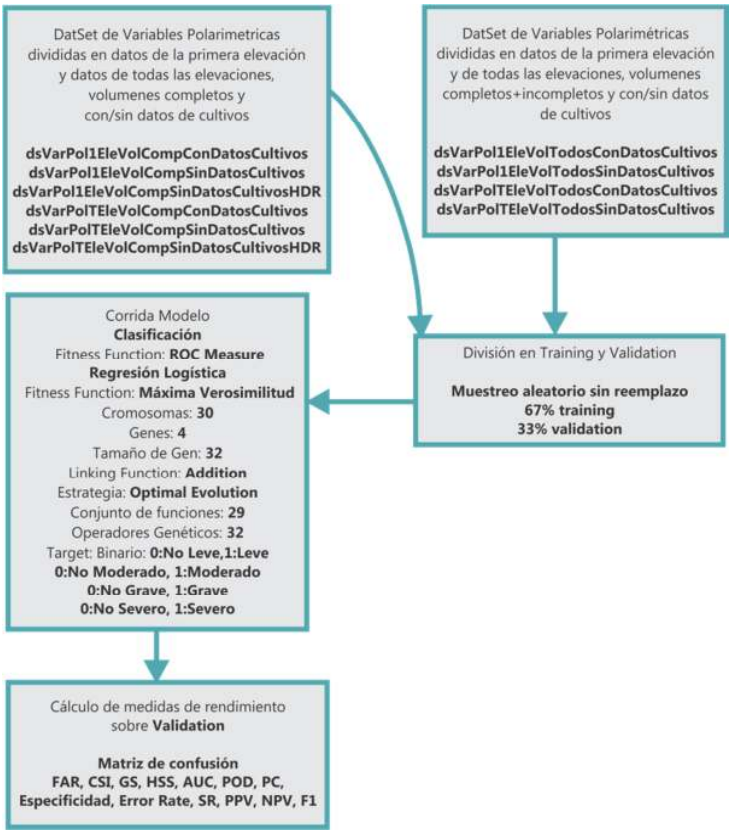


Figura 25. Esquema de tareas realizadas en la etapa de modelado del target Daño.

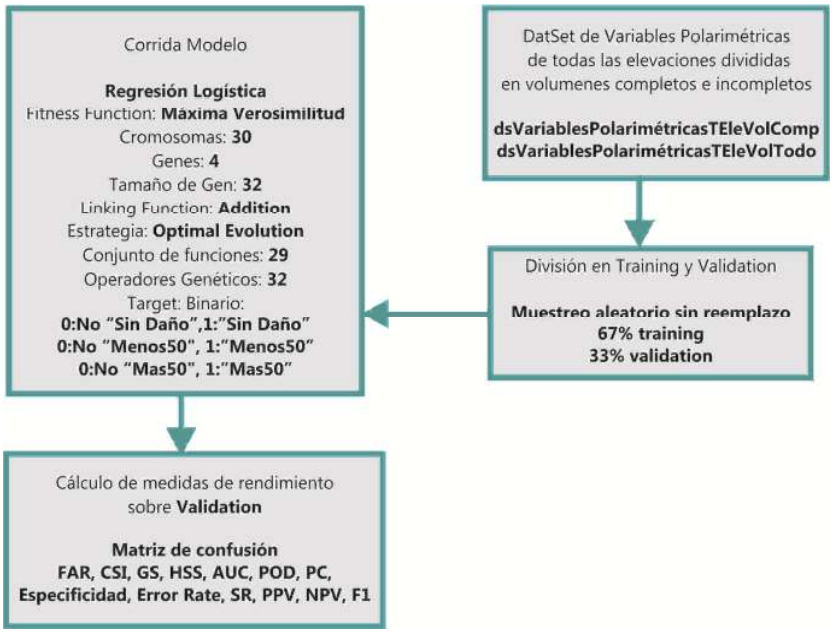


Figura 26. Esquema de tareas realizadas en la etapa de modelado del target Daño con tres clases.

2.4. Implementación

El modelo se va a usar para clasificar una serie de localizaciones nuevas; que puede ir de un solo caso a 245.935 casos, que corresponden a todos los pixeles que están dentro del área de cobertura del radar. La implementación implica realizar las siguientes tareas independiente del target y de la cantidad de localizaciones a clasificar: a) descargar los volúmenes del radar, b) convertir los volúmenes a archivos ASCII, c) convertir de los archivos ASCII a GeoTIFF y d) realizar los cálculos de las variables seleccionadas en los modelos para cada localización a clasificar. Estos cálculos implican: d.1) generar resúmenes de Z, Rho_{HV} y Z_{DR} con diferentes frecuencias temporales (cada 10 minutos y diarios) y con diferentes cálculos (máximos, mínimos, promedios y totales) y d.2) calcular variables derivadas de Z y de Z_{DR}. A partir de este paso se pueden tomar diversos caminos definidos por: a) la gran cantidad de datos involucrados, b) los tiempos de procesamiento, c) los recursos de hardware y software necesarios y d) nivel de automatización de la implementación. Para este trabajo se analizaron dos enfoques, pensando en una implementación que permita clasificar **toda el área del radar** para el target **Granizo**:

- *Opción 1:* recorrer todas las imágenes que genera el radar en un día, obtener los datos correspondientes a cada pixel, guardarlos en una base de datos relacional y realizar los cálculos de resúmenes de los modelos con sentencias T-SQL (este método es muy similar al utilizado en el armado de los dataset).
- *Opción 2:* generar imágenes compuestas, por medio de algebra lineal, que contengan los resúmenes de las variables seleccionadas para los modelos y posteriormente procesar estos resúmenes de acuerdo a las necesidades del modelo.

2.4.1. Opción 1: Base de datos relacional

Se programó el script de Python *datos10minutosImgCompleta.py* que toma como parámetros, la fecha a procesar, el directorio donde se encuentran los archivos GeoTIFF de esa fecha y si la información se almacena en una base de datos o en un archivo de texto. El algoritmo recorre todos los archivos correspondientes y extrae los valores de cada pixel, almacenándolos en el medio seleccionado. Utiliza el script *Identify.py*.

Se diseñó una base de datos relacional, con una tabla por cada variable de radar (Z, Z_{DR}, Rho_{HV}). Se programaron 6 vistas con T-SQL para calcular los resúmenes de cada variable (máximo, mínimo, promedio y total). La figura 28 muestra el esquema de funcionamiento de esta implementación.

2.4.2. Opción 2: Imágenes compuestas

Las imágenes compuestas se realizan utilizando todas las elevaciones (perfil vertical de la tormenta) o solo algunas. El script de Python *GIC.py* (Generar Imagen Compuesta) calcula estas imágenes a partir de los argumentos indicados por el usuario (tabla 16 y figura 27), por medio de operaciones entre matrices. El resumen diario de una elevación, toma las 144 imágenes correspondientes a un día y a esa elevación y las transforma en matrices de datos. Con esas matrices realiza una comparación (mayor y menor) o un cálculo (promedio y suma), elemento con elemento (pixel con pixel) de cada matriz y almacena los resultados en una matriz de resumen que convierte en una imagen (en la figura 27 se detalla como: "cálculo horizontal"). En el caso de los resúmenes temporales, se toman las 12 imágenes de cada escaneo y se realiza la misma operación que para el resumen diario (en la figura

27 se detalla como: “cálculo vertical”), si se hace para todas las tomas de 10 minutos, se procesan 1.728 matrices. Finalmente, para un resumen diario y de todas las elevaciones, se realiza un cálculo horizontal sobre las matrices de resúmenes obtenidas en el cálculo vertical (en la figura 27 aparece como resumen diario). Para estas operaciones se utilizan las funciones: fmin o min, fmax o max, mean o average y sum que provee la librería numpy [137]. Estas operaciones evitan tener que utilizar dos bucles para recorrer cada matriz, logrando que disminuya la complejidad computacional de la implementación.

Tabla 16. Argumentos script GIC.py.

Argumentos posicionales:	
Path_img	Ubicación de los archivos raster a procesar para generar la imagen compuesta.
Fecha	Fecha a procesar, formato: aaaammdd
Extensión	Extensión de los archivos a procesar, el valor por defecto es “tif”
Variable a procesar	Posibles valores: dBZ, ZDR, RhoHV, KDP, PhiDP, E, EW, EWs. Valor por defecto: dBZ
Argumentos opcionales:	
-h, --help	Muestra la ayuda
-mto	Calcula imágenes compuestas por cada paso de toma de datos (default: 10 minutos). Para todas las elevaciones.
-ele	Número de elevación a procesar. Posibles valores: 1 a 12. Si no se indica se procesan todas.
-d	Indica que hace falta generar las imágenes horarias para la imagen compuesta de 24 horas y todas las elevaciones.
-maxi	Genera la imagen compuesta con el valor máximo.
-mini	Genera la imagen compuesta con el valor mínimo
-prom	Genera la imagen compuesta con el valor promedio.
-tot	Genera la imagen compuesta con el valor total (suma).

Para este trabajo se generaron imágenes de resumen diario (24 horas) para cada una de las fechas a procesar (tabla 3), con los valores máximos, mínimos, promedios y totales de las variables Z, Z_{DR}, Rho_{HV}, E y H_{DR}. Como E y H_{DR} no son tomadas por el radar, previo a realizar los resúmenes, se calcularon usando los script de Python *EK.py* y *HDR.py*. Se generaron 52.269 archivos GeoTIFF de resúmenes que ocupan 101 Gb.

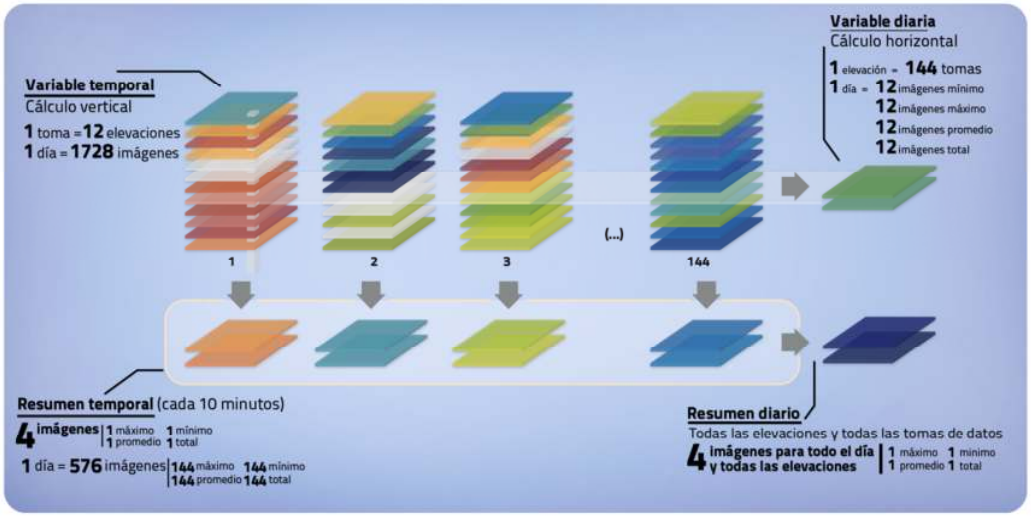


Figura 27. Esquema del cálculo de las imágenes compuestas temporales y diarias.

Para obtener los valores de las imágenes compuestas se programó el script de Python *datos10minutosResumenes.py* que obtiene los valores del pixel correspondiente a cada par de coordenadas geográficas pasadas como parámetros desde las imágenes resumen. Utiliza el script *Identify.py*.

Los valores obtenidos en las imágenes de resumen se validaron utilizando el script *PuntosfromRaster10minutos.py*, comparando los datos de ambos cálculos. La figura 28 muestra el esquema de funcionamiento de esta implementación.

Finalmente se programó el script de Python *AplicaModelo.py* que recibe como parámetros: la fecha a procesar y utiliza las funciones generadas automáticamente por la herramienta GeneXProTools que contiene el código fuente desarrollado por la metodología GEP del modelo seleccionado para la implementación. Este script se utiliza en ambas posibilidades de implementación (figura 28).

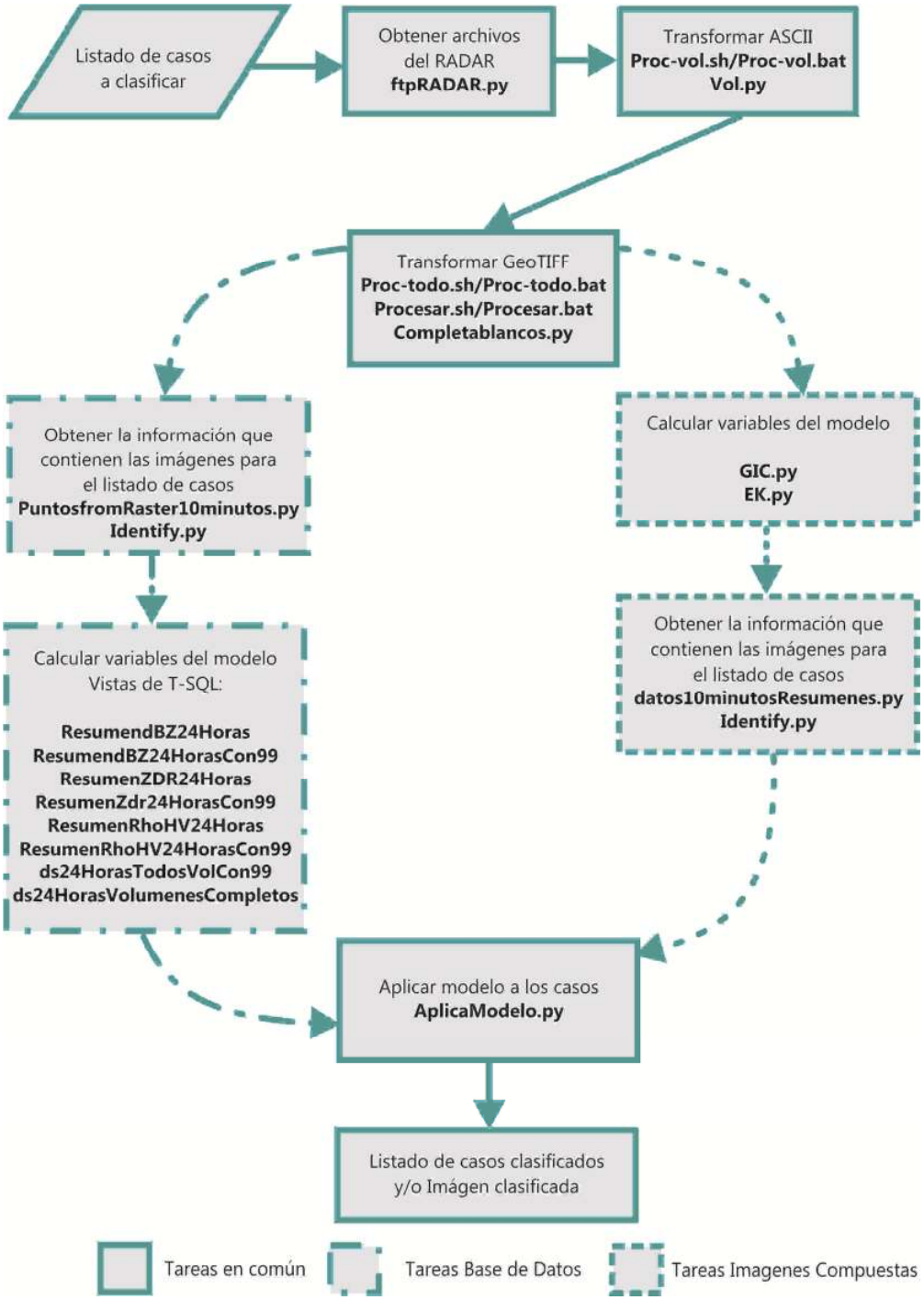


Figura 28. Esquema de tareas en dos alternativas de implementación.

Capítulo 3. Resultados y Discusión

3.1. Resultados Target Granizo

Este capítulo finaliza de explicar las tareas de "Modelar" y describe las actividades de las etapas "Evaluar" e "Implementar" del proceso CRISP-DM.

3.1.1. Características de los DataSet

La figura 29 muestra los diagramas de caja donde se aprecia la diferencia en la distribución de las variables Z (dBZ), Rho_{HV} y Z_{DR}, entre los casos de las fechas que tienen los volúmenes completos con respecto a las que no. Además el ANOVA de un factor mostró diferencias significativas en los valores de dBZ (Pr(>F) = <2e-16), Z_{DR} (Pr(>F) = <2e-16) y Rho_{HV} (Pr(>F) = <2e-16) entre los volúmenes completos y los incompletos. Esta evidencia apoya la decisión de construir una serie de datasets solo con los casos cuyas fechas tenga todos los volúmenes presenten y otros con todos los casos (tengan o no todos los volúmenes presentes).

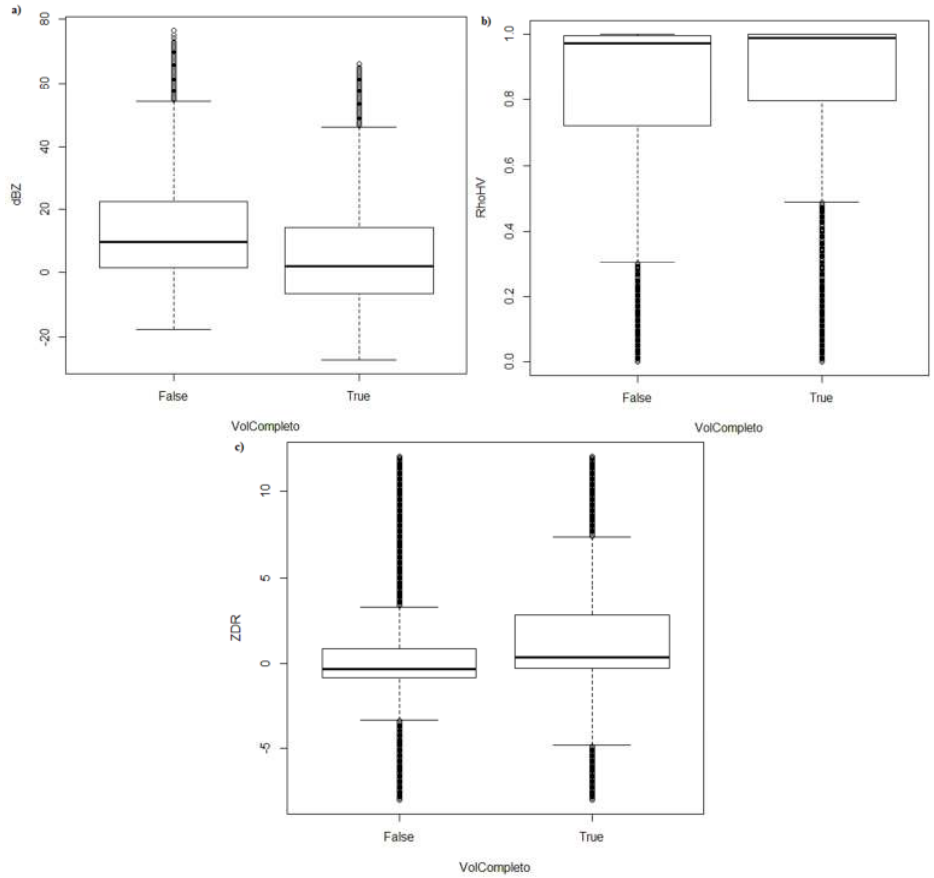


Figura 29. Diagramas de Caja de las variables: a) Z, b) Rho_{HV} y c) Z_{DR}.

3.1.1.1. Reflectividad

La figura 30 presenta los diagramas de caja de nueve variables generadas a partir de Z (para todas las elevaciones) y su distribución con respecto del target granizo. En todas las variables, los casos positivos presentan mayores valores de Z y mayor cantidad de presencia de ecos de reflectividad (>0 dBZ) y ecos fuertes (45, 50, 55 y 60 dBZ), lo cual coincide con los valores propuestos por [13] [63] [77] y [81]. El promedio de los valores máximos de Z para los casos positivos es de 47.5 dBZ (figura 30.a), lo que sugiere que se debería aumentar el umbral de 45 dBZ propuesto en la configuración del radar de INTA Anguil [34]; se propone el uso de 50 dBZ para la zona bajo estudio, utilizando el mismo valor que [32] y [33] usan con el radar de INTA Pergamino. Este valor es menor al sugerido por los mismos autores para el radar de INTA Paraná (60-65 dBZ).

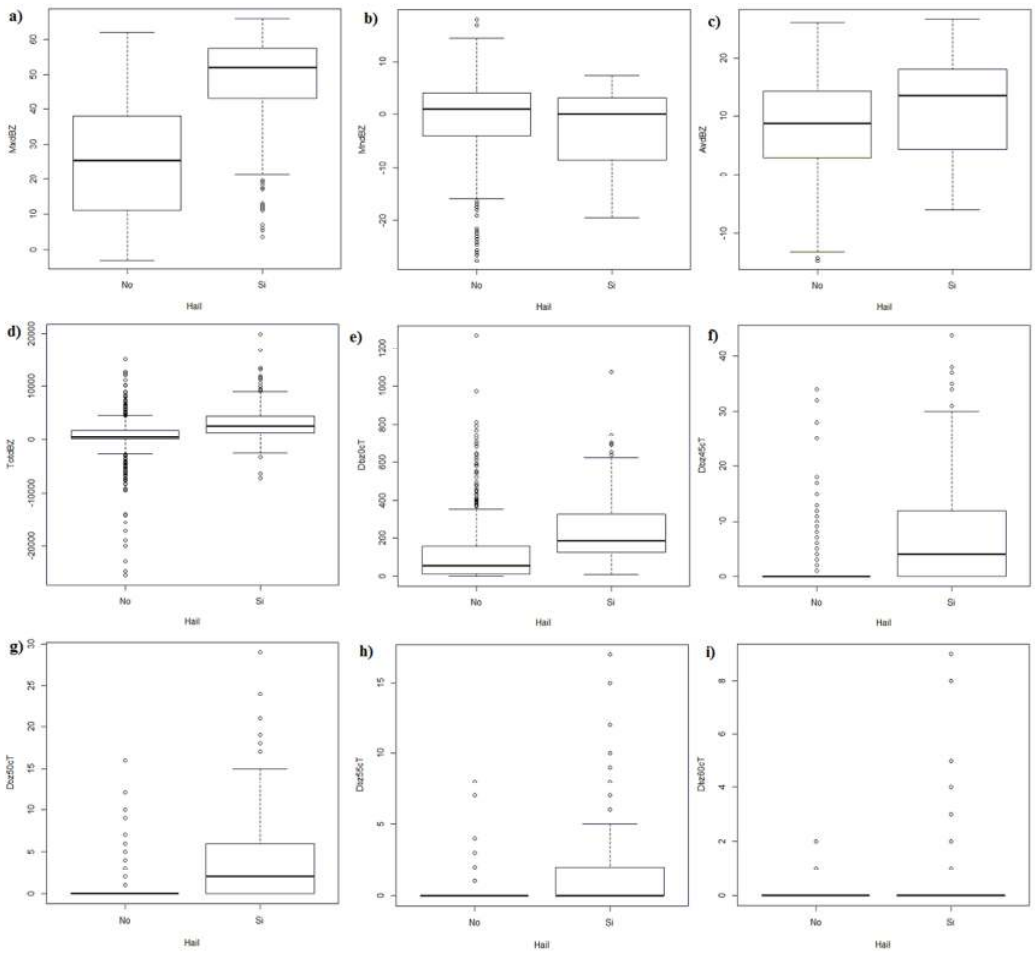


Figura 30. Diagramas de Caja de las variables: a) MxdBZ, b) MndBZ y c) AvdBZ, d) TotdBZ, e) Dbz0cT, f) Dbz45cT, g) Dbz50cT, h) Dbz55cT, i) Dbz60cT.

3.1.1.2. Rho_{HV}

La figura 31 muestra los diagramas de caja de las variables de resúmenes de Rho_{HV} para todas las elevaciones. Con excepción del máximo (MxRho) el resto de las variables muestran diferencia entre las clases del target; para la variable de mínimos (MnRho) los valores se alejan marcadamente de 1 en ambas clases, siendo los casos positivos los que menor valor presentan. La variable promedio (AvRho) muestra una mayor dispersión en los casos negativos y también una caída en la media de AvRho para los casos positivos. Estos comportamientos coinciden con lo mencionado en la bibliografía para la banda C ([13] [63] [77] [81]) y para Argentina ([31]). Los valores menores a 0,95 presentes en este trabajo se pueden deber a áreas de granizo mezclado con grandes gotas de lluvia, que generan una gran variedad de formas y maneras de caer de los hidrometeoros de acuerdo a [63].

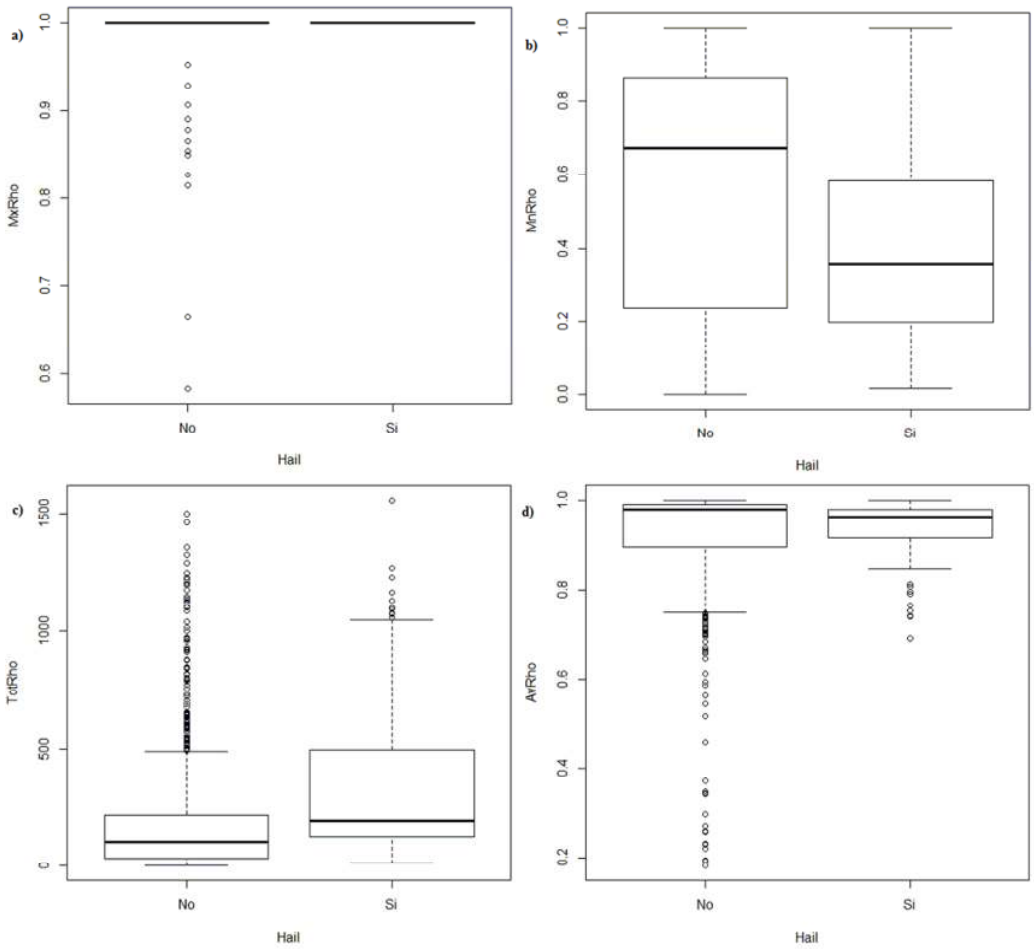


Figura. 31. Diagramas de Caja de las variables: a) MxRho, b) MnRho y c) TotRho, d) AvRho

3.1.1.3. Z_{DR}

Los valores máximos y mínimos de Z_{DR}, presentan diferencias importantes entre las clases del target. El promedio de los valores mínimos es de -5.14 dB (figura 32.b), el promedio de los valores máximos es de 5.69 dB (figura 32.a) y el promedio de las medias es de 0.54 dB (figura 32.d). Estos promedios de los máximos y los mínimos coinciden con los valores observados por [63] [77] [81] y [82] para banda C, donde los valores son elevados (> 4 dB); mientras que el promedio de las medias coincide con el comportamiento observado para banda S, donde los valores son cercanos a cero ([56] [63] [64] [65] [80]). Ésta marcada variación en los valores de Z_{DR} fue observada por [13], que encontró valores de -5 dB hasta 5 dB pasando por los valores cercanos a 0, ante la presencia de diferentes formas de granizo (grande, derritiéndose, mezclado con lluvia, etc.). Este comportamiento también lo observaron [31] y [33] en el radar de INTA Paraná ante la presencia de granizo.

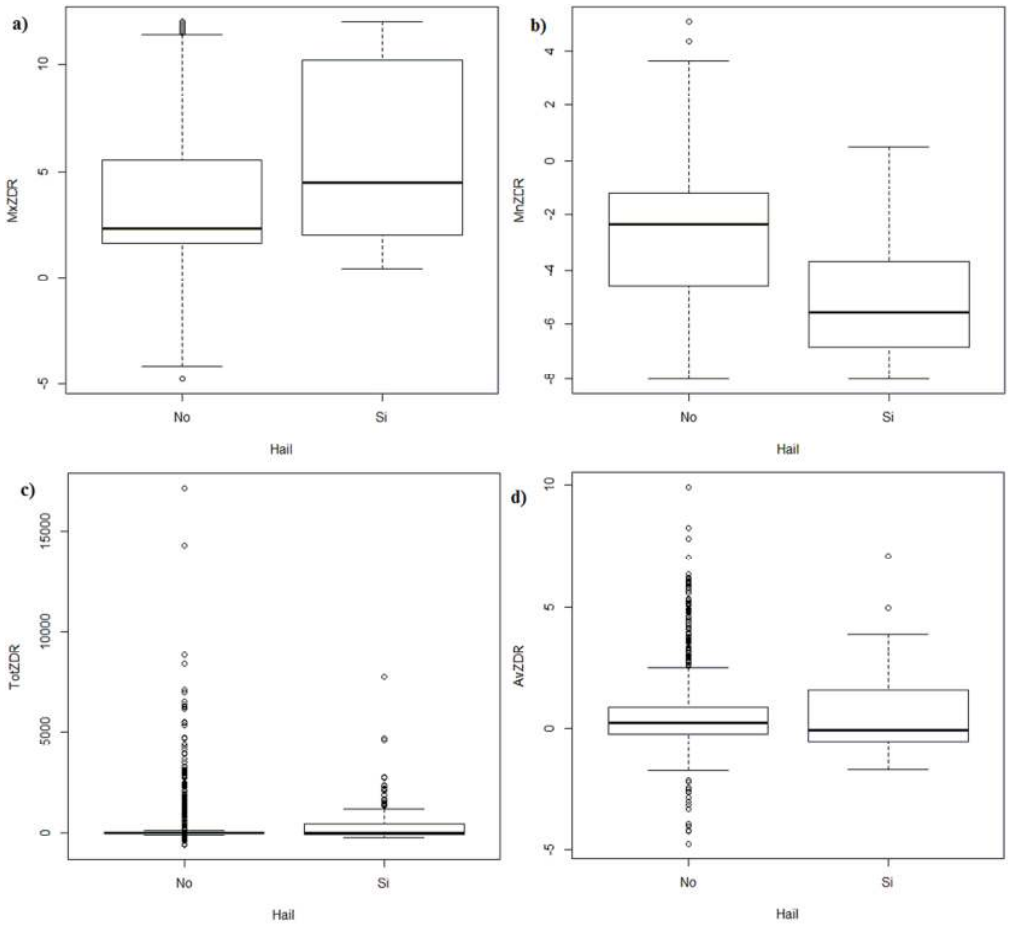


Figura. 32. Diagramas de Caja de las variables: a) MxZDR, b) MnZDR y c) TotZDR, d) AvZDR

3.1.2. Modelos Obtenidos

Todos los modelos presentan rendimientos muy buenos en la clasificación de los casos en el dataset de testing (tablas 17 y 18). Los modelos con variables polarimétricas son mejores que aquellos que usan las variables derivadas de Z, lo cual coincide con los antecedentes ([30] [56] [59] [62] [63] [64] [65] [66]). De los seis modelos construidos con variables polarimétricas, los modelos 4 y 6 "empatan" en su performance (cada uno tiene seis medidas de rendimiento con el mejor valor); ambos modelos corresponden a datos de todas las elevaciones y sin estandarizar. Para las variables derivadas de Z, el mejor modelo es el 10, que utiliza variables de todas las elevaciones, estandarizadas.

Tabla 17. Resultados modelos regresión logística de Variables Polarimétricas (dataset de testing)

	Primera Elevación			**Todas las Elevaciones**		
Ejecución	**Modelo 1 RL**	**Modelo 2 RL Std.**	**Modelo 3 RL Completo**	**Modelo 4 RL**	**Modelo 5 RL Std.**	**Modelo 6 RL Completo**
Variables	12	12	12	12	12	12
Training	946	946	1.032	1.029	1.029	1.032
Testing	473	473	1.983	514	514	1.983
Variables usadas						
	MxdBZ1 MndBZ1 TotdBZ1 AvdBZ1 MnRho1 MxZDR1 MnZDR1 TotZDR1 AvZDR1	MxdBZ1 TotdBZ1 MxRho1 MnRho1 AvRho1 MxZDR1 MnZDR1 TotZDR1 AvZDR1	MxdBZ TotdBZ AvdBZ MnRho TotRho AvRho MxZDR MnZDR AvZDR	MxdBZ MxRho MnRho TotRho MxZDR MnZDR AvZDR	MxdBZ MndBZ TotdBZ AvdBZ MnRho AvRho MxZDR MnZDR TotZDR	MxdBZ MndBZ TotdBZ AvdBZ MxRho TotRho AvRho MxZDR MnZDR TotZDR AvZDR
Medida	**Resultados dataset de Testing**					
FAR	0,2364	0,2258	0,2092	0,2353	0,3077	0,1441
CSI	0,5714	0,6443	0,6859	0,6691	0,6081	0,7388
GS	0,4700	0,5481	0,5100	0,5946	0,5194	0,5888
HSS	0,6394	0,7081	0,6755	0,7458	0,6837	0,7412
AUC	0,8779	0,8881	0,9065	0,9018	0,8905	0,9190
POD	0,6942	0,7934	0,8380	0,8426	0,8333	0,8437
PC	0,8668	0,8879	0,8411	0,9125	0,8872	0,8739
Especific	0,9261	0,9205	0,8434	0,9310	0,9015	0,8961
Error Rate	0,1332	0,1121	0,1589	0,0875	0,1128	0,1261
SR,PPV	0,7636	0,7742	0,7908	0,7647	0,6923	0,8559
NPV	0,8981	0,9284	0,8805	0,9570	0,9531	0,8868
F1	0,7272	0,7837	0,8137	0,8018	0,7563	0,8498
Referencias		Mejor valor				

Tabla 18. Resultados modelos regresión logística de Variables Derivadas de Z (dataset de Testing)

	Primera Elevación		**Todas las Elevaciones**	
Ejecución	**Modelo 7 RL**	**Modelo 8 RL Std.**	**Modelo 9 RL**	**Modelo 10 RL Std.**
Variables	13	13	19	19
Training	986	986	1.066	1.066
Testing	492	492	533	533
Variables usadas				
	MxdBZ1 Dbz451 Dbz551 TotEWt AvEWt	MxdBZ1 TotdBZ1 Dbz01 Dbz451 Dbz501 Dbz601 TotEWt AvEWt MxEWt MnEWt	Dbz0cT Dbz45cT Dbz50cT Dbz0T Dbz1T Dbz45T Dbz50T Dbz60T MndBZ TotEWt MxEWt	Dbz45cT Dbz55cT Dbz1T Dbz45T Dbz55T MxdBZ MndBZ TotEWt MnEWt
Medida	**Resultados dataset de Testing**			
FAR	0,3540	0,3333	0,2957	0,3413
CSI	0,4679	0,4837	0,4909	0,5253
GS	0,3584	0,3771	0,3857	0,4267
HSS	0,5276	0,5477	0,5567	0,5981
AUC	0,8288	0,8387	0,8132	0,8649
POD	0,6293	0,6379	0,6183	0,7217
PC	0,8313	0,8394	0,8424	0,8593
Especificidad	0,8936	0,9016	0,9154	0,8971
Error Rate	0,1687	0,1606	0,1576	0,1407
SR,PPV	0,6460	0,6667	0,7043	0,6587
NPV	0,8865	0,8898	0,8804	0,9214
F1	0,6375	0,6520	0,6585	0,6888
Referencias		Mejor valor		

Los dos modelos que tienen la mejor performance se seleccionaron y se nombraron con los vocablos mapuches Kurá (modelo 4) que significa Piedra y Pire (modelo 6) que significa Nieve-Granizo, con el objetivo de darles una identidad territorial asociada al área de estudio y al lugar donde se desarrollaron; permitiendo además que puedan ser identificados y recordados fácilmente por los usuarios. Las tabla 19 resume el rendimiento por medio de POD y FAR[33] de otros algoritmos de clasificación de granizo. De los diez algoritmos analizados, hay solo dos que tienen un rendimiento mejor que los logrados en este trabajo: algoritmo HCA y modelos discriminantes.

[33] Se presentan estas medidas porque son las disponibles en las publicaciones correspondientes.

Tabla 19. Rendimiento de POD y FAR de algoritmos de clasificación de granizo.

Método	**POD**	**FAR**	**Trabajo**
Z > 55 dBZ	66,2%	48,9%	[75]
Z_{CORR} > 55 dBZ	88.7%	67.8%	[75]
Dry + wet hail (fuzzy logic)	90.8%	67.6%	[75]
Dry hail (fuzzy logic)	71%	43.9%	[75]
HCA	100%	11%	[64] y [66]
HDA	88%	39%	[64] y [66]
Logistic regresión	67%	14%	[97]
Discriminant model	86%	12%	[14]
Logistic regresión	84%	15%	[14]
Logistic regresión	64%	11%	[17]
Kurá	***84%***	***23%***	
Pire	***84%***	***14%***	

La tabla 20 muestra en orden decreciente el valor de PC[34] de veinte algoritmos de clasificación de granizo, se observa que los modelos Kurá y Pire poseen los dos rendimientos más altos.

Tabla 20. Rendimiento de PC de algoritmos de clasificación de granizo.

Método	**PC**	**Trabajo**
Pire	0,86	
Kurá	0,84	
Rough Set Methods (10 clases)*	0,82	[92] y [57]
Support Vector Machine	0,81	[22]
Support Vector Machine (10 clases)*	0,80	[92] y [57]
Fuzzy c-means*	0,80	[90]
MPL (binary splits)	0,78	[90]
Neural Network (10 clases)*	0,78	[92] y [57]
Radial Basis Function (10 clases)*	0,77	[92] y [57]
Multilayer perceptron*	0,76	[90]
Genetic Algorithm	0,75	[19]
Rough Set Methods (4 clases)*	0,75	[92] y [57]
Classic Radial Basis Function	0,74	[22]
Hybrid Radial Basis Function	0,72	[22]
Rough Set Methods	0,72	[95]
Pseudo-inverse	0,70	[90]
Inductive decisión tree*	0,70	[90]
FMF	0,70	[138]
Support Vector Machine (4 clases)*	0,68	[92]
Neural Network (4 clases)*	0,63	[92] y [57]
Radial Basis Function (4 clases)*	0,54	[92] y [57]
K nearest neighbor	0,41	[90]
* Los trabajos mostraban resultados de diferentes configuraciones de estas metodologías, en esta tabla solo se muestra el modelo con el mejor rendimiento.		

[34] Se presenta esta medida de rendimiento porque es la disponible en las publicaciones correspondientes.

Las figuras 33 y 34 presentan los árboles de expresión de los modelos Kurá y Pire. Las tablas 21 y 22 muestran las funciones logísticas finales, el detalle de cómo se construye cada variable y los valores de las constantes en pseudo código. Del análisis de las variables incluidas en los modelos, se destaca para clasificar el target:

a) El valor máximo de Z (seleccionada en 9 de los 10 modelos),
b) La presencia de los ecos de 45 dBZ seleccionado en los cuatro modelos derivados de Z. Los otros ecos analizados (50, 55 y 60) fueron seleccionados en la mitad de los modelos,
c) El máximo y el mínimo de Z_{DR} se seleccionaron en los 6 modelos polarimétricos,
d) El mínimo de Rho_{HV} seleccionado en 5 de 6 modelos y
e) El total de E, elegida en todos los modelos de las variables derivadas de Z.

Esta selección de variables es coherente con los antecedentes: a mayor valor de Z (y por ende de E), menores valores de Rho_{HV} y valores extremos de Z_{DR}, mayor es la probabilidad de presencia de granizo ([13] [31] [33] [47] [63] [77] [81] [82]).

Tabla 21. Función logística para la probabilidad de granizo positivo del Modelo Kurá y detalles de su construcción.

<table>
<tr><th colspan="4">Función logística</th></tr>
<tr><td colspan="4">$P(Y=1) = \frac{1}{1 + \exp(0.112107575724697 * y + -6.45746730193226)}$</td></tr>
<tr><th colspan="4">Armado de y</th></tr>
<tr><td colspan="4">y = (gepGOE2A((AvZDR*gepGOE2C(-1.6825800347911, 8.18572069460128)), -7.17642750328074)+(3.55021820734275-gepLT2C((MxRho-MxdBZ), -1.6825800347911)))

y = y + (9.3725394451735+(((gepGOE2G(gepLT2A(MnZDR,TotRho),dMnRho)+(dMxZDR-dMnZDR))/2.0)-(gepAND1(-19.3568254646443,dMxZDR)-dAvZDR)))

y = y + ((((gepAND1(-9.27060762352367,MxRho)+((-6.30753501998962+ -7.53898739585559)/2.0))/2.0)-gepGOE2G(gepAND1(MxZDR,dMnZDR),TotRho))*gepAND2(((-5.29526657918027+MxdBZ)/2.0),AvZDR))

y = y + (MnRho-(gepGOE2E(8.75606555375835,exp(MnRho))*MnRho))</td></tr>
<tr><th colspan="4">Funciones lógicas para el armado de y</th></tr>
<tr><th>gepAND1</th><th>gepAND2</th><th>gepLT2A</th><th>gepLT2C</th></tr>
<tr><td>Recibe dos valores x e y:
Si (x < 0) y (y < 0) devuelve 1, si no devuelve 0</td><td>Recibe dos valores x e y:
Si (x >= 0) y (y >= 0) devuelve 1, si no devuelve 0.</td><td>Recibe dos valores x e y:
Si (x < y) devuelve x, si no devuelve y.</td><td>Recibe dos valores x e y:
Si (x < y) devuelve (x+y), si no devuelve (x-y)</td></tr>
</table>

gepGOE2A	gepGOE2C	gepGOE2E	gepGOE2G
Recibe dos valores x e y: Si (x >= y): devuelve x, si no devuelve y	Recibe dos valores x e y: Si (x >= y): devuelve (x+y), si no (x-y)	Recibe dos valores x e y: Si (x >= y): devuelve (x+y), si no devuelve (x*y)	Recibe dos valores x e y: Si (x >= y): devuelve (x+y), si no devuelve atan(x*y)

Tabla 22. Función logística para la probabilidad de granizo positivo del Modelo Pire y detalles de su construcción.

Función logística
$P(Y=1) = \frac{1}{1 + \exp(6.48084911389186E\text{-}05*y + -4.63926726110945)}$
Armado de *y*
y = gepLT2C((gepLT2G((9.64666513565478*(1.0-TotdBZ)),(1.0-gepLT2G(MxZDR,MnZDR)))*AvRho),AvdBZ) y = y + pow(exp(exp(gepGOE2G(gepLT2C(gepLT2B(TotdBZ,TotZDR),exp(-4.10822324594867)),gepAND2(gepLT2E(TotRho,TotZDR),pow(MxdBZ,3.0))))),4.0) y = y + gepGOE2C(pow((MnZDR/gepLT2G(TotRho,TotZDR)),2.0),gepLT2E(gepLT2E(gepLT2G(MxZDR,MxdBZ),(MxRho-MxZDR)),(1.0-MxdBZ))) y = y + gepGOE2G(((((gepGOE2G(MxZDR,AvRho)*MndBZ)+pow(MxdBZ,3.0))/2.0)+(gepLT2G(AvRho, 5.10116885891293)*gepLT2G(AvZDR,AvRho))),AvZDR)

Funciones lógicas para el armado de *y*			
gepAND2	**gepLT2B**	**gepLT2C**	**gepLT2E**
Recibe dos valores x e y: Si (x >= 0) y (y >= 0): devuelve 1 si no 0	Recibe dos valores x e y: Si (x < y): entonces devuelve 1, si no devuelve 0	Recibe dos valores x e y: Si (x < y) devuelve (x+y), si no (x-y)	Recibe dos valores x e y: Si (x < y) devuelve (x+y), si no devuelve (x*y)
gepLT2G	**gepGOE2C**	**gepGOE2G**	
Recibe dos valores x e y: Si (x < y): devuelve (x+y), si no devuelve atan(x*y)	Recibe dos valores x e y: Si (x >= y) devuelve (x+y), si no devuelve (x-y)	Recibe dos valores x e y: Si (x >= y): devuelve (x+y), si no devuelve atan(x*y)	

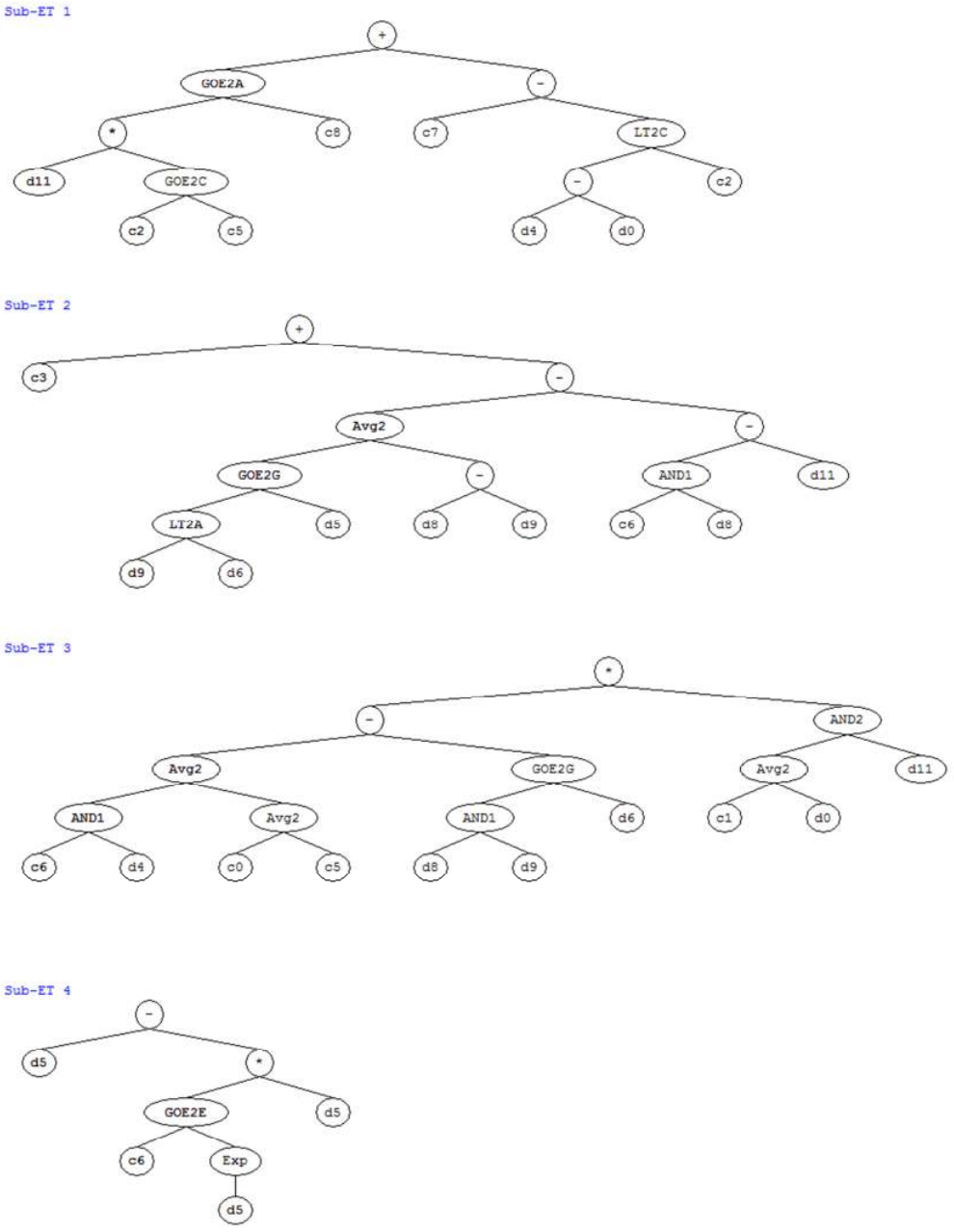


Figura 33. Árbol de Expresión del modelo Kurá (*Variables:* MxdBZ = d0, MxRho = d4, MnRho = d5, TotRho = d6, MxZDR = d8, MnZDR = d9, AvZDR = d11. *Constantes:* Sub-ET 1: C8 = -7.17642750328074, C7 = 3.55021820734275, C2 = -1.6825800347911, C5 = 8.18572069460128, Sub-ET2: C3 = 9.3725394451735, C6 = -19.3568254646443, Sub-ET 3: C1 = -5.29526657918027, C6 = -9.27060762352367, C0 = -6.30753501998962, C5 = -7.53898739585559, Sub-ET 4: G4C6 = 8.75606555375835)

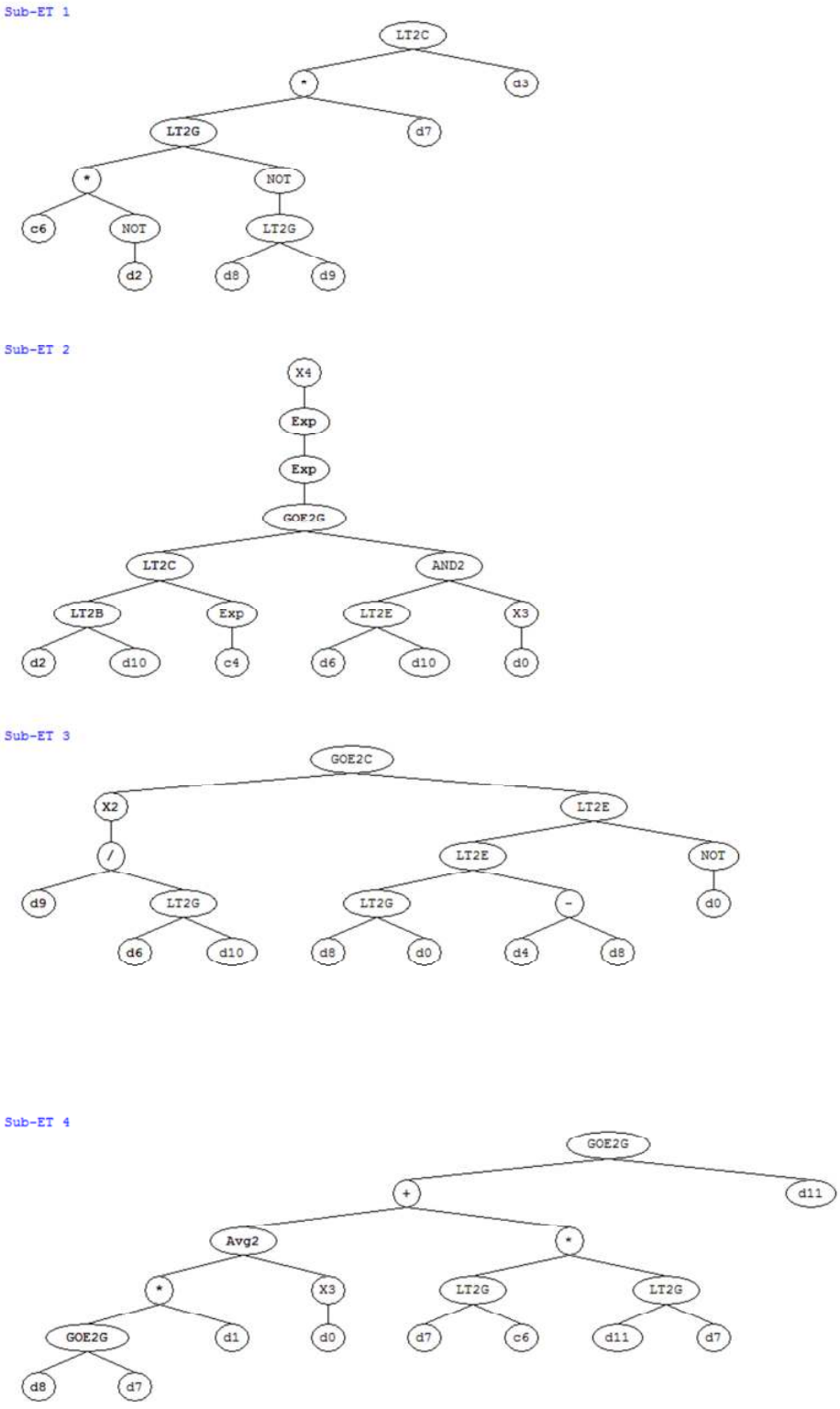


Figura 34. Árbol de Expresión del modelo Pire (*Variables:* MxdBZ = d0, MndBZ = d1, TotdBZ = d2, AvdBZ = d3, MxRho = d4, TotRho = d6, AvRho = d7, MxZDR = d8, MnZDR = d9, TotZDR = d10, AvZDR = d11. *Constantes:* Sub-ET 1: C6 = 9.64666513565478, Sub-ET 2: C4 = -4.10822324594867, Sub-ET 4: C6 = 5.10116885891293).

3.2. Resultados Target Daño

3.2.1. Características de los DataSet

3.2.1.1. Reflectividad

La tabla 23 presenta las estadísticas básicas por clase de daño de las variables de resumen de Z. Las cuatro variables presentan los menores valores registrados para la clase "Sin Daño", como los casos sin daño pertenecen en su gran mayoría a casos sin granizo o de granizo muy leve, es lógico que los registros de Z para esta clase son los de menor valor [55] [58] [67].

Tabla 23. Estadísticas básicas de las variables de resumen de Z en el dataset de todos los volúmenes, todas las elevaciones y sin datos de cultivos.

Variable	***MxdBZ***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	53,9	51,6	54,3	29,2	49,4
Máximo	66.0	72,0	72,0	70,0	69,0
Mínimo	30,0	7,0	17,0	-1,5	23,5
Variable	***MndBZ***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	2,9	0,1	0,4	-0,2	1,1
Máximo	12,5	14,0	12,5	17,0	13,0
Mínimo	-13,5	-19,5	-15,0	-27,5	-18,0
Variable	***AvDBZ***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	17,1	13,8	14,6	9,2	14,9
Máximo	24,1	32,1	25,9	30,3	24,8
Mínimo	4,1	-6,0	0,8	-14,7	-5,0
Variable	***TotDBZ***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	2.224,3	2.901,7	3.120,0	889,9	1.780,1
Máximo	8.085,5	13.443,5	13.052,5	15.053,5	5.705,5
Mínimo	572,5	-7.285,0	88,0	-25.415,0	-6.389,5

La clase "Grave" se diferencia de las demás porque sus valores mínimos son los más elevados, mientras que las clases "Leve", "Moderado" y "Severo" se comportan de manera muy similar en las variables MxdBZ, MndBZ y AvdBZ; el análisis de ANOVA indica que no hay diferencias significativas entre estas tres clases con ninguna de estas tres variables (MxdBZ: Pr(>F)= 0,121, MndBZ: Pr(>F)= 0,726 y AvdBZ: Pr(>F)= 0,57), mientras que TotdBZ presenta diferencias significativas (Pr(>F)= 0,00273) para la clase "Severo".

3.2.1.2. RhoHV

La variable MxRho no presenta diferencias entre las clases. MnRho presenta los mayores valores para la clase "Grave" (tabla 24). Un análisis ANOVA para cada variable entre las clases "Leve", "Moderado" y "Severo" confirma que no hay diferencias significativas (MxRho: Pr(>F)= 0,877, MnRho: Pr(>F)= 0,274, AvRho: Pr(>F)= 0,339 y TotRho: Pr(>F)= 0,941).

Tabla 24. Estadísticas básicas de las variables de resumen de RhoHV en el dataset de todos los volúmenes, todas las elevaciones y sin datos de cultivos.

Variable	***MxRho***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	1,00	1,00	1,00	0,99	1,00
Máximo	1,00	1,00	1,00	1,00	1,00
Mínimo	1,00	1,00	1,00	0,43	1,00
Variable	***MnRho***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	0,655	0,361	0,381	0,552	0,428
Máximo	0,925	0,957	0,976	1,000	0,886
Mínimo	0,024	0,008	0,020	0,000	0,024
Variable	***AvRho***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	0,9481	0,9099	0,9128	0,9214	0,8909
Máximo	0,9920	0,9973	0,9955	1,0000	0,9907
Mínimo	0,6608	0,6222	0,6109	0,1820	0,6584
Variable	***TotRho***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	174,17	298,60	290,56	185,57	286,38
Máximo	944,34	1.267,78	1.097.10	1.501.16	1.056.75
Mínimo	35,31	28,24	31,85	0,918	61,15

3.2.1.3. Z_{DR}

La clase "Grave" (mayor % de daño) presenta los valores de Z_{DR} más cercanos a cero, sumado a valores de Z > 50 dBZ (tabla 23) concuerda con [63] que sugiere la posibilidad de observar esta anti correlación de Z y Z_{DR} en banda C, ante la presencia de granizos grandes y en mucha cantidad, que dominan Z y aparecen como esferas perfectas. El resto de las clases presenta un promedio por encima de 4 dB, acorde al comportamiento de la banda C en presencia de granizo [13] [63] [77] [81] [82] (tabla 25). Estas tres variables tampoco presentan diferencias significativas entre las clases "Leve", "Moderado" y "Severo" (MxZDR: Pr(>F)=0,381, MnZDR: Pr(>F)= 0,118 y

AvZDR: Pr(>F)=0,689). La variable TotZDR presenta diferencia significativa (Pr(>F)= 0,00624) para la clase "Severo".

Tabla 25. Estadísticas básicas de las variables de resumen de Z_{DR} en el dataset de todos los volúmenes, todas las elevaciones y sin datos de cultivos.

Variable	***MxZDR***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	3,082	5,332	4,957	4,002	4,483
Máximo	12,000	12,000	12,000	12,000	11,921
Mínimo	0,819	-0,126	-0,126	-4,772	0,504
Variable	***MnZDR***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	-3,856	-5,069	-4,800	-3,118	-4,478
Máximo	-1,543	-0,835	-0,992	5,071	-1,150
Mínimo	-8,000	-8,000	-8,000	-8,000	-7,921
Variable	***AvZDR***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	-0,2023	0,2521	0,1677	0,4728	0,4153
Máximo	2,6227	4,9732	3,2044	9,9004	7,1355
Mínimo	-1,2052	-1,6749	-1,6579	-4,7720	-1,1770
Variable	***TotZDR***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	110,52	301,02	317,97	282,94	786,70
Máximo	1.845,36	4.682,02	2.842,33	17.107,89	7.706,37
Mínimo	-198,68	-247,28	-209,06	-672,50	-181,14

3.2.1.4. H_{DR}

El promedio de los valores máximos de H_{DR} muestran una marcada diferencia de la clase "Sin Daño" con respecto del resto, coincidiendo con la bibliografía donde se marca que a mayores valores de H_{DR}, mayor probabilidad de granizo y mayor tamaño de granizo ([59], [80]). Los máximos valores de MxHDR están por debajo de los 50 dB que encontró [33] asociado al daño "Grave" en el radar de INTA Paraná (tabla 26). El análisis ANOVA presenta a las variables AvHDR (Pr(>F)=0.0364), MxHDR (Pr(>F)=0.0246) y TotHDR (Pr(>F)=1.32e-06) con diferencias significativas entre las clases "Grave", "Leve", "Moderado" y "Severo".

Tabla 26. Estadísticas básicas de las variables de resumen de H_{DR} en el dataset de todos los volúmen es, todas las elevaciones y sin datos de cultivos.

Variable	***MxHDR***				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	33,3	24,1	28,1	3,1	16,5
Máximo	42,0	41,5	42,5	38,5	42,0
Mínimo	9,0	-19,0	-11,0	-58,5	-22,5

Variable	*MnHDR*				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	-61,3	-55,9	-56,1	-52,4	-65,0
Máximo	-18,0	-12,5	-15,0	3,0	-45,5
Mínimo	-73,0	-78,0	-73,0	-86,0	-77,5
Variable	*AvHDR*				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	-15,601	-16,951	-12,637	-20,711	-27,461
Máximo	-2,822	3,720	0,059	3,656	0,160
Mínimo	-27,126	-60,995	-36,483	-66,078	-62,488
Variable	*TotHDR*				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	-11.3	-7.3	-6.4	-5.2	-26.8
Máximo	-175,0	479,9	8,2	474,9	19,2
Mínimo	-24.929,2	-57.078,7	-28.072,0	-111.540,8	-67.487,7

3.2.1.5. E

Esta variable es calculada con una función de peso, que asigna el cero cuando la Z es menor a un umbral especificado, por esta razón la variable MnEWt y los mínimos de las otras tres variables valen 0 en todos los casos y no se presentan en la tabla 27. El análisis de ANOVA de las tres variables, muestra diferencias significativas (Pr(>F)=0,0961) entre las clases solo para TotEWt.

Tabla 27. Estadísticas básicas de las variables de resumen de E en el dataset de todos los volúmenes, todas las elevaciones y con datos de cultivos.

Variable	*MxEWt*				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	442,26	772,46	987,55	646,67	437,37
Máximo	2.931,2	10.613,1	8.905,6	9.565,9	4.891,8
Variable	*AvEWt*				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	13,43	19,92	20,43	15,98	9,96
Máximo	75,13	550,28	109,62	446,75	119,84
Variable	*TotEWt*				
Medida	**Grave**	**Leve**	**Moderado**	**Sin Daño**	**Severo**
Promedio	774,77	1.130,47	1.605,00	1.139,39	633,17
Máximo	6.635,33	12.659,61	11.400,67	12.955,88	5.675,55

3.2.1.6. Fenología, Cultivo y Tipo Cultivo

Las variables Fenología, Cultivos y TipoCultivo son categóricas, el estado fenológico se agrupo en tres categorías (tabla 28), el estado "Reproductivo" es el que

mayor cantidad de casos presenta en total y se debe a las fechas incluidas en el análisis. En Cultivos se detectaron ocho categorías (tabla 29), de las cuales seis corresponden a cultivos de cosecha, una a forrajeras y una a otros usos del suelo. "Girasol" y "Soja" son las dos categorías con mayor presencia, nuevamente debido a las fechas incluidas en el estudio. La variable TipoCultivo agrupa los cultivos en cosecha Fina o Gruesa, siendo esta última la de mayor cantidad de casos. El detalle de la distribución según el target se presenta en las tablas 28 a 30.

La cantidad de casos presentes por clase y categorías pueden ser insuficientes para analizar el aporte que este tipo de variable puede hacer al modelo.

Tabla 28. Estadísticas básicas de las variables de resumen de Fenología en el dataset de todos los volúmenes, todas las elevaciones y con datos de cultivos.

Fenología	Sin Daño	Leve	Moderado	Severo	Grave
Madurez	0 (0%)	14 (74%)	2 (11%)	3 (16%)	0 (0%)
Reproductivo	73 (16%)	299 (64%)	50 (11%)	27 (6%)	21 (4%)
Vegetativo	16 (10%)	97 (60%)	10 (6%)	13 (8%)	26 (16%)
Sin Dato	8 (20%)	26 (65%)	1 (3%)	3 (8%)	2 (5%)

Tabla 29. Estadísticas básicas de las variables de resumen de Cultivos en el dataset de todos los volúmenes, todas las elevaciones y con datos de cultivos.

Cultivos	Sin Daño	Leve	Moderado	Severo	Grave
Avena	0 (0%)	25 (81%)	1 (3%)	310%	2 (6%)
Cebada	1 (2%)	32 (67%)	11 (23%)	12%	3 (6%)
Girasol	28 (17%)	108 (67%)	12 (7%)	85%	5 (3%)
Monte	8 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Maíz	13 (18%)	38 (51%)	6 (8%)	7 (9%)	10 (14%)
Sorgo	6 (38%)	10 (63%)	0 (0%)	0 (0%)	0 (0%)
Soja	30 (16%)	114 (60%)	17 (9%)	11 (6%)	17 (9%)
Trigo	11 (7%)	108 (66%)	16 (10%)	16 (10%)	12 (7%)

Tabla 30. Estadísticas básicas de las variables de resumen de Tipo Cultivos en el dataset de todos los volúmenes, todas las elevaciones y con datos de cultivos.

Tipo Cultivo	Sin Daño	Leve	Moderado	Severo	Grave
Fina	12 (5%)	166 (68%)	28 (12%)	20 (8%)	17 (7%)
Gruesa	77 (18%)	270 (61%)	35 (8%)	26 (6%)	32 (7%)
Sin Dato	8 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

3.2.2. Modelos Obtenidos

3.2.2.1. Target con cinco clases

La tablas 31 a 34 presentan los resultados de los modelos obtenidos en los dataset de testing que en general fallan en diferenciar las clases de daño. En la tabla 35 se muestran los cuatro mejores modelos y se aprecia que presentan una importante cantidad de falsos positivos (FAR > 40%) para todas las clases. El agregado de las variables de cultivos parece importante para clasificar las categorías "Moderado" y "Severo", debido a la reducida cantidad de casos con los que se construyen y validan estos modelos, y a la subespecificación en el cálculo de las variables es necesario generar un conjunto de datos con mayor cantidad de casos e información más detalla para profundizar el análisis del aporte de estas variables. Los cuatro modelos presentan una buena AUC (> 74%), y excelentes valores de Especificidad (>87%), Error Rate (< 13%), PC (>86%) y NPV (>95%); que evidencian un buen rendimiento en la identificación de los casos negativos; comportamiento esperable dada la mayor proporción de este tipo de casos. Todos los modelos utilizaron los dataset con los datos de los volúmenes completos. La clase "Grave" es la única que presenta dos modelos que "empatan" en su performance con mejores valores en seis medidas de rendimiento cada uno, uno de los modelos utilizó variables referidas a la primera elevación y el otro a variables calculadas con todas las elevaciones. Las clases "Leve" y "Moderado" seleccionaron variables construidas con datos de todas las elevaciones, mientras que "Severo" eligió variables que resumían datos de la primera elevación. Las clases "Moderado" y "Severo" son las únicas que utilizaron datos relacionados con los cultivos. Solamente la clase "Grave" utilizó variables relacionadas a H_{DR}, coincidiendo con la observación realizado por [33] donde H_{DR} es un buen indicador del daño más alto. Se destaca que en todos los modelos generados con variables de cultivos se seleccionó E, apoyando la idea de [47] que esta variable es un buen indicador del daño y que mayores valores de Z están relacionados con mayor daño en terreno, concordando en este aspecto con [33]. No se presentaron diferencias entre los modelos generados con la función "ROC Measure" de aquellos generados con la función "Máxima Verosimilitud".

Tabla 31. Resultados del modelo de clasificación para la clase LEVE (Testing).

	Datos de volúmenes completos						Datos de volúmenes completos e incompletos			
	Primera Elevación			Todas las Elevaciones			Primera Elevación		Todas las elevaciones	
	Con Datos Cultivos	Sin Datos Cultivos		Con Datos Cultivos	Sin Datos Cultivos		Con Datos Cultivos	Sin Datos Cultivos	Con Datos Cultivos	Sin Datos Cultivos
		Sin H_{DR}	Con H_{DR}		Sin H_{DR}	Con H_{DR}				
Variables	19	12	16	19	12	12	19	12	18	12
Training	188	896	896	188	976	976	449	1.314	461	1.408
Testing	94	447	447	94	488	488	224	657	230	703
Variables seleccionadas	MxdBZ1 MndBZ1 TotdBZ1 MnRho1 TotRho1 AvRho1 MxZDR1 TotZDR1 AvZDR1 TotEWt MxEWt Cultivos	MxdBZ1 MndBZ1 TotdBZ1 AvdBZ1 MnRho1 TotRho1 AvRho1 MnZDR1 TotZDR1	MxdBZ1 AvdBZ1 MnRho1 AvRho1 MnZDR1 TotZDR1 TotHDR	MndBZ TotdBZ AvdBZ MnZDR TotZDR AvZDR Fenologia tipoCultivo Cultivos TotEWt AvEWt	MxdBZ, MndBZ, TotdBZ, AvdBZ, MxRho, MnRho, TotRho, MxZDR MnZDR TotZDR	MxdBZ MndBZ, TotdBZ, AvdBZ, MnRho, TotRho, AvRho, MnZDR, TotZDR, AvZDR MxHDR	MndBZ1,AvdBZ1, MxRho1, MnRho1, TotRho1, AvRho1, MxZDR1, MnZDR1, TotZDR1, Fenologia, Cultivos, TotEWt, AvEWt	MxdBZ1, MndBZ1, TotdBZ1, MxRho1, AvRho1, MnZDR1, TotZDR1, AvZDR1	MndBZ, AvdBZ, MxRho, TotRho, AvRho, MxZDR, TotZDR, AvZDR, Cultivos, TotEWt, AvEWt	MxdBZ, AvdBZ, MxRho, MnRho, MxZDR, MnZDR, TotZDR, AvZDR
FAR	0.2152	0.6595	0.6026	0.1944	0.4828	0.5714	0.3107	0.4859	0.2394	0.5299
CSI	0.7381	0.3347	0.3563	0.7073	0.4545	0.3696	0.6289	0.4555	0.7044	0.4245
GS	0.2055	0.1862	0.2333	0.1977	0.3681	0.2806	0.1445	0.3057	0.1876	0.2928
HSS	0.3410	0.3139	0.3783	0.3302	0.5381	0.4382	0.2526	0.4682	0.3159	0.4529
AUC	0.7402	0.8388	0.8039	0.6742	0.8636	0.8435	0.6176	0.8238	0.6857	0.8441
POD	0.9254	0.9518	0.7750	0.8529	0.7895	0.7286	0.8777	0.8000	0.9051	0.8138
PC	0.7660	0.6488	0.7494	0.7447	0.8525	0.8217	0.6786	0.7671	0.7391	0.7724
Especi	0.3704	0.5797	0.7439	0.4615	0.8641	0.8373	0.3529	0.7565	0.3750	0.7616
Error Rate	0.2340	0.3512	0.2506	0.2553	0.1475	0.1783	0.3214	0.2329	0.2609	0.2276
SR,PPV	0.7848	0.3405	0.3974	0.8056	0.5172	0.4286	0.6893	0.5141	0.7606	0.4701
NPV	0.6667	0.9814	0.9381	0.5455	0.9570	0.9485	0.6383	0.9216	0.6429	0.9403
F1	0.8493	0.5016	0.5254	0.8286	0.6477	0.5396	0.7722	0.6259	0.8266	0.5959

Tabla 32. Resultados del modelo de clasificación para la clase MODERADO (Testing).

	Datos de volúmenes completos						Datos de volúmenes completos e incompletos			
	Primera Elevación			Todas las Elevaciones			Primera Elevación		Todas las elevaciones	
	Con Datos Cultivos	Sin Datos Cultivos		Con Datos Cultivos	Sin Datos Cultivos		Con Datos Cultivos	Sin Datos Cultivos	Con Datos Cultivos	Sin Datos Cultivos
		Sin H_{DR}	Con H_{DR}		Sin H_{DR}	Con H_{DR}				
Variables	19	12	16	19	12	12	19	12	18	12
Training	188	896	896	188	976	976	449	1.314	461	1.408
Testing	94	447	447	94	488	488	224	657	230	703
Variables seleccionadas	MxdBZ1 MndBZ1 TotdBZ1 AvdBZ1 MnRho1 TotRho1 AvRho1 MnEWt TipoCultivo Cultivos	MxdBZ1 MndBZ1 TotdBZ1 AvdBZ1 MxRho1 MnRho1 TotRho1 AvRho1 MxZDR1 MnZDR1 AvZDR1	MxdBZ1 TotdBZ1 AvdBZ1 MnRho1 TotRho1 MxZDR1 MnZDR1TotZDR1 AvZDR1 AvHDR MxHDR, MnHDR	TotdBZ AvdBZ MxRho MxZDR MnZDR AvZDR Fenologia tipoCultivo Cultivos TotEWt AvEWt MxEWt	MxdBZ MndBZ TotdBZ AvdBZ MxRho MnRho TotRho AvRho MxZDR MnZDR TotZDR	MxdBZ MndBZ TotdBZ MxRho MnRho TotRho AvRho MnZDR AvZDR MxHDR	MxdBZ1, AvdBZ1, MnRho1, TotRho1, AvRho1, MxZDR1, MnZDR1, TotZDR1, Fenologia, TipoCultivo, Cultivos, AvEWt, MxEWt	MxdBZ1, MndBZ1, TotdBZ1, AvdBZ1, MxRho1, MnRho1, TotRho1, AvRho1, MxZDR1, MnZDR1, TotZDR1, AvZDR1	MxdBZ, MnRho, MxZDR, MnZDR, AvZDR, Fenologia, Cultivos, TotEWt	MxdBZ, MndBZ, AvdBZ, MxRho, TotRho, AvRho, MxZDR, MnZDR, TotZDR, AvZDR
FAR	0.6923	0.9275	0.9412	0.5263	0.8919	0.9149	0.8161	0.9277	0.9167	0.8808
CSI	0.2353	0.0704	0.0571	0.4091	0.1013	0.0842	0.1758	0.0710	0.0759	0.1168
GS	0.1821	0.0561	0.0383	0.3359	0.0783	0.0672	0.0989	0.0497	0.0258	0.0822
HSS	0.3080	0.1062	0.0738	0.5028	0.1452	0.1259	0.1800	0.0947	0.0502	0.1519
AUC	0.7398	0.8954	0.7825	0.7947	0.8477	0.8914	0.7620	0.8440	0.6044	0.8339
POD	0.5000	0.7143	0.6667	0.7500	0.6154	0.8889	0.8000	0.8000	0.4615	0.8519
PC	0.8617	0.8523	0.7785	0.8617	0.8545	0.8217	0.6652	0.7610	0.6826	0.7525
Especi	0.8953	0.8545	0.7808	0.8780	0.8611	0.8205	0.6520	0.7601	0.6959	0.7485
Error Rate	0.1383	0.1477	0.2215	0.1383	0.1455	0.1783	0.3348	0.2390	0.3174	0.2475
SR,PPV	0.3077	0.0725	0.0588	0.4737	0.1081	0.0851	0.1839	0.0723	0.0833	0.1192
NPV	0.9506	0.9947	0.9913	0.9600	0.9879	0.9975	0.9708	0.9939	0.9557	0.9922
F1	0.3809	0.1316	0.1081	0.5806	0.1839	0.1553	0.2990	0.1326	0.1412	0.2091

Tabla 33. Resultados del modelo de clasificación para la clase SEVERO (Testing).

	Datos de volúmenes completos						Datos de volúmenes completos e incompletos			
	Primera Elevación			Todas las Elevaciones			Primera Elevación		Todas las elevaciones	
	Con Datos Cultivos	Sin Datos Cultivos		Con Datos Cultivos	Sin Datos Cultivos		Con Datos Cultivos	Sin Datos Cultivos	Con Datos Cultivos	Sin Datos Cultivos
		Sin H_{DR}	Con H_{DR}		Sin H_{DR}	Con H_{DR}				
Variables	19	12	16	19	12	12	19	12	18	12
Training	188	896	896	188	976	976	449	1.314	461	1.408
Testing	94	447	447	94	488	488	224	657	230	703
Variables seleccionadas	TotdBZ1 AvdBZ1 MxRho1 TotRho1 AvRho1 MnZDR1 TotZDR1 AvZDR1 TotEWt AvEWt MnEWt Cultivos	TotdBZ1 AvdBZ1 TotRho1 AvRho1 MxZDR1 TotZDR1 AvZDR1	MxdBZ1 MndBZ1 TotdBZ1 AvdBZ1 MnRho1 MxZDR1 MnZDR1 TotZDR1 AvZDR1 AvHDR MxHDR MnHDR TotHDR	MxdBZ MndBZ TotdBZ AvdBZ TotRho AvRho MnZDR TotZDR AvZDR tipoCultivo, Cultivos TotEWt MxEWt	MndBZ TotdBZ AvdBZ MxRho MnRho AvRho TotZDR AvZDR	MndBZ AvdBZ TotRho AvRho MxZDR MnZDR AvZDR AvHDR MxHDR MnHDR TotHDR	MxRho1 TotRho1 MxZDR1 MnZDR1 Fenologia Cultivos AvEWt	MxdBZ1 MndBZ1 TotdBZ1 AvdBZ1 MxRho1 MnRho1 TotRho1 AvRho1 MxZDR1 MnZDR1 TotZDR1 AvZDR1	MndBZ AvdBZ MxRho MnRho TotRho AvRho MxZDR TotZDR AvZDR Fenologia tipoCultivo Cultivos TotEWt MnEWt	MxdBZ MndBZ AvdBZ MxRho MxZDR MnZDR AvZDR
FAR	0.6000	0.9250	0.8333	0.8667	0.8125	0.8750	0.8462	0.9083	0.9158	0.8961
CSI	0.3333	0.0732	0.1538	0.1333	0.1667	0.1176	0.1500	0.0885	0.0784	0.0930
GS	0.3151	0.0650	0.2587	0.1145	0.1590	0.1073	0.0948	0.0694	0.0188	0.0730
HSS	0.4792	0.1221	0.1486	0.2055	0.2744	0.1938	0.1733	0.1297	0.0370	0.1360
AUC	0.9084	0.8516	0.9504	0.9022	0.6546	0.7659	0.7690	0.8521	0.5668	0.7589
POD	0.6667	0.7500	0.6667	1.0000	0.6000	0.6667	0.8571	0.7143	0.5333	0.4706
PC	0.9574	0.9150	0.9754	0.8617	0.9693	0.9385	0.6964	0.8432	0.5913	0.8890
Especi	0.9670	0.9165	0.9775	0.8587	0.9731	0.9419	0.6857	0.8460	0.5953	0.8994
Error Rate	0.0426	0.0850	0.0246	0.1383	0.0307	0.0615	0.3036	0.1568	0.4087	0.1110
SR,PPV	0.4000	0.0750	0.1667	0.1333	0.1875	0.1250	0.1538	0.0917	0.0842	0.1039
NPV	0.9888	0.9975	0.9775	1.0000	0.9958	0.9956	0.9863	0.9927	0.9481	0.9856
F1	0.500	0.1364	0.2666	0.2353	0.2857	0.2105	0.2609	0.1626	0.1454	0.1702

Tabla 34. Resultados del modelo de clasificación para la clase GRAVE (Testing).

	Datos de volúmenes completos						Datos de volúmenes completos e incompletos			
	Primera Elevación			Todas las Elevaciones			Primera Elevación		Todas las elevaciones	
	Con Datos Cultivos	Sin Datos Cultivos		Con Datos Cultivos	Sin Datos Cultivos		Con Datos Cultivos	Sin Datos Cultivos	Con Datos Cultivos	Sin Datos Cultivos
		Sin H_{DR}	Con H_{DR}		Sin H_{DR}	Con H_{DR}				
Variables	19	12	16	19	12	12	19	12	18	12
Training	188	896	896	188	976	976	449	1.314	461	1.408
Testing	94	447	447	94	488	488	224	657	230	703
Variables seleccionadas	MndBZ1, AvdBZ1, MxRho1, AvZDR1, TotEWt, Fenologia	TotdBZ1, MxRho1, MnRho1, TotRho1, MxZDR1, TotZDR1, AvZDR1	MndBZ1 TotdBZ1 AvdBZ1 MnRho1 TotRho1 AvRho1, MxZDR1 MnZDR1 AvZDR1 AvHDR MxHDR MnHDR TotHDR	MndBZ, TotdBZ, TotRho, AvRho, MxZDR, MnZDR, TotZDR, Cultivos, TotEWt, AvEWt, MxEWt, MnEWt	MxdBZ, TotdBZ, AvdBZ, MxRho, TotRho, AvRho, MxZDR, MnZDR, TotZDR, AvZDR	MxdBZ TotdBZ MxRho TotRho MxZDR TotZDR MnHDR TotHDR	MndBZ1, TotdBZ1, MnRho1, AvRho1, MnZDR1, AvZDR1, Fenologia, Cultivos, AvEWt	MxdBZ1, MndBZ1, TotdBZ1, AvdBZ1, MxRho1, TotRho1, AvRho1, MnZDR1, TotZDR1, AvZDR1	MxdBZ, MndBZ, AvdBZ, MxRho, TotRho, MxZDR, TotZDR, AvZDR, Fenologia, tipoCultivo, Cultivos, AvEWt, MxEWt, MnEWt	MxdBZ, MndBZ, AvdBZ, MnRho, TotRho, MnZDR, TotZDR, AvZDR
FAR	0.9474	0.8889	0.7143	0.6667	0.6000	0.9231	0.6087	0.8690	0.8364	0.7586
CSI	0.0435	0.1071	0.2857	0.1667	0.2857	0.0769	0.2903	0.1209	0.1552	0.2222
GS	-0.0005	0.0994	0.2825	0.1486	0.2815	0.0712	0.2480	0.0981	0.1112	0.2024
HSS	-0.0010	0.1808	0.4406	0.2587	0.4393	0.1330	0.3974	0.1786	0.2001	0.3366
AUC	0.6315	0.7895	0.9932	0.9028	0.7438	0.9656	0.7798	0.7819	0.8237	0.9408
POD	0.2000	0.7500	1.0000	0.2500	0.5000	1.0000	0.5294	0.6111	0.7500	0.7368
PC	0.7660	0.9441	0.9888	0.9468	0.9898	0.9262	0.9018	0.8782	0.7870	0.9303
Especi	0.7978	0.9458	0.9888	0.9778	0.9938	0.9258	0.9324	0.8858	0.7890	0.9357
Error Rate	0.2340	0.0559	0.0112	0.0532	0.0102	0.0738	0.0982	0.1218	0.2130	0.0697
SR,PPV	0.0526	0.1111	0.2857	0.3333	0.4000	0.0769	0.3913	0.1310	0.1636	0.2414
NPV	0.9467	0.9976	1.0000	0.9670	0.9959	1.0000	0.9602	0.9878	0.9829	0.9922
F1	0.0833	0.1935	0.4444	0.2857	0.4444	0.1428	0.4500	0.2157	0.2686	0.3636

Tabla 35. Mejores modelos seleccionados para cada clase (Testing).

	Leve	**Moderado**	**Severo**	**Grave**	
	Volúmenes Completos				
	Todas las Elevaciones		Primera Elevación		TE
	Sin Cultivos	Con Cultivos		Sin Cultivos	
	Sin H_{DR}			Con H_{DR}	Sin H_{DR}
Variables	10	12	12	13	10
Variables seleccionadas	MxdBZ, MndBZ, TotdBZ, AvdBZ, MxRho, MnRho, TotRho, MxZDR MnZDR TotZDR	TotdBZ, AvdBZ, MxRho, MxZDR, MnZDR, AvZDR, Fenologia, tipoCultivo, Cultivos, TotEWt,AvEWt, MxEWt	TotdBZ1, AvdBZ1, MxRho1, TotRho1, AvRho1, MnZDR1, TotZDR1, AvZDR1, TotEWt, AvEWt, MnEWt, Cultivos	MndBZ1,TotdBZ1, AvdBZ1, MnRho1, TotRho1, AvRho1, MxZDR1, MnZDR1, AvZDR1, AvHDR, MxHDR, MnHDR, TotHDR	MxdBZ, TotdBZ, AvdBZ, MxRho, TotRho, AvRho, MxZDR, MnZDR, TotZDR, AvZDR
FAR	0.4300	0.5263	0.6000	0.7143	0.6000
CSI	0.4790	0.4091	0.3333	0.2857	0.2857
GS	0.4005	0.3359	0.3151	0.2825	0.2815
HSS	0.5720	0.5028	0.4792	0.4406	0.4393
AUC	0.8636	0.7947	0.9084	0.9932	0.7438
POD	0.7500	0.7500	0.6667	1.0000	0.5000
PC	0.8730	0.8617	0.9574	0.9888	0.9898
Especi	0.8956	0.8780	0.9670	0.9888	0.9938
Error Rate	0.1270	0.1383	0.0426	0.0112	0.0102
SR,PPV	0.5700	0.4737	0.4000	0.2857	0.4000
NPV	0.9510	0.9600	0.9888	1.0000	0.9959
F1	0.6477	0.5806	0.5000	0.4444	0.4444

3.2.2.2. Target con tres clases

La tabla 36 presenta los resultados de los modelos generados para el target daño con tres clases en el dataset de testing. La clase "Más de 50%", tampoco presentó buenos rendimientos (FAR=65%, POD=58%), mientras que las clases "Sin Daño" y "Menos del 50%" mostraron muy buenos valores de FAR < 13% y POD > 94%, reforzando la idea que estas herramientas son útiles para generar modelos sobre el daño en cultivos en dataset con más casos. Se destaca que la variable H_{DR} fue seleccionada en todos los modelos, por lo que se debe incluir en futuros estudios.

En el anexo 3 se presentan los árboles de expresión del mejor modelo de cada clase y las funciones logísticas finales.

Tabla 36. Resultados modelos Daño con tres clases (Testing).

	Datos de volúmenes completos			Datos de volúmenes completos e incompletos		
	Variables Polarimétricas			Variables Polarimétricas		
	Todas las elevaciones			Todas las elevaciones		
	Sin Cultivos			Sin Cultivos		
	Con HDR			Con HDR		
Clase	Sin Daño	Menos de 50%	Más de 50%	Sin Daño	Menos de 50%	Más de 50%
Variables	16	16	16	16	16	16
Training	976	976	976	1.407	1.407	1.407
Testing	488	488	488	703	703	703
Variables seleccionadas	AvdBZ, MxRho, MnRho, TotRho, MnZDR, TotZDR, MxHDR	MxdBZ, MndBZ, AvdBZ, TotZDR, AvHDR, MnHDR	MxdBZ, TotdBZ, AvRho, MnZDR, TotZDR, AvHDR, MxHDR, MnHDR, TotHDR	MxdBZ, AvRho, MnZDR, AvHDR, MxHDR, MnHDR, TotHDR	MxdBZ, AvdBZ, MxRho, MnRho, AvZDR, AvHDR, MxHDR	MxdBZ, MndBZ, AvdBZ, MnRho, AvRho, MxZDR, TotZDR, AvZDR, MxHDR, MnHDR, TotdBZ
FAR	0.0957	**0.1342**	0.9333	0.1359	0.4078	*0.6557*
CSI	0.8750	**0.8562**	0.0625	0.8056	0.4473	*0.2763*
GS	0.4388	0.3006	0.0552	0.4160	**0.3290**	*0.2453*
HSS	0.6100	0.4622	0.1047	0.5875	**0.4952**	*0.3940*
AUC	0.8342	**0.8400**	0.8171	0.8632	0.8337	*0.8322*
POD	0.9643	**0.9872**	0.5000	0.9225	0.6463	*0.5833*
PC	0.8893	**0.8668**	*0.9385*	0.8407	0.8137	0.9218
Especi	0.5833	0.3750	*0.9421*	0.6350	**0.8646**	0.9400
Error Rate	0.1107	**0.1332**	*0.0615*	0.1593	0.1863	0.0782
SR,PPV	0.9043	**0.8658**	0.0667	0.8641	0.5922	*0.3443*
NPV	0.8000	0.8780	*0.9956*	0.7651	**0.8893**	0.9766
F1	0.9333	**0.9225**	0.1176	0.8923	0.6180	*0.4329*

3.3. Resultados Implementación

3.3.1. Opción 1: Base de Datos

El esquema de implementación se presenta en la figura 35. Los pasos 1 a 3 se detallaron en las secciones 2.2.2 y 2.4.1 de este trabajo. En el paso 4 se toman los datos de las imágenes GeoTIFF, si se tiene un listado de casos a clasificar, solo se leen los datos correspondientes a los píxeles donde se ubican geográficamente esos casos (script de Python *ProcesarImg10minutos.py*); de lo contrario se lee toda la imagen (script de Python *ProcesarImgCompleta10minutos.py)*.

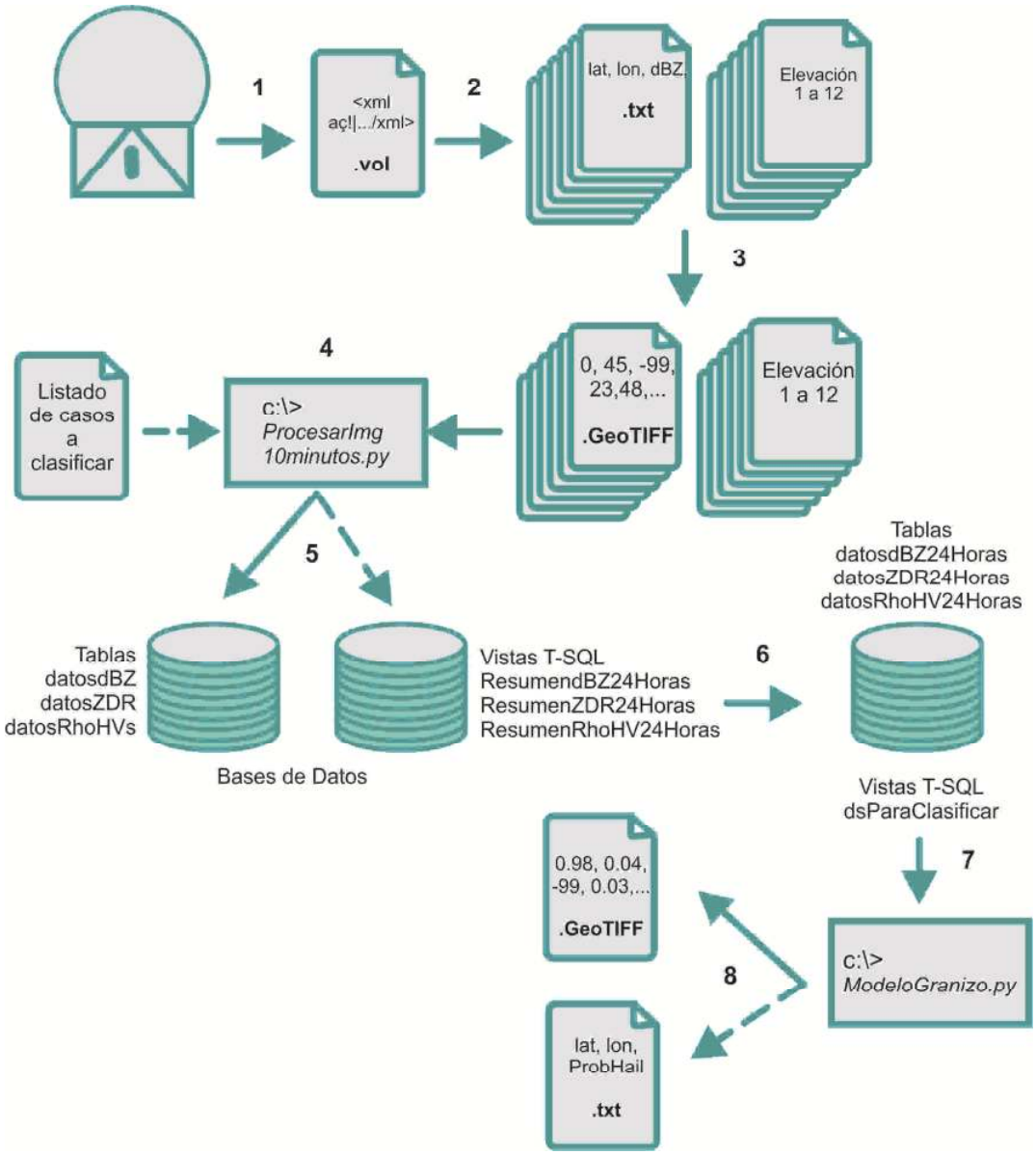


Figura 35. Esquema de la solución de implementación del modelo de clasificación de Granizo.

Los datos se almacenan en una base de datos creada para la implementación, en el Anexo 2 se presentan los detalles. Los datos de una sola elevación completa (144 tomas) requieren un espacio mayor a los 10 Gb que permite administrar la herramienta SQL Server Express R2 en una sola base de datos (figura 36.a), esta situación requirió generar otra base de datos, con el mismo esquema, que se utiliza cuando la primera base de datos se llena. Cada base de datos puede almacenar hasta 142 elevaciones completas y su procesamiento requiere 52 horas de cómputo en promedio (paso 5).

Posteriormente, se ejecutó una tarea de exportación de los resultados de las vistas T-SQL *ResumendBZ24Horas, ResumenZDR24Horas* y *ResumenRhoHV24Horas* a las tablas *datosdBZ24Horas, datosZDR24Horas, datosRhoHV24Horas* (paso 6). Esta tarea fue necesaria porque la gran cantidad de datos que se deben procesar para generar las variables de entrada al modelo, genera un error de *time out*, al hacer una consulta o una vista (figura 36.b). Esta exportación consume 15 minutos de procesamiento.

Figura 36. Errores de espacio (a) y de tiempos de espera (b) debido al volumen de datos a trabajar.

El paso 7 es la implementación del modelo, el script de Python *ModeloGranizo.py,* lee las variables de entrada desde la vista *dsParaClasificar* y genera un archivo de texto cuando debe clasificar un listado de casos o una imagen GeoTIFF si debe clasificar toda el área del radar (paso 8). La figura 37.a y 37.b muestra como se ve el resultado de clasificar un listado de casos y la figura 37.c y 37.d muestra cómo se ven las imágenes del resultado de clasificar toda el área del radar. Un listado de 3.000 casos, con el modelo 3 (tabla 17), fue procesado en 2,5 minutos, el listado de 180.250 casos, con el modelo 1 (tabla 17), fue procesado en 47 minutos y un listado de 244.944 casos, con el modelo 3 (tabla 17), se procesó en 65 minutos. El tiempo total de procesamiento para esta solución es de 3,5 días aproximadamente, el detalle se presenta en la tabla 37.

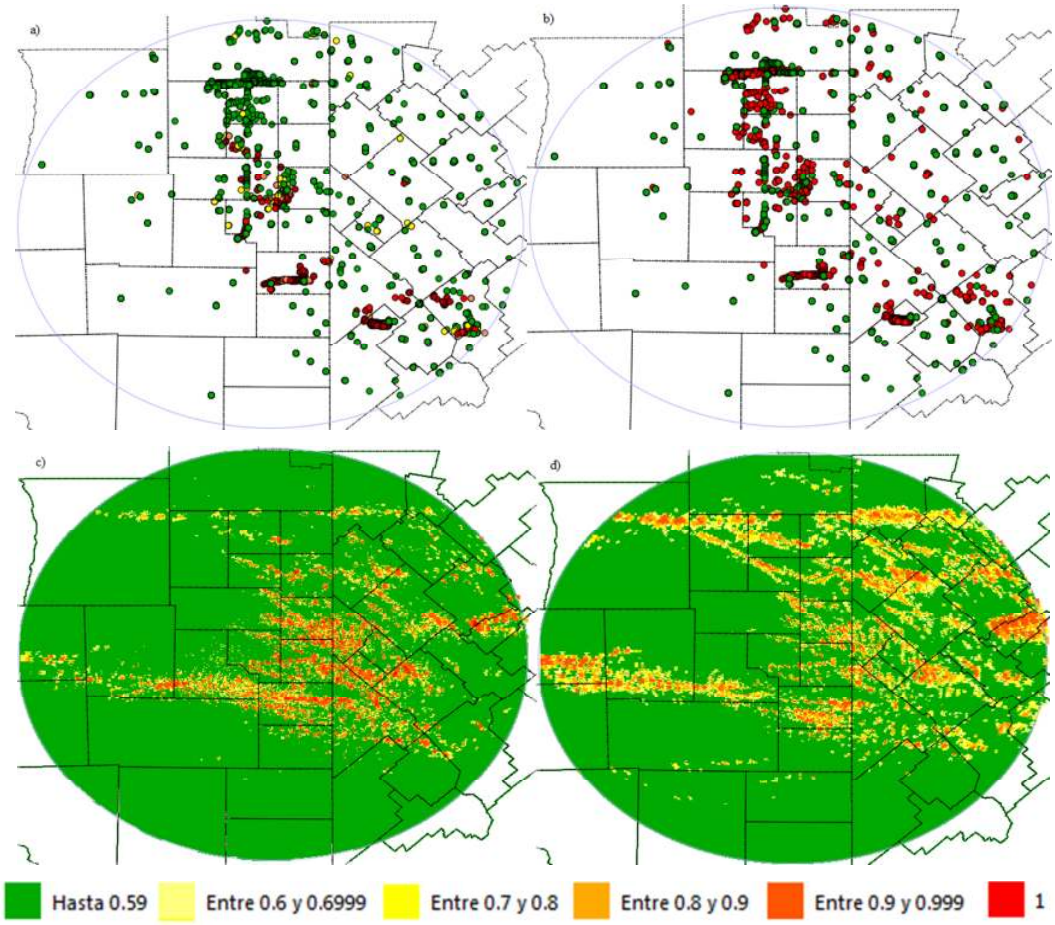


Figura 37. Resultados de la implementación de los modelos de clasificación de granizo. a) Resultado modelo primera elevación con -99.00. b) Verdad de campo de los lotes. c) Imagen completa del 10-12-2012 modelo granizo primera elevación (modelo 1). d) Imagen completa del 10-12-2012 modelo granizo primera elevación con -99.00 (modelo 3).

Tabla 37. Detalle del tiempo de procesamiento aproximado para cada paso de la implementación con base de datos del modelo de clasificación de granizo (figura 35).

Tarea	Descripción	Minutos de procesamiento
Paso 1	Descarga .vol	2 minutos
Paso 2	Conversión de .vol a ASCII (432 vol con 12 elevaciones)	882 minutos
Paso 3	Conversión de ASCII a GeoTIFF (144 ASCII, solo primera elevación)	1.008 minutos.
Paso 4	Datos GeoTIFF a base de datos (imagen	3.120 minutos

	completa de una sola elevación)	
Paso 5	Generar variables de resumen de 24 horas y exportar a una tabla.	15 minutos.
Paso 6	Aplicar modelo a 244.944 casos	65 minutos
Total		**5.092 minutos** **84,86 horas** **3,5 días**

3.3.2. Opción 2: Imágenes Compuestas

El esquema de implementación de esta opción se detalla en la figura 38. Nuevamente los pasos 1 a 3 se detallaron en las secciones 2.2.2 y 2.4.1 de este trabajo. El script *GIC.py* se utilizó para generar los valores máximos, mínimos, promedios y totales de la primera elevación del 10-12-2012, estos totales se calcularon sin los valores perdidos y reemplazando los valores perdidos por -99.00. La figura 39 muestra seis ejemplos del resultado de imágenes compuestas de máximos y mínimos de Z, Rho_{HV} y Z_{DR}, correspondientes a la primera elevación (0.5º) y 24 horas para tres fechas del 2012 con tormentas. La figura 40 muestra el PPI de las 12 elevaciones del 15-01-2011 a las 23.40 h, también generado con *GIC.py*. Todas estas imágenes de ejemplo se calcularon sin valores perdidos. Los valores calculados por medio de este script se compararon con los valores obtenidos por el cálculo con sentencias T-SQL. El único cálculo que presentó diferencias fue el del promedio, este problema puede deberse al manejo de la precisión de los números con coma flotante; este inconveniente está documentado para numpy [35] y se debe tener en cuenta a la hora de generar y calibrar los modelos. Este paso insume un minuto de procesamiento. A pesar de las diferencias menores en los valores de promedio, se aplicó el modelo 3 para medir la velocidad de procesamiento. El algoritmo consumió 2,5 minutos de cómputo. La imagen clasificada se presenta en la figura 41. Esta solución consume en total 1,3 días de procesamiento, lo cual mejora la velocidad de procesamiento de la solución de base de datos.

[35] Ver: http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html

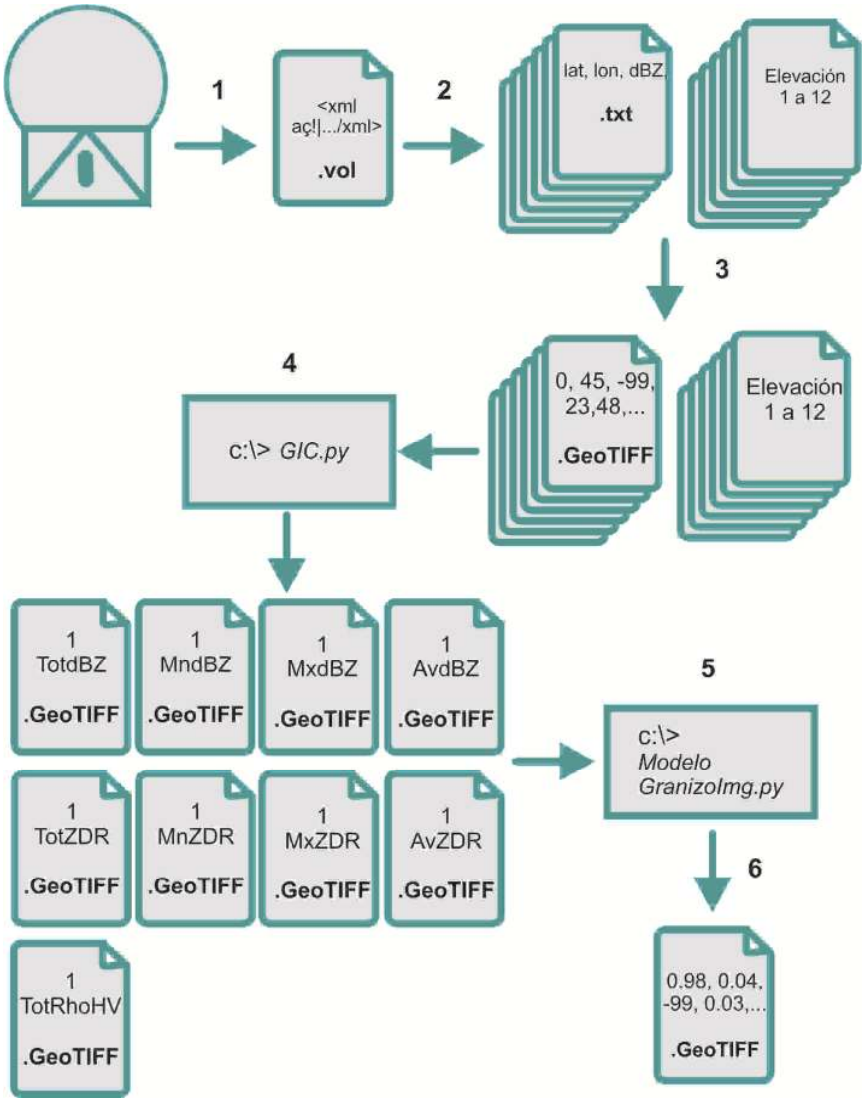


Figura 38. Esquema de la solución de implementación del modelo de clasificación de Granizo.

Tabla 38. Detalle del tiempo de procesamiento aproximado para cada paso de la implementación con imágenes compuestas del modelo de clasificación de granizo.

Tarea (esquema figura 37)	**Descripción**	**Minutos de procesamiento**
Paso 1	Descarga .vol	2 minutos
Paso 2	Conversión de .vol a ASCII (432 vol con 12 elevaciones)	882 minutos
Paso 3	Conversión de ASCII a GeoTIFF (144 ASCII, solo primera elevación)	1.008 minutos.
Paso 4	Generar imágenes compuestas de máximo, mínimo, promedio y total de dBZ, Rho_{HV} y	1 minuto

	Z_{DR}.	
Paso 5 y 6	Aplicar modelo a 244.944 casos y generar imagen clasificada	2,5 minutos
Total		**1895,5 minutos 31, 5horas 1,3 días**

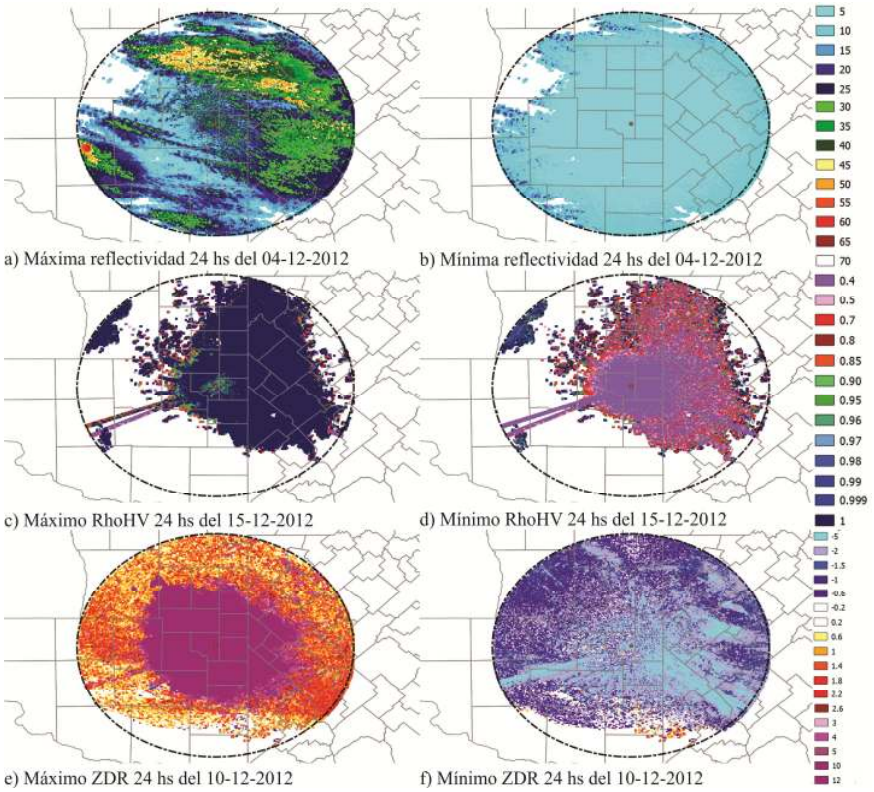


Figura. 39. Ejemplos de las imágenes compuestas de la primera elevación (0.5º) y 24 horas generadas con el script *GIC.py* para tres fechas del 2012 con tormentas.

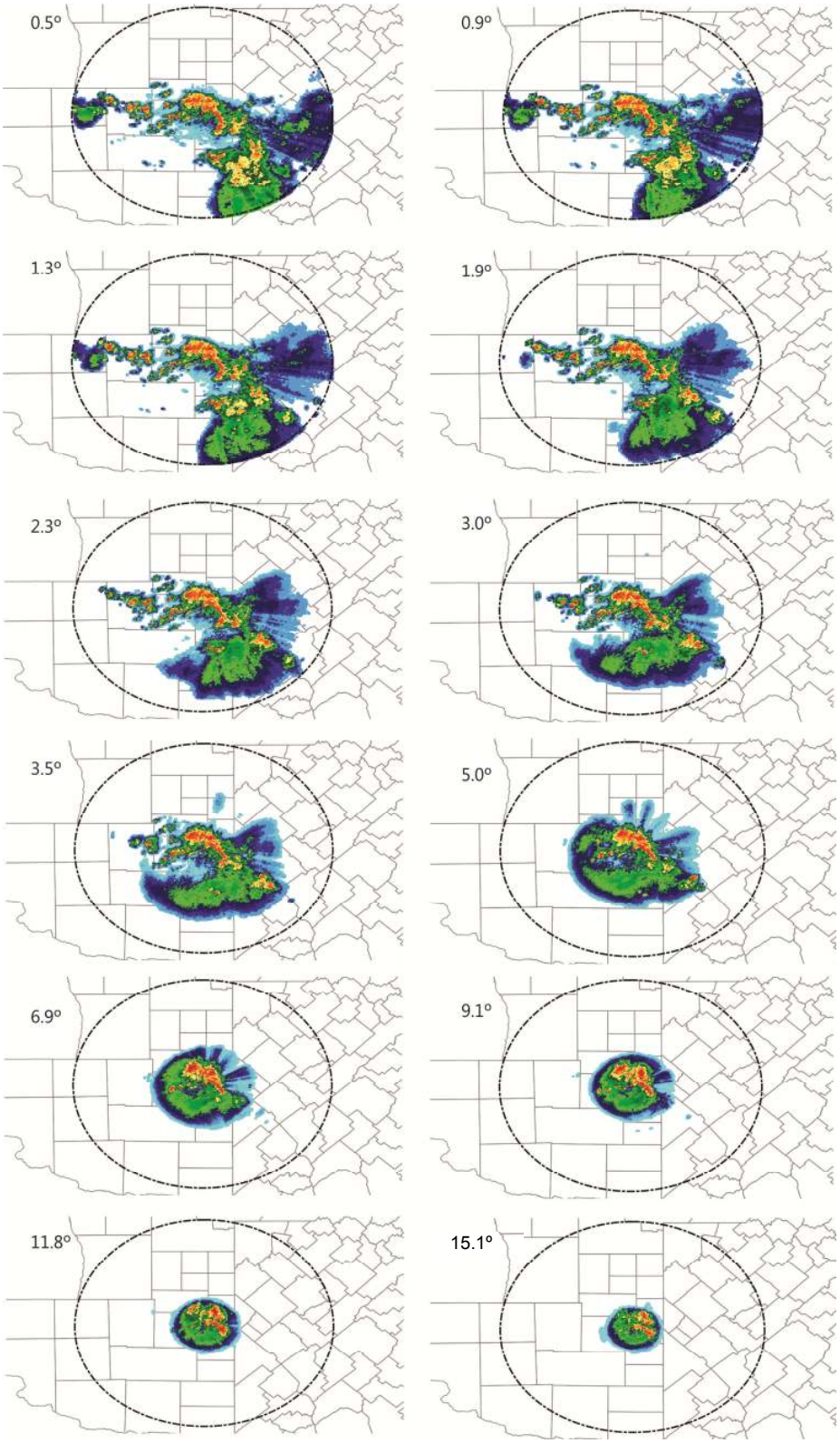


Figura 40. PPI de las 12 elevaciones del 15-01-2011 a las 23.40 h.

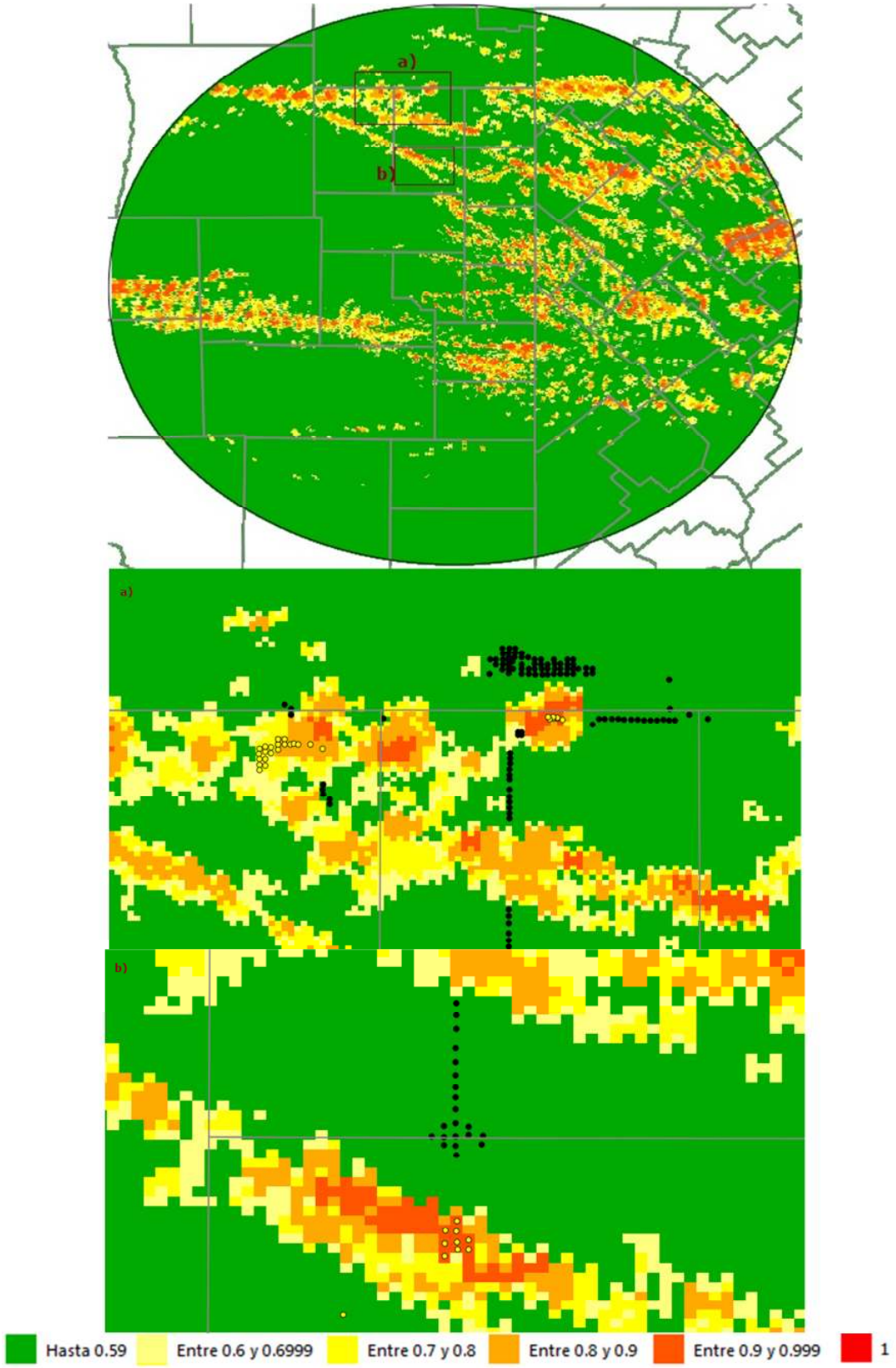


Figura 41. Resultados de la implementación de la clasificación de granizo: imagen completa del 10-12-2012 modelo 3 de granizo. a) y b) Detalle de la clasificación de la imagen y los puntos de verdad de campo: negro sin registro de granizo, amarrillo con registro de granizo.

Capítulo 4. Conclusiones y Recomendaciones

4.1. Conclusiones

Se lograron desarrollar dos modelos de clasificación de granizo en superficie con información de un radar meteorológico de banda C con una alta performance. Los dos modelos seleccionados se nombraron como Kurá y Pire, por las palabras mapuches correspondientes a Piedra y Granizo/Nieve. Los valores de POD (84%), FAR (23% y 14%) y PC (86% y 84%) de estos dos modelos fueron superados sólo por dos algoritmos (de un total de 30 modelos ya existentes internacionalmente).

El daño también fue tratado como problemas binarios con cinco y tres clases. Además de las variables polarimétricas del radar se incorporaron variables del cultivo. Los modelos obtenidos para las cinco clases fallan en diferenciarlas. Muestran una importante cantidad de falsos positivos (FAR > 40%) para todas las clases. Estos resultados eran esperados por la subespecificación de los datos del radar intradía, la poca cantidad de casos para algunas de las clases y la falta de variables relacionadas al cultivo. Sin embargo, se obtuvieron buenos resultados en otras medidas de rendimiento (AUC > 74%, PC > 86%, NPV >95%, Error Rate < 13% y Especificidad >87%) lo que sugiere que se pueden usar estas herramientas para análisis futuros en un conjunto de datos más grande y completo. En el segundo análisis del daño con tres clases, la clase correspondiente al mayor porcentaje de daño, no presentó buenos rendimientos (FAR=65%, POD=58%), este resultado también se esperaba, por las mismas razones mencionadas en el párrafo anterior. Las otras dos clases se pueden diferenciar muy satisfactoriamente (FAR < 13% y POD > 94%), reforzando la idea que estas herramientas son útiles para generar modelos que clasifiquen el daño en cultivos.

Se pudieron utilizar como variables de entrada para los modelos, resúmenes (máximos, mínimos, promedios, totales y cantidades) de las variables Z, Z_{DR} y Rho_{HV}. También se pudieron calcular las variables E y H_{DR} derivadas de Z y Z_{DR}. Para la identificación de granizo, los modelos con variables polarimétricas funcionaron mejor que aquellos que usan las variables de simple polarización. Se observó que, a mayor valor de Z, menores valores de Rho_{HV} y valores extremos de Z_{DR}, mayor es la probabilidad de presencia de granizo. Se establece que el umbral de Z para determinar presencia de granizo cuando se trabaje con simple polarización, se debe fijar entre 50 y 55 dBZ para este radar. En el caso del Daño a cultivos, H_{DR} y E se presentan como un buen indicador del mismo.

Mayores valores de Z están relacionados con mayor daño en terreno. Para el daño más alto Z_{DR} presenta valores más cercanos a cero, que puede asociarse a la presencia de granizos grandes y en mucha cantidad, que dominan Z y aparecen como esferas perfectas; para el resto de las clases Z_{DR} presenta un promedio por encima de 4 dB, acorde al comportamiento de la banda C en presencia de granizo.

El problema planteado cuenta con la características de una problemática a resolver con DM debido: a) al gran volumen de información generada por el radar, b) la dificultad de realizar un diseño experimental en la recolección de los datos de campo (ocurrencia del granizo y daño en los cultivos), c) la necesidad de buscar relaciones no lineales como soluciones al problema y d) La baja frecuencia del caso positivo con relación al negativo. La aplicación de CRISP-DM fue adecuada para el problema a resolver y se cumplió que la etapa de preparación de los datos insumió el 70% del tiempo del proceso.

El método GEP (regresión logística) resultó ser una muy buena herramienta para clasificar la ocurrencia de granizo en superficie. Para el caso del daño, tanto la regresión logística como la clasificación arrojaron resultados similares. Para este último target se debe completar un mayor set de datos.

El software que permite la descarga, lectura, transformación a formatos estándares y abiertos y procesamiento de los datos del radar es un importante subproducto de este trabajo que aporta una herramienta que permite minimizar el uso de software propietario para el manejo de los datos y que facilita el acceso a los datos del radar para futuros estudios. Este software se utilizó exitosamente durante todo el desarrollo de este trabajo, tanto para plataforma Linux como Windows. Todo el código fuente y el detalle de su uso están publicados en https://github.com/INTA-Radar. Como el desarrollo se realizó con herramientas libres, se asegura la transparencia en el tratamiento de los datos, la posibilidad de extender y mejorar el desarrollo realizado y la disponibilidad de estas herramientas para los diferentes usuarios. El software desarrollado brinda herramientas a los usuarios de los datos de la red que permiten el acceso y procesamiento de los mismos, propiciando el uso de formatos estándares y abiertos de datos para facilitar la interacción tanto entre las herramientas de software elegidas como entre los usuarios que las utilizan, incrementando el uso y aprovechamiento de los radares meteorológicos del INTA.

El sistema de información web y la base de datos de reportes de ocurrencia de granizo en superficie se publicaron en internet en la dirección

http://rian.inta.gob.ar/DanioGranizo con el objetivo que los mismos puedan ser utilizados para otros estudios.

Se pudieron implementar tres modelos, logrando clasificar una serie de lotes e imágenes completas. Estas implementaciones demostraron inconvenientes con el almacenamiento de los datos del radar y su posterior procesamiento, tardando como mínimo 1,3 días en obtener una clasificación de una imagen completa.

4.2. Recomendaciones

Tanto para el target Granizo en superficie, como Daño en cultivos y a partir de los resultados de este primer análisis exploratorio, sería conveniente en futuros trabajos, poder aislar el momento de ocurrencia de la tormenta y trabajar solo con los datos correspondientes a ese momento, evitando de esta manera suavizar los valores de las variables del radar. Otra característica a determinar, para evitar la subespecificación, son las elevaciones que realmente aportan a la identificación de la caída en superficie y a la diferenciación del daño. También sería muy importante darle continuidad a la actualización y ampliación de la información de verdad de campo utilizando el sistema de información desarrollado para tal fin. Además, sería interesante realizar una comparación de los resultados del algoritmo ZHAIL del software Rainbow, con los resultados de los modelos generados en este trabajo para las mismas fechas y zona de estudio.

Como el daño a los cultivos está estrechamente relacionado con el cultivo y su estado fenológico, para lograr mejores modelos sería conveniente analizar los datos por cultivo o tipos de cultivos similares y por momento de ocurrencia asociándolo a la fenología del mismo. La información podría ser provista por compañías de seguro, que cuentan con el detalle del cultivo afectado y su estado fenológico.

A medida que aumenta la distancia desde el radar, el haz emitido toma datos a mayor altura (pudiendo perder datos de la tormenta) y se ensancha (aumentando el volumen de muestreo). En la estimación de otras variables como la lluvia, la distancia influye en la precisión de la misma, por lo que el análisis de este aspecto se debería abarcar en futuros estudios.

La distribución espacial del granizo puede ser una variable de entrada importante, en futuros trabajos se debería estudiar la forma de incorporar la misma a los modelos y

analizar su influencia. Esto incluye la posibilidad de estudiar el valor que toman las variables del radar no solo en el pixel correspondiente a la localización de verdad de campo, si no a aquellos píxeles circundantes a la misma.

Si bien en este trabajo el objetivo era utilizar datos de un solo sensor remoto, otra variable de interés es la temperatura provista por radio sondeos, para poder determinar la isoterma de cero grado que puede ayudar a mejorar las clasificaciones. Un estudio futuro de múltiples sensores sería conveniente.

Las variables Phi_{DP} y K_{DP} se deberían incluir en futuros modelos para analizar su aporte a los target buscados, una vez que se pueda programar su lectura y procesamiento.

Se destaca que la variable H_{DR} fue seleccionada en todos los modelos en los cuales se utilizó, por lo que también se debe incluir en futuros estudios; si se quiere generar un modelo que solo dependa de Z, se debe incluir también la variable E con diferentes configuraciones y métodos de cálculo.

Para que el modelo sea más útil es necesario encontrar una manera eficiente de almacenar y catalogar los datos disponibles del radar; el uso de bases de datos no estructuradas con productos como Mongo DB o Hadoop se presentan como una alternativa a estudiar. La velocidad de procesamiento también es una característica que se debe mejorar; como estos problemas se pueden paralelizar, es conveniente analizar el uso de metodologías como Map Reduce. Se debería trabajar en lograr una implementación completamente automática de los modelos obtenidos.

Finalmente, sería provechoso generar los modelos con otras técnicas de DM que se mencionan en los antecedentes, con los objetivos de buscar la más eficiente para este tipo de información ya sea por una mejora en la performance de los modelos o por el uso de menos cantidad de variables o información, permitiendo que la clasificación se procese más rápidamente.

5. Bibliografía

[1] G. A. Casagrande, G. T. Vergara, and Y. Bellini Saibene, "Cartas agroclimáticas actuales de temperaturas, heladas y lluvia de la provincia de La Pampa [Argentina]. Recent agroclimatic maps of temperature, frost and rainfall in La Pampa [Argentina].," *Revista de la Facultad de Agronomía*, vol. 17, Nov. 2006.

[2] E. Ponce de Leon, "Granizo." Servicio Meterologico Nacional, 1985.

[3] D. Ligier, "Documento Base del Programa Nacional: Ecorregiones." INTA, 2009.

[4] R. N. Mezher and P. A. Mercuri, "Análisis espacial y temporal de la ocurrencia de eventos de granizo sobre Argentina," *XV CONGRESSO BRASILEIRO DE METEROLOGIA*, agosto de 2008.

[5] R. N. Mezher, M. Doyle, and V. Barros, "Climatology of hail in Argentina," *Atmospheric Research*, vol. 114–115, pp. 70–82, Oct. 2012.

[6] R. N. Mezher, P. A. Mercuri, and N. N. Gattinoni, "Distribución espacio-temporal del granizo en Argentina," *Reunión Argentina de Agrometeorología. 12. 2008 10 08-10, 8 al 10 de octubre de 2008. San Salvador de Jujuy. AR.*, 2008.

[7] R. N. Mezher, N. N. Gattinoni, and P. A. Mercuri, "VARIABILIDAD ESTACIONAL DE LA OCURRENCIA DE GRANIZO EN EL CENTRO, ESTE Y NORESTE DE ARGENTINA," *Reunión Argentina de Agrometeorología. 12. 2008 10 08-10, 8 al 10 de octubre de 2008. San Salvador de Jujuy. AR.*, 2008.

[8] A. Rodríguez and H. Ciappesoni, "SIstema NAcional de Radares Meteorológicos. SINARAME.," presented at the Simposio de Radarización y Sistemas de Alerta Hidrometeorológicas del Cono Sur., Buenos Aires, Argentina, 2012.

[9] R. Hohl, H.-H. Schiesser, and I. Knepper, "The use of weather radars to estimate hail damage to automobiles: an exploratory study in Switzerland," *Atmospheric research*, vol. 61, no. 3, pp. 215–238, 2002.

[10] K. Casellas, G. Parellada, L. Longo, J. Portillo, P. Calonge, and E. Cristeche, "Memoria de La XXX Jornada de Perspectivas Agropecuarias 'Gestión del riesgo agropecuario: herramientas para mitigar y transferir el riesgo climático,'" Ciudad Autónoma de Buenos Aires, May 2012.

[11] C. Bustos and H. Videla, "Modelo estadístico de predicción de tormentas a corto plazo para la provincia de Mendoza," in *Anales del XI Congreso Argentino de Meteorología. Catuogno, GA*, 1982.

[12] J. L. Sánchez, L. López, E. García-Ortega, and B. Gil, "Nowcasting of kinetic energy of hail precipitation using radar," *Atmospheric Research*, vol. 123, pp. 48–60, 2013.

[13] P. Tabary, B. Fradon, A. J. Illingworth, and G. Vulpiani, "Hail detection and quantification with a C-band polarimetric radar: Challenges and promises," in *34th Conference on Radar Meteorology*, 2009, pp. 5–9.

[14] L. López and J. L. Sánchez, "Discriminant methods for radar detection of hail," *Atmospheric Research*, vol. 93, no. 1, pp. 358–368, 2009.

[15] L. López, E. García-Ortega, and J. L. Sánchez, "A short-term forecast model for hail," *Atmospheric Research*, vol. 83, no. 2–4, pp. 176–184, Feb. 2007.

[16] Wikipedia contributors, "Meteoro (meteorología)." Wikimedia Foundation, Inc., 09-Jun-2012.

[17] J. Billet, M. DeLisi, B. G. Smith, and C. Gates, "Use of Regression Techniques to Predict Hail Size and the Probability of Large Hail," *Weather and Forecasting*, vol. 12, no. 1, pp. 154–164, Mar. 1997.

[18] M. Alexiuk, P. C. Li, N. Pizzi, and W. Pedrycz, "Classification of Hail and Tornado Storm Cells Using Neural Networks," in *1999 IEEE Western Canada Conference and Exhibition*, pp. 15–21.
[19] P. C. Li, N. Pizzi, W. Pedrycz, D. Westmore, and R. Vivanco, "Severe storm cell classification using derived products optimized by genetic algorithms," in *Electrical and Computer Engineering, 2000 Canadian Conference on*, 2000, vol. 1, pp. 445–448.
[20] I. Holleman, *Hail detection using single-polarization radar*. Ministerie van Verkeer en Waterstaat, Koninklijk Nederlands Meteorologisch Instituut, 2001.
[21] C. Marzban and A. Witt, "A Bayesian neural network for severe-hail size prediction," *Weather and Forecasting*, vol. 16, no. 5, pp. 600–610, 2001.
[22] L. Ramirez, W. Pedrycz, and N. Pizzi, "Severe storm cell classification using support vector machines and radial basis function approaches," in *Electrical and Computer Engineering, 2001. Canadian Conference on*, 2001, vol. 1, pp. 87–91.
[23] D. J. Gagne, A. McGovern, and J. Brotzge, "Classification of convective areas using decision trees," *Journal of Atmospheric and Oceanic Technology*, vol. 26, no. 7, pp. 1341–1353, 2009.
[24] P. J. Visser and J. van Heerden, "Comparisons of hail kinetic energy derived from radar reflectivity with crop damage reports over the eastern Free State," *WATER SA-PRETORIA-*, vol. 26, no. 1, pp. 91–96, 2000.
[25] F. S. Marzano, D. Scaranari, and G. Vulpiani, "Supervised Fuzzy-Logic Classification of Hydrometeors Using C-Band Weather Radars," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 11, pp. 3784–3799, Nov. 2007.
[26] M. Ceperuelo Mallafré, "Identificación y caracterización del granizo mediante el radar meterologico. Modelos de predicción del ciclo de vida de las células convectivas.," Tesis Doctoral., 2008.
[27] R. C. Pérez, "Caracterización de las celdas graniceras que producen da\ no severo en la provincia de Mendoza utilizando mediciones del radar meteorológico de Banda C," 2006.
[28] S. C. Simonelli and M. Nicolini, "Modelo estadístico de Pronóstico de Convección para la zona norte de la Provincia de Mendoza," 2000.
[29] V. Makitov, "Radar measurements of integral parameters of hailstorms used on hail suppression projects," *Atmospheric Research*, vol. 83, no. 2–4, pp. 380–388, Feb. 2007.
[30] G. B. Foote, T. W. Krauss, and V. Makitov, "Hail metrics using conventional radar," in *Proc., 16th Conference on Planned and Inadvertent Weather Modification*, 2005.
[31] R. N. Mezher, S. Banchero, and Y. N. Bellini Saibene, "Identificación de granizo con la utilización de variables polarimétricas de los radares de Paraná y Anguil, el radar de Pergamino y daño en cultivos.," in *Congreso Argentino de Meteorología. 11. 2012 05-06 28-01, 28 de mayo al 1 de junio de 2012. Mendoza. AR.*, 2012.
[32] R. N. Mezher and P. A. Mercuri, "Uso de la red de radares de INTA para la detección de granizo," *XIII Reunión Argentina y VI Latinoamericana de Agrometeorología.*, Oct. 2010.
[33] R. N. Mezher, L. Vidal, and P. Salio, "Hailstorms Analysis using Polarimetric Weather Radars and Microwave Sensors in Argentina," *6th European Conference on Severe Storms (ECSS 2011)*, 26082011.
[34] "Instruction Manual. Rainbow 5. Products and Algorithms.," Germatronik, 2007.
[35] A. Waldvogel, B. Federer, and P. Grimm, "Criteria for the detection of hail cells," *Journal of Applied Meteorology*, vol. 18, no. 12, pp. 1521–1525, 1979.

[36] K. Skripnikova, D. Rezacova, and K. Skripnikova, "Testing radar-based hail detection criteria."
[37] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. The MIT press, 2001.
[38] J. M. Ale, "Data Mining y Cambio de Creencias: Aplicación de Lógicas Rebatibles al Problema de la Actualización de Reglas.," *VI Congreso Internacional en Innovación Tecnológica Informática, CIITI, Capítulo Buenos Aires.*, 2006.
[39] J. M. Gutiérrez, R. Cano, A. S. Cofi\ no, and C. M. Sordo, "Redes probabilísticas y neuronales aplicadas a las ciencias atmosféricas," *INM, Ministerio de Medio Ambiente, Madrid*, 2004.
[40] J. Bartok, O. Habala, P. Bednar, M. Gazak, and L. Hluchy, "Data mining and integration for predicting significant meteorological phenomena," *Procedia Computer Science*, vol. 1, no. 1, pp. 37–46, 2010.
[41] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Verlag, 2008.
[42] R. Nisbet, J. Elder, J. F. Elder, and G. Miner, *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.
[43] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.
[44] N. Almada, F. Díaz, L. Osorio, J. Blatter, H. Rodriguez, J. J. De Battista, N. Arias, and C. Bocchio, "Evaluación del daño por granizo en soja," *Boletín Técnico EEA Concepción del Uruguay.*, vol. 46, p. 3, 2005.
[45] R. W. Katz and R. R. Garcia, "Statistical relationships between hailfall and damage to wheat," *Agricultural Meteorology*, vol. 24, pp. 29–43, 1981.
[46] P. Simeonov, "An Overview of Crop Hail Damage and Evaluation of Hail Suppression Efficiency in Bulgaria," *J. Appl. Meteor.*, vol. 35, no. 9, pp. 1574–1581, Sep. 1996.
[47] H. H. Schiesser, "Hailfall: the relationship between radar measurements and crop damage," *Atmospheric Research*, vol. 25, no. 6, pp. 559–582, 1990.
[48] J. L. Sánchez, R. Fraile, J. L. De La Madrid, M. T. De La Fuente, P. Rodríguez, and A. Castro, "Crop damage: The hail size factor," *Journal of Applied Meteorology*, vol. 35, no. 9, pp. 1535–1541, 1996.
[49] "Resultados de la 'Encuesta Censal sobre los Seguros en el Sector Agropecuario y Forestal,'" Superintendencia de Seguros de la Nación, 3545, Apr. 2013.
[50] "Infografía: El granizo | EROSKI CONSUMER," *EROSKI CONSUMER*, 27-Jul-2012. [Online]. Available: http://www.consumer.es/web/es/medio_ambiente/naturaleza/2009/06/07/185799.php. [Accessed: 27-Jul-2012].
[51] L. E. Maderey Rascón, J. Roman, and others, *Principios de hidrogeografía. Estudio del ciclo hidrológico*. Unam, 2005.
[52] Wikipedia contributors, "Metaestabilidad." Wikimedia Foundation, Inc., 19-Jun-2012.
[53] "Thunderstorm: hail production -- Britannica Online Encyclopedia," *Encyclopedia Britannica*, 27-Jul-2012. [Online]. Available: http://www.britannica.com/EBchecked/media/161933/Hail-producing-thunderstorm-in-cross-section. [Accessed: 27-Jul-2012].
[54] H. E. Brooks, J. W. Lee, and J. P. Craven, "The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data," *Atmospheric Research*, vol. 67, pp. 73–94, 2003.

[55] R. E. Rinehart, *Radar for meteorologists*. Dept. of Atmospheric Sciences, Center for Aerospace Sciences, University of North Dakota, 1997.
[56] V. N. Bringi and V. Chandrasekar, *Polarimetric Doppler weather radar: Principles and applications*. Cambridge Univ Pr, 2001.
[57] J. F. Peters, Z. Suraj, S. Shan, S. Ramanna, W. Pedrycz, and N. Pizzi, "Classification of meteorological volumetric radar data using rough set methods," *Pattern Recognition Letters*, vol. 24, no. 6, pp. 911–920, Mar. 2003.
[58] *Fundamentos de radar meteorológico*. USA: University Corporation for Atmospheric Research (UCAR), 20120910.
[59] K. Aydin, T. A. Seliga, and V. Balaji, "Remote sensing of hail with a dual linear polarization radar," *Journal of Climate and Applied Meteorology*, vol. 25, no. 10, pp. 1475–1484, 1986.
[60] "Doppler Radar Frequently Asked Questions." National Weather Services - NOAA.
[61] T. Schuur, "Radar: Frequently Asked Questions - Summary." National Severe Storms Laboratory. Polarimetric Radar Research., 17-Feb-2003.
[62] A. Ryzhkov, D. Zrnic, J. Krause, M. Kumjian, and S. Ganson, "Small Hail Dry Hail? Wet Hail? Giant Hail," Jul. 2012.
[63] M. E. Anderson, L. D. Carey, W. A. Petersen, and K. R. Knupp, "C-band Dual-polarimetric Radar Signatures of Hail," Jul. 2012.
[64] P. L. Heinselman and A. V. Ryzhkov, "Validation of polarimetric hail detection," 2010.
[65] E. A. Brandes and A. V. Ryzhkov, "Hail detection with polarimetric radar," 2004, vol. 5.
[66] A. V. Ryzhkov, T. J. Schuur, D. W. Burgess, P. L. Heinselman, S. E. Giangrande, and D. S. Zrnic, "The joint polarization experiment. polarimetric Rainfall Measurement and Hidrometeor Classification," *Bull. Amer. Meteor. Soc*, vol. 86, pp. 809–824, 2005.
[67] J. Boettcher, "Dual Polarization RADAR Operation Course," *Dual Polarization RADAR Operation Course*. [Online]. Available: http://www.wdtb.noaa.gov/courses/dualpol/RDA/Lesson1/player.html.
[68] WW2010, "Radars Basics.Online Guides. Remote Sensing. Radars." University of Illinois, 2010.
[69] I. Holleman, H. R. Wessels, J. R. Onvlee, and S. J. Barlag, "Development of a hail-detection-product," *Physics and Chemistry of the Earth Part B Hydrology Oceans and Atmosphere*, vol. 25, pp. 1293–1297, 2000.
[70] L. Baldini, E. Gorgucci, V. Chandrasekar, and W. Peterson, "Implementation of CSU hydrometeor classification scheme for C-band polarimetric radars," 2005.
[71] R. Kaltenboeck and A. Ryzhkov, "Comparison of polarimetric signatures of hail at S and C bands for different hail sizes," *Atmospheric Research*, no. 0, Jul. 2012.
[72] S. Boodoo, D. Hudak, M. Leduc, A. V. Ryzhkov, N. Donaldson, and D. Hassan, "Hail detection with a C-band dual polarization radar in the canadian Great Lakes region.," *34th Conference on Radar Meteorology*, 2009.
[73] P. P. Alberoni, D. S. Zrnic, A. V. Ryzhkov, and L. Guerrieri, "Use of a fuzzy logic classification scheme with a C-band polarimetric radar: first results," 2002, vol. 324, p. 327.
[74] S. P. Williamson, *DOPPLER RADAR METEOROLOGICAL OBSERVATIONS. PART C WSR-88D PRODUCTS AND ALGORITHMS.* Washington, DC: OFFICE OF THE FEDERAL COORDINATOR FOR METEOROLOGICAL SERVICES AND SUPPORTING RESEARCH, 2006.

[75] P. Tabary, C. Berthet, P. Dupuy, J. Figueras i Ventura, B. Fradon, J. F. Georgis, R. Hogan, F. Kabeche, and J. P. Wasselin, "Hail detection and quantification with C-band polarimetric radars: results from a two-year objective comparison against hailpads in the south of France," in *Proceedings of 6th European Conference on Radar in Meteorology and Hydrology, Sibiu, Romania, 6-10 Sept*, 2010.
[76] F. S. Marzano, D. Scaranari, M. Montopoli, and G. Vulpiani, "Supervised classification and estimation of hydrometeors from C-band dual-polarized radars: A Bayesian approach," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 46, no. 1, pp. 85–98, 2008.
[77] D. Zrnic, "Dual Polarization for Weather Observations, Clasification of Hydrometeors and Measurement of Rain.," presented at the Simposio de Radarización y Sistema de Alertas Hidrometeológicos del Cono Sur, Buenos Aires, Argentina, 11-Sep-2012.
[78] J. J. Gourley, P. Tabary, and J. Parent du Chatelet, "A fuzzy logic algorithm for the separation of precipitating from nonprecipitating echoes using polarimetric radar observations," *Journal of Atmospheric and Oceanic Technology*, vol. 24, no. 8, pp. 1439–1451, 2007.
[79] H. Liu and V. Chandrasekar, "Classification of hydrometeors based on polarimetric radar measurements: Development of fuzzy logic and neuro-fuzzy systems, and in situ verification," *Journal of Atmospheric and Oceanic Technology*, vol. 17, no. 2, pp. 140–164, 2000.
[80] A. Ryzhkov, D. Zrnic, J. Krause, M. Kumjian, and S. Ganson, "Discrimination Between Large And Small Hail. Final Report." NSSL, 2010.
[81] P. T. May and T. D. Keenan, "Four-dimensional microphysical data from Darwin," in *13th ARM Science Team Meeting Proceedings*, 2003.
[82] T. Keenan, "Hydrometeor classification with a C-band polarimetric radar," *Australian Meteorological Magazine*, vol. 52, no. 1, pp. 23–31, 2003.
[83] S. A. Changnon Jr, "Hailfall characteristics related to crop damage," *Journal of Applied Meteorology*, vol. 10, no. 2, pp. 270–274, 1971.
[84] G. M. Morgan Jr and N. G. Towery, "CROP DAMAGE-HAILPAD PARAMETER STUDY IN ILLINOIS," *Illinois State Water Survey, Urbana*, vol. 101, p. 55, 1976.
[85] A. Waldvogel, W. Schmid, and B. Federer, "The Kinetic Energy of Hailfalls. Part I: Hailstone Spectra," *J. Appl. Meteor.*, vol. 17, no. 4, pp. 515–520, Abril 1978.
[86] H. Seino and others, "On the characteristics of hail size distribution related to crop damage," *J. Agric. Meteorol*, vol. 36, pp. 81–88, 1980.
[87] R. Hohl, H.-H. Schiesser, and D. Aller, "Hailfall: the relationship between radar-derived hail kinetic energy and hail damage to buildings," *Atmospheric Research*, vol. 63, no. 3, pp. 177–207, 2002.
[88] A. Witt, M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. . Mitchell, and K. W. Thomas, "An enhanced hail detection algorithm for the WSR-88D," *Weather and Forecasting*, vol. 13, no. 2, pp. 286–303, 1998.
[89] R. M. Hohl and M. Beniston, "Relationship between hailfall intensity and hail damage on ground, determined by radar and lightning observations," *Doctorial thesis. University Freiburg, Switzerland*, 2001.
[90] M. Alexiuk, N. Pizzi, and W. Pedrycz, "Classification of volumetric storm cell patterns," in *Electrical and Computer Engineering, 1999 IEEE Canadian Conference on*, 1999, vol. 2, pp. 1081–1085.

[91] V. Lakshmanan, T. Smith, G. Stumpf, and K. Hondl, "The warning decision support system-integrated information," *Weather and Forecasting*, vol. 22, no. 3, pp. 596–612, 2007.
[92] Z. Suraj, J. F. Peters, and W. Rzasa, "A Comparison of Different Decision Algorithms Used in Volumetric Storm Cells Classification," *Fundamenta Informaticae*, vol. 51, no. 1, pp. 201–214, Enero 2002.
[93] M. B. Richman, B. Santosa, and T. B. Trafalis, "Feature selection of radar-derived tornado attributes with support vector machines," in *4th Conference on Artificial Intelligence Applications to Environmental Sciences, San Diego, CA., Amer. Meteor. Soc*, 2005.
[94] V. Lakshmanan and T. Smith, "Data mining storm attributes from spatial grids," 2010.
[95] Z. Suraj and W. Rzasa, "Volumetric storm cell classification with the use of rough set methods," *Zeszyty Naukowe WSIiZ*, vol. 1, 2001.
[96] E. G. Tsagalidis, K. G. Tsitouridis, G. Evangelidis, and D. A. Dervos, "Hail Size Estimation and Prediction using Data Mining Techniques."
[97] E. Collino, P. Bonelli, and L. Gilli, "ST-AR (STorm-ARchive): A project developed to assess the ground effects of severe convective storms in the Po Valley," *Atmospheric Research*, vol. 93, no. 1–3, pp. 483–489, Jul. 2009.
[98] X. Li, R. Ramachandran, J. Rushing, S. Graves, K. Kelleher, S. Lakshmivarahan, D. Kennedy, and J. Levit, "Mining nexrad radar data: An investigative study," in *American Meteorology Society annual meeting*, 2004.
[99] X. Li, B. Plale, N. Vijayakumar, R. Ramachandran, S. Graves, and H. Conover, "Real-time storm detection and weather forecast activation through data mining and events processing," *Earth Science Informatics*, vol. 1, no. 2, pp. 49–57, 2008.
[100] Y. Bellini Saibene and G. Casagrande, "Radar Meteorológico en la EEA Anguil," *Horizonte Agropecuario*, La Pampa, Argentina, p. 13, Apr-2010.
[101] J.-P. Tuovinen, A.-J. Punkka, J. Rauhala, H. Hohti, and D. M. Schultz, "Climatology of Severe Hail in Finland: 1930–2006," *Mon. Wea. Rev.*, vol. 137, no. 7, pp. 2238–2249, Jul. 2009.
[102] D. Changnon and S. A. Changnon, "Surrogate data to estimate crop-hail loss," *Journal of applied meteorology*, vol. 36, no. 9, pp. 1202–1210, 1997.
[103] "Microsoft Excel," *Wikipedia, la enciclopedia libre*. 05-Dec-2014.
[104] "KML," *Wikipedia, la enciclopedia libre*. 07-Oct-2014.
[105] "Archivo de texto," *Wikipedia, la enciclopedia libre*. 17-Sep-2014.
[106] E. Saltikoff, J.-P. Tuovinen, J. Kotro, T. Kuitunen, and H. Hohti, "A Climatological Comparison of Radar and Ground Observations of Hail in Finland," *Journal of Applied Meteorology and Climatology*, vol. 49, no. 1, pp. 101–114, Jan. 2010.
[107] P. Bonelli, P. Marcacci, E. Bertolotti, E. Collino, and G. Stella, "Nowcasting and assessing thunderstorm risk on the Lombardy region (Italy)," *Atmospheric Research*, vol. 100, no. 4, pp. 503–510, Jun. 2011.
[108] R. Prieto, R. Herrera, P. Doussel, L. Gimeno, P. Ribera, R. García, and E. Hernández, "Looking for periodicities in the hail intensity in the Andes region," *ATMOSFERA*, vol. 14, no. 2, Oct. 2009.
[109] R. N. Mezher, V. Barros, and P. A. Mercuri, "Climatologia de eventos de granizo en la Region Pampeana." 2010.
[110] Yanina Bellini Saibene, L. Schaab, L. Ramos, M. L. Belmonte, and M. E. Fuentes, "Anuario RIAN-RIAP 2009-2010," *Boletín de Divulgación Tecnica EEA Anguil*, vol. 105, p. 54, Oct. 2011.

[111] A. Ferreyra, "Determinación de zona afectada por granizo en el partido de Chacabuco (Pcia. de Buenos Aires) el día 20 de noviembre de 2011, mediante imágenes satelitales.," INTA, Pergamino, Buenos Aires, Diciembre de 2011.
[112] K. L. Ortega, T. M. Smith, K. L. Manross, A. G. Kolodziej, K. A. Scharfenberg, A. Witt, and J. J. Gourley, "The Severe Hazards Analysis and Verification Experiment," *Bull. Amer. Meteor. Soc.*, vol. 90, no. 10, pp. 1519–1530, Oct. 2009.
[113] V. N. Bringi, J. Vivekanandan, and J. D. Tuttle, "Multiparameter Radar Measurements in Colorado Convective Storms. Part II: Hail Detection Studies," *J. Atmos. Sci.*, vol. 43, no. 22, pp. 2564–2577, Nov. 1986.
[114] "Transact-SQL," *Wikipedia, la enciclopedia libre*. 12-Oct-2014.
[115] "Gematronik Weather Radar Systems." Selex Systems Integration GmbH, 2010.
[116] T. HARTMANN, M. S. AMBURRINO, and F. BAREILLES, "Análisis preliminar de datos obtenidos por la red de radares del INTA para el estudio de precipitaciones en la Región Pampeana," *39 Congreso Argentino de Agroinformatica. 2. Jornadas Argentinas de Informática. JAIIO*, Sep. 2010.
[117] "Instruction Manual. Rainbow 5. File Format.," in *Instruction Manual. Rainbow 5.*, Gematronik, 2007.
[118] Santiago Banchero, *Procesamiento Datos Radar Meterológico Polarimétrico*. Buenos Aires, Argentina: INTA, 2010.
[119] "ASCII," *Wikipedia, la enciclopedia libre*. 25-Nov-2014.
[120] "GeoTIFF," *Wikipedia, la enciclopedia libre*. 17-Mar-2014.
[121] "Tagged Image File Format," *Wikipedia, the free encyclopedia*. 10-Aug-2014.
[122] ESRI, *ArcGIS 9.2 Desktop Help*. 2007.
[123] "WGS84," *Wikipedia, la enciclopedia libre*. 04-Dec-2014.
[124] doxygen, "Gdal_rasterize," *Gdal*, 16-Aug-2014. [Online]. Available: http://www.gdal.org/gdal_rasterize.html.
[125] C. Ferreira, "Gene expression programming: a new adaptive algorithm for solving problems," *arXiv preprint cs/0102027*, 2001.
[126] X. Li, C. Zhou, W. Xiao, and P. C. Nelson, "Prefix gene expression programming," in *Proc. Genetic and Evolutionary Computation Conference, Washington*, 2005, pp. 25–31.
[127] Z. Xie, X. Li, B. Di Eugenio, P. C. Nelson, W. Xiao, and T. M. Tirpak, "Using gene expression programming to construct sentence ranking functions for text summarization," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004, p. 1381.
[128] C. Ferreira, "Designing neural networks using gene expression programming," in *Applied Soft Computing Technologies: The Challenge of Complexity*, Springer, 2006, pp. 517–535.
[129] C. Ferreira, *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence (Studies in Computational Intelligence)*. Springer-Verlag New York, Inc., Secaucus, NJ, 2006.
[130] C. Ferreira, "Getting Started with Classification," *GeneXproTools Tutorials – A Gepsoft Web Resource.*, 10-Sep-2013. [Online]. Available: http://www.gepsoft.com/tutorials/GettingStartedWithClassification.htm. [Accessed: 01-Dec-2015].
[131] C. Ferreira, "Getting Started with Logistic Regression.," *GeneXproTools Tutorials – A Gepsoft Web Resource.*, 10-Nov-2013. [Online]. Available: http://www.gepsoft.com/tutorials/GettingStartedWithLogisticRegression.htm. [Accessed: 01-Dec-2015].

[132] D. W. Hosmer and S. Lemeshow, *Applied logistic regression*, Second., vol. 354. Wiley-Interscience, 2000.
[133] A. Ng, "Machine Learning: Logistic Regression.," presented at the Curso de Machine Learning, Coursera, 30-Mar-2014.
[134] N. / N. S. W. P. Center, "Space Weather Prediction Center - Forecast Verification Glossary." [Online]. Available: http://www.swpc.noaa.gov/forecast_verification/Glossary.html#G. [Accessed: 16-Nov-2013].
[135] G. León Aristizábal, "Verificación de los modelos Meterológicos." Instituto de Hidrologia, Meteorologia y Estudios Ambientales - IDEAM, Mar-2005.
[136] C. Ferreyra, "GeneXproTools Help. Modeling made easy.," *Gene X proo Tools Knowledge Base*. [Online]. Available: http://www.gepsoft.com/gxpt4kb/default.htm. [Accessed: 08-Dec-2014].
[137] A. Martelli, *Python in a Nutshell*. O'Reilly Media, Inc., 2006.
[138] D. Kolarič, "Hail detection methods using radar data."

Anexo 1. Glosario de acrónimos y siglas

ASCII: American Standard Code for Information Interchange. Código Estándar Estadounidense para el Intercambio de Información
ASP.NET: Active Server Page. Servidor de páginas activo. EL sufijo .NET hace referencia al nombre comercial dado por Microsoft a este producto.
AUC: Area Under the Curve. Area bajo la curva.
CAPE: energía convectiva potencial disponible.
CAPPI: Constant Altitude PPI. Indicador de posición en un plano de altitud constante.
CC: coeficiente de correlación
CRISP-DM: Cross Industry Standard Process for Data Mining.
CSI: Critical Success Index. También llamado TS.
dBZ: decibelios.
DER: Diagrama de Entidad Relación.
DM: Data Mining. Minería de Datos.
DP: doble polarización
E: energía cinética.
EEA: Estación Experimental Agropecuaria.
ER: Error rate
ETOP: Echo Top. Eco superior.
FAR: False Alarm Ratio. Taza de falsas alarmas.
GDAL: Geospatial Data Abstraction Library
GeoTIFF: Geo Tagged Image File Format.
GEP: Gene Expression Programming
GHz: Gigahertz (unidad de medida de frecuencia).
GIS: Geographic Information System. Sistema de Información Geográfica
GPS: Global Positioning System. Sistema de posicionamiento global.
GS: Gilbert Skill Score
HDA: Hail Detection Algorithm. Algoritmo de detección de granizo.
H_{DR}: Hail Differential Reflectivity
HSS: Heidke Skill Score
IDE: Integrated Development Enviroment
INTA: Instituto Nacional de Tecnología Agropecuaria.
KDD: Knowledge Discovery in Databases. Descubrimiento del conocimiento en base de datos.
K_{DP}: fase específica diferencial. Specific Differential Phase.
KML: Keyhole Markup Language.
L_{DR}: Linear Depolarization Ratio.
MHz: Megahertz (unidad de medida de frecuencia).
MODIS: Moderate-Resolution Imaging Spectroradiometer. Espectroradiómetro de imágenes de media resolución
NDVI: Índice de Vegetación de Diferencia Normalizada o Indice Verde.
.NET: Microsoft dot NET Framework
NPV: Negative Predictive Value
PC: Percent Correct o Accuracy

Phi_{DP}: Fase de propagación diferencial. Differential Propagation Phase.
POD: Probability Of Detection. También conocido como Sensibilidad, Recall, Exhaustividad
POH: Probabilit of Hail. Probabilidad de Granizo.
PPI: Plan Position Indicator. Indicador de posición en un plano.
PPV: Positive Predictive Value. También conocido como Precision, SR (Success Ratio)
QGIS: Quantum Geographic Information System
RADAR: Radio Detection and Ranking.
RHI: Range Height Indicator. Indicador de muestreo vertical.
Rho_{HV}: Coeficiente de correlación copolar. Correlation Coefficient.
RIAN: Red de Información Agropecuaria Nacional.
SD: Texture Parameter.
SHI: Severe Hail Index (Indice de granizo severo)
SIIA: Sistema Integrado de Información Agropecuaria
SIG: Sistema de Información Geográfico
SMN: Servicio Meteorológico Nacional
SP: Simple polarización
SQL: Structured Query Language. Lenguaje estructurado de consulta.
SR: Success Ratio, También conocido como Precision y PPV.
TS: Threat Score. También llamado CSI.
TSQL: Transac-SQL
TIFF: Tagged Image File Format
TXT: TeXTo. Extensión de archivos de texto plano.
VIL: Vertically Integrated Liquid. Líquido integrado en la vertical.
W(Z): función de peso de Z.
WBZ: Wet Bulb Zero
WGS84: World Geodetic System 84.
XLS: Extensión de archivos Microsoft Excel en versiones anteriores o iguales a Excel 2003.
XML: Extensible Markup Languaje.
Z: Reflectividad.
Z_{DR}: Reflectividad diferencial. Differential Reflectivity.
Zh: Reflectividad horizontal.
ZHAIL: Z-based Hail Warning. Alarma de granizo basada en Z (reflectividad).
Zmax: maxima reflectividad.
Zmin: minima reflectividad.
Zv: Reflectividad vertical.
ΦDP: Phi_{DP}. Fase de propagación diferencial. Differential Propagation Phase.
ρHV: Rho_{HV}. Coeficiente de correlación copolar. Correlation Coefficient

Anexo 2. Sistema de Información y Base de Datos

Relevamiento de los datos a campo

Se diseñó y compiló una base de datos que se realizó en SQL Server 2008 Express R2 con extensión para el tratamiento de datos geográficos. La base de datos quedó compuesta por nueve tablas que se pueden agrupar en:

a) Datos de tormentas: la tabla Eventos contiene la fecha y horarios de los eventos de granizo ocurridos en el período bajo estudio. La tabla TipoEvento detalla si la tormenta fue de granizo, viento, lluvia o sus combinaciones.
b) Datos del radar: las tablas datosZDR, datosRhoHV y datosdBZ contienen la información de las variables del radar para una localización, en cada toma de datos (atributo "horario") y cada elevación (atributo elevación).
c) Datos de medios de comunicación: la tabla Reportes contiene el detalle del artículo desde el cual se sacó información de ocurrencia de granizo. La tabla Medios contiene la información para identificar la fuente de cada artículo y la tabla TipoMedio indica si la fuente es un periódico local, nacional, una radio, etc.
d) Datos de localizaciones: la tabla LocalizacionXEvento contiene los datos de cada punto de campo georeferenciado, con la información de ocurrencia de granizo, daño ocasionado, cultivo presente, fuente, entre otros datos.

La figura 42 presenta el diagrama de entidad relación (DER) de la base de datos. Esta base de datos se administra por medio de un sistema de información desarrollado con la plataforma web ASP.NET, utilizando Visual Basic .NET como lenguaje de programación. Se puede acceder al mismo desde http://rian.inta.gob.ar/DanioGranizo. La figura 43 muestra la pantalla del sistema web donde se mapean las localizaciones de verdad de campo.

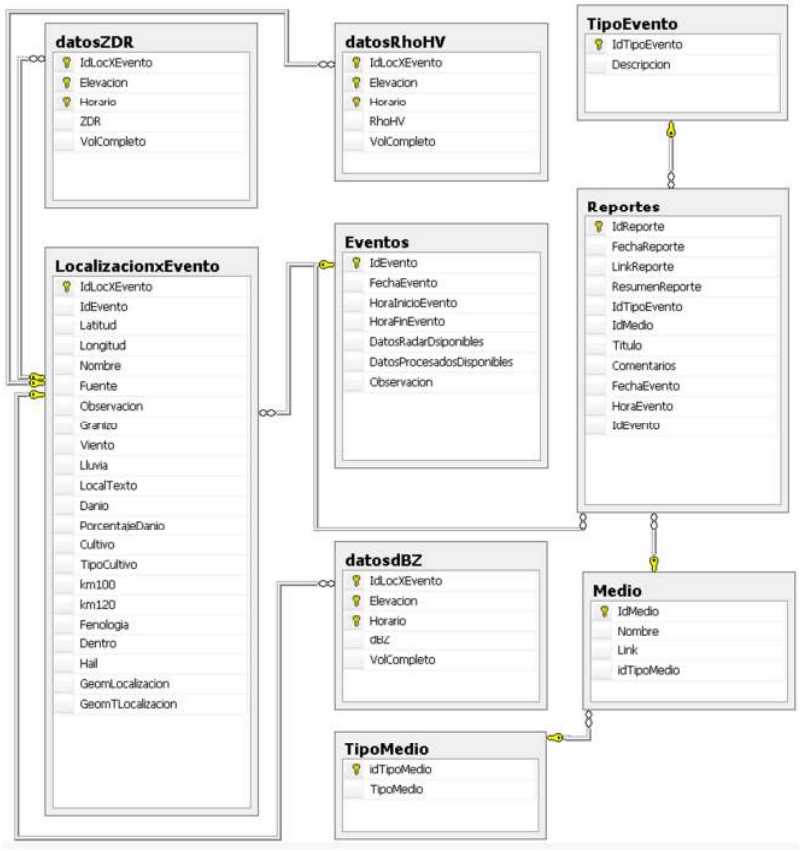


Figura 42. Diagrama de Entidad Relación.

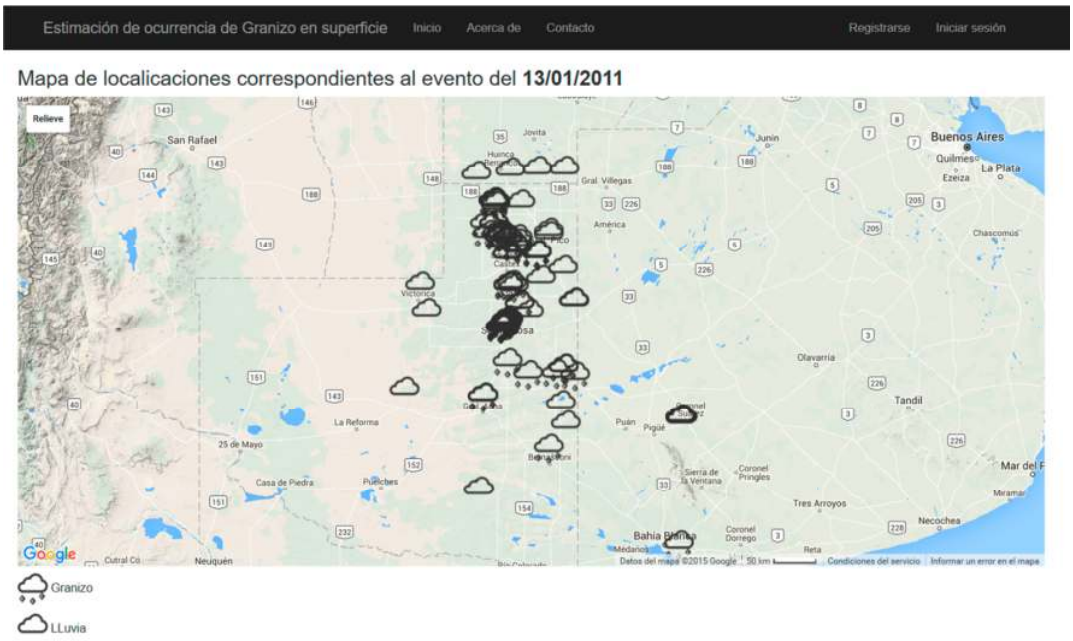


Figura 43 Ejemplo de visualización en un mapa de las localizaciones relevadas a campo.

Implementación

Se diseñaron tres bases de datos (BD) para la prueba de implementación. En la figura 44 se detallan el esquema de tablas de cada una el DER. Las bases de datos ImplementacionRadarGranizo e ImplementacionRadarGranizo2 tienen el mismo formato y las mismas tablas (figura 44.b). La BD ImplementacionRadarGranizo2 solo se utiliza si la BD ImplementacionRadarGranizo se llena con los datos a procesar. Estas bases de datos también tienen las vistas que hacen los resúmenes de 24 horas de las variables Z, Z_{DR} y Rho_{HV}.

La BD Implementación tiene las tablas que almacenan las variables de resumen que luego son variables de entrada para los modelos (figura 44.c) y la vista que genera el DataSet como lo necesita el modelo para ser aplicado a los datos.

Figura 44. Diagrama de Entidad Relación de las bases de datos de la prueba de implementación.

Anexo 3. Detalle de modelos GEP target Daño

Target con 5 clases

Tabla 39. Función logística para la probabilidad de granizo positivo del Modelo Leve y detalles de su construcción.

Función logística			
$P(Y=1) = \frac{1}{1 + \exp(-y)}$			
Armado de *y*			
y = (gepLT2G(log(pow(gepLT2G(d[TotZDR],d[MnZDR]),4.0)),(gepLT2B((G1C2*G1C2),d[MnRho])*(d[MxdBZ]-d[MnZDR])))-d[MndBZ]) y = y + pow((gepLT2B((((d[AvZDR]-d[TotRho])+gepOR2(d[AvRho],G2C3))/2.0),(G2C7*d[MxZDR]))*(gepOR1(G2C6,d[MnZDR])*pow(G2C2,4.0))),4.0) y = y + (gepLT2A((G3C2*d[AvdBZ]),(((1.0-d[MxRho])*d[TotdBZ])*(d[MxdBZ]-d[TotdBZ])))-d[MnRho]) y = y + pow(gepGOE2G(gepLT2G(gepLT2A(gepLT2B(d[TotdBZ],d[TotZDR]),((d[MxdBZ]+G4C4)/2.0)),gepOR2(pow(d[TotZDR],3.0),d[MndBZ])),gep3Rt(gep3Rt(d[MxZDR]))),4.0) # Model standardization and shift y = (y - 4.26482181902775E+15) / 1.64285525803045E+15 – [(4.89701893582712E+15 - 4.26482181902775E+15) / 1.64285525803045E+15]			
Funciones lógicas para el armado de *y*			
gep3Rt	**gepOR1**	**gepOR2**	**gepLT2A**
Recibe (x): if (x < 0.0): return -pow(-x,(1.0/3.0)), else: return pow(x,(1.0/3.0))	Recibe (x, y): if ((x < 0.0) or (y < 0.0)): return 1.0, else: return 0.0	Recibe (x, y):if ((x >= 0.0) or (y >= 0.0)):return 1.0, else: return 0.0	Recibe (x, y): if (x < y): return x, else: return y
gepLT2B	**gepLT2G**	**gepGOE2G**	
Recibe (x, y):if (x < y): return 1.0, else: return 0.0	Recibe (x, y):if (x < y): return (x+y), else: return atan(x*y)	Recibe (x, y):if (x >= y):return (x+y), else: return atan(x*y)	

Tabla 40. Función logística para la probabilidad de granizo positivo del Modelo Moderado y detalles de su construcción.

<table>
<tr><td colspan="4">Función logística</td></tr>
<tr><td colspan="4">$P(Y=1) = \frac{1}{1 + \exp(-y)}$</td></tr>
<tr><td colspan="4">Armado de y</td></tr>
<tr><td colspan="4">y = gepGOE2A(pow((log(d[Fenologia])*(G1C7/d[MnZDR])),2.0),(G1C9+gepGOE2A(((G1C5+d[MnZDR])/2.0),gepAND2(G1C9,d[AvdBZ]))))

y = y + gepGOE2B(gepGOE2G(d[MxRho],gepLT2B(gepGOE2A(d[Fenologia],d[AvEWt]),exp(d[AvEWt]))),gepLT2C(((d[MnZDR]+d[Cultivos])*d[tipoCultivo]),(d[TotEWt]*G2C7)))

y = y + (G3C6*exp(((exp(d[Fenologia])+(d[Cultivos]+d[Cultivos]))-(gepGOE2G(d[MxZDR],G3C1)*gepGOE2A(G3C3,d[AvdBZ])))))

y = y + gep3Rt(gepLT2G(gepGOE2G(((d[TotdBZ]*d[Fenologia])/gepLT2G(d[AvZDR],G4C4)),pow(d[MxEWt],2.0)),(pow(G4C5,4.0)*(G4C7+G4C7))))

Model standardization and shift
y = (y - -5.41092664054064E+48) / 4.26044538324689E+49 – [(31.1777647177883 - -5.41092664054064E+48) / 4.26044538324689E+49]</td></tr>
<tr><td colspan="4">Funciones lógicas para el armado de y</td></tr>
<tr><td>gep3Rt</td><td>gepAND2</td><td>gepLT2B</td><td>gepLT2C</td></tr>
<tr><td>Recibe (x):
if (x < 0.0):
return -pow(-x,(1.0/3.0)), else:
return pow(x,(1.0/3.0))</td><td>Recibe (x, y): if ((x >= 0.0) and (y >= 0.0)): return 1.0, else: return 0.0</td><td>Recibe (x, y):
if (x < y): return 1.0, else: return 0.0</td><td>Recibe (x, y):
if (x < y): return (x+y), else: return (x-y)</td></tr>
<tr><td>gepLT2G</td><td>gepGOE2A</td><td>gepGOE2B</td><td>gepGOE2G</td></tr>
<tr><td>Recibe (x, y):
if (x < y): return (x+y), else: return atan(x*y)</td><td>Recibe def (x, y): if (x >= y): return x, else: return y</td><td>Recibe (x, y):
if (x >= y): return 1.0, else: return 0.0</td><td>Recibe (x, y):
if (x >= y):
return (x+y), else:
return atan(x*y)</td></tr>
</table>

Tabla 41. Función logística para la probabilidad de granizo positivo del Modelo Severo y detalles de su construcción.

<table>
<tr><th colspan="4">Función logística</th></tr>
<tr><td colspan="4">$P(Y=1) = \frac{1}{1 + \exp(-y)}$</td></tr>
<tr><th colspan="4">Armado de y</th></tr>
<tr><td colspan="4">y = (((d[AvZDR1]-(d[Cultivos]*d[MnEWt]))*gepOR2((d[AvdBZ1]-G1C2),G1C5))-gepLT2B((d[MxRho1]+d[TotdBZ1]),d[TotEWt]))

y = y + (gepGOE2E(gepLogi(pow(d[MnZDR1],4.0)),gepLogi((d[TotRho1]-d[AvRho1])))+sqrt(gepLT2B(gepLT2G(G2C0,d[MnZDR1]),d[Cultivos])))

y = y + gep3Rt(d[AvEWt])

y = y + gepLT2G(((d[MxRho1]*d[TotZDR1])*gepLT2A(pow(d[AvdBZ1],2.0),G4C1)),((log(G4C4)+gepGOE2E(d[MnZDR1],d[AvZDR1]))/2.0))

Model standardization and shift
y = (y - -15.927857625588) / 40.9724435273808 – [(5.73717757079161 - -15.927857625588) / 40.9724435273808]</td></tr>
<tr><th colspan="4">Funciones lógicas para el armado de y</th></tr>
<tr><th>gep3Rt</th><th>gepOR2</th><th>gepLT2A</th><th>gepLT2B</th></tr>
<tr><td>Recibe (x):
if (x < 0.0):
return -pow(-x,(1.0/3.0)), else:
return pow(x,(1.0/3.0))</td><td>Recibe (x, y):
if ((x >= 0.0) or (y >= 0.0)): return 1.0, else: return 0.0</td><td>Recibe (x, y):
if (x < y): return x,
else: return y</td><td>Recibe (x, y):
if (x < y): return 1.0,
else: return 0.0</td></tr>
<tr><th>gepLT2G</th><th>gepGOE2E</th><th>gepLogi</th><th></th></tr>
<tr><td>Recibe (x, y):
if (x < y): return (x+y), else: return atan(x*y)</td><td>Recibe (x, y):
if (x >= y): return (x+y), else: return (x*y)</td><td>Recibe (x):
if (abs(x) > 709.0):
return 1.0 / (1.0 + exp(abs(x) / x * 709.0)), else:
return 1.0 / (1.0 + exp(-x))</td><td></td></tr>
</table>

Tabla 42. Función logística para la probabilidad de granizo positivo del Modelo Grave y detalles de su construcción.

Función logística
$P(Y=1) = \frac{1}{1 + \exp(-y)}$
Armado de *y*
y = gepGOE2E(exp(d[MxRho]),((((G1C6+G1C6)/2.0)+G1C4)*(gepLT2B(d[MxdBZ],d[TotZDR])+gepGOE2E(d[MnZDR],d[TotRho]))))
y = y + gepGOE2E((((gepLT2C(((d[MxdBZ]+G2C7)/2.0),d[AvRho])*pow(d[MnZDR],4.0))+gepGOE2A((1.0-d[AvdBZ]),d[MnZDR]))/2.0),d[MxZDR])
y = y + d[AvdBZ]
y = y + (G4C6*(d[AvZDR]*gepGOE2E((d[AvRho]+(d[TotZDR]+G4C1)),(gepAND1(G4C1,d[MxdBZ])-d[TotdBZ]))))
Model standardization and shift y = (y - -46776170.5132249 E) / 355746155.98962 – [(27597.1053189181 - 46776170.5132249) / 355746155.98962]

Funciones lógicas para el armado de *y*			
gepAND1	**gepLT2B**	**gepLT2C**	**gepGOE2A**
Recibe (x, y): if ((x < 0.0) and (y < 0.0)): return 1.0, else: return 0.0	Recibe (x, y): if (x < y): return 1.0, else: return 0.0	Recibidos (x, y): if (x < y): return (x+y), else: return (x-y)	Recibidos (x, y): if (x >= y): return x, else: return y.
gepGOE2E			
Recibe (x, y): if (x >= y): return (x+y), else: return (x*y)			

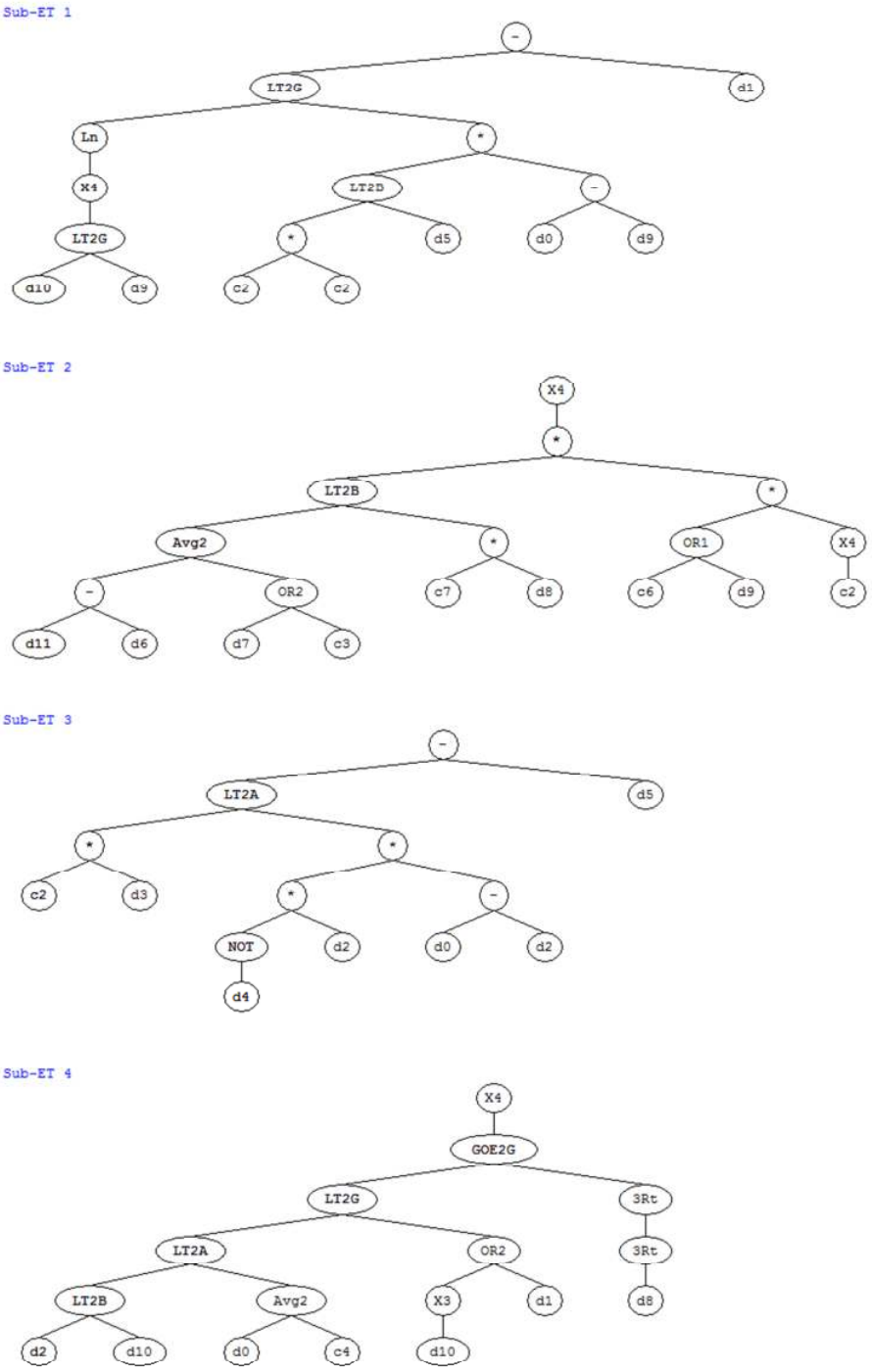


Figura 45. Árbol de expresión para la clase "Leve". Variables: MxdBZ =d0, MndBZ = d1, TotdBZ = d2, AvdBZ = d3, MxRho = d4, MnRho = d5, TotRho = d6, AvRho = d7, MxZDR =d8, MnZDR = d9, TotZDR = d10, AvZDR = d11. Constantes: G1C2 = -0.210586870937223, G2C7 = -4.55305187566759, G2C6 = 4.3667619861446, G2C2 = -9.56358531449324, G2C3 = 3.51664784691916, G3C2 = 8.61670966312089, G4C4 = -9.6226383251442.

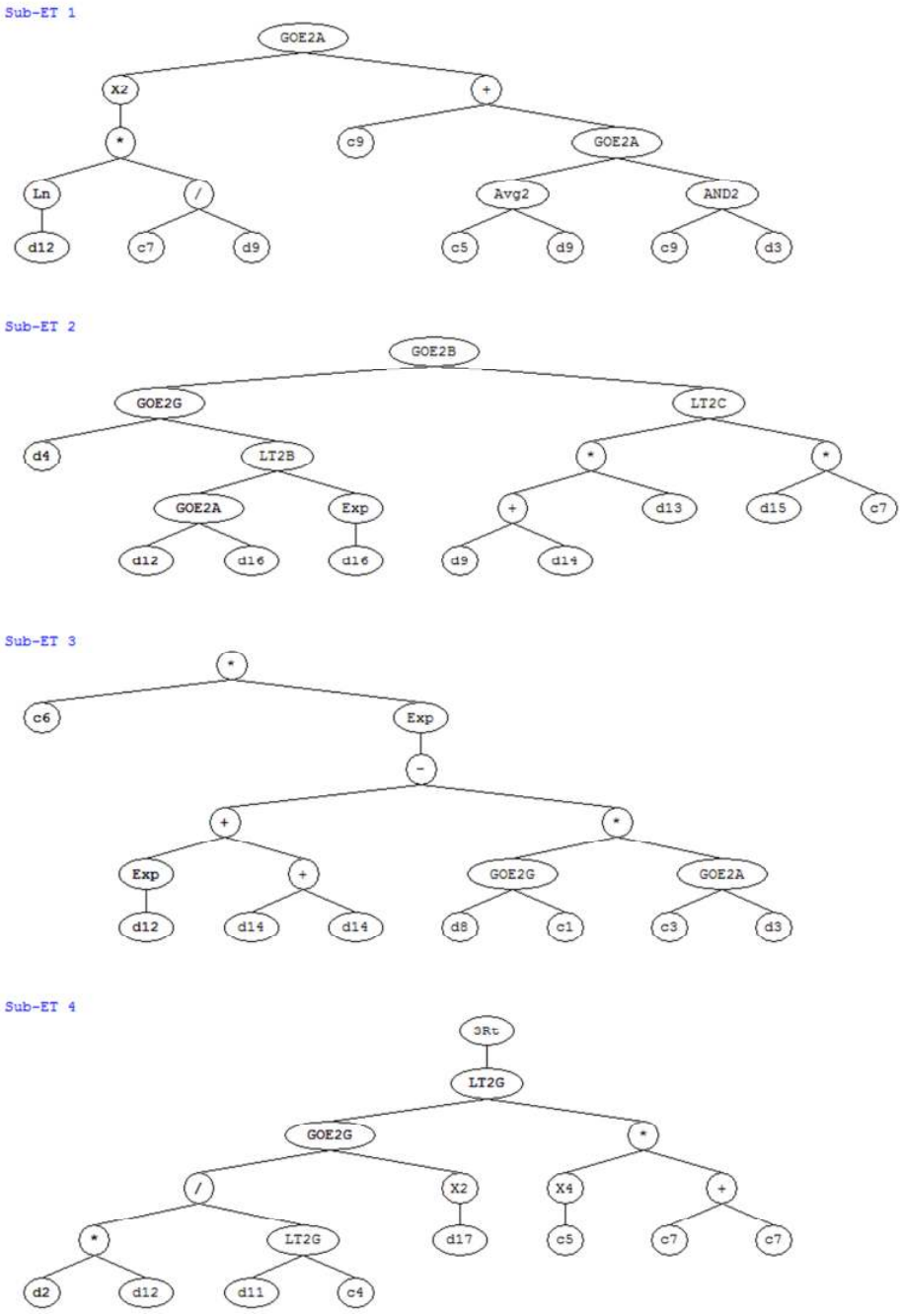


Figura 46. Árbol de expresión para la clase "Moderado". Variables: TotdBZ = d2, AvdBZ = d3, MxRho = d4, MxZDR = d8, MnZDR = d9, AvZDR = d11, Fenologia = d12, tipoCultivo = d13, Cultivos = d14, TotEWt = d15, AvEWt = d16, MxEWt = d17, Constantes: G1C9 = 6.31703247198675, G1C7 = 5.58267459781778, G1C5 = 5.98754749110385, G2C7 = 4.62935270241401, G3C6 = -2.22083193456832, G3C1 = 1.84240241706595, G3C3 = -5.47471541489914, G4C5 = -6.30174664454085, G4C7 = 4.30652855159154, G4C4 = 0.100848588352388

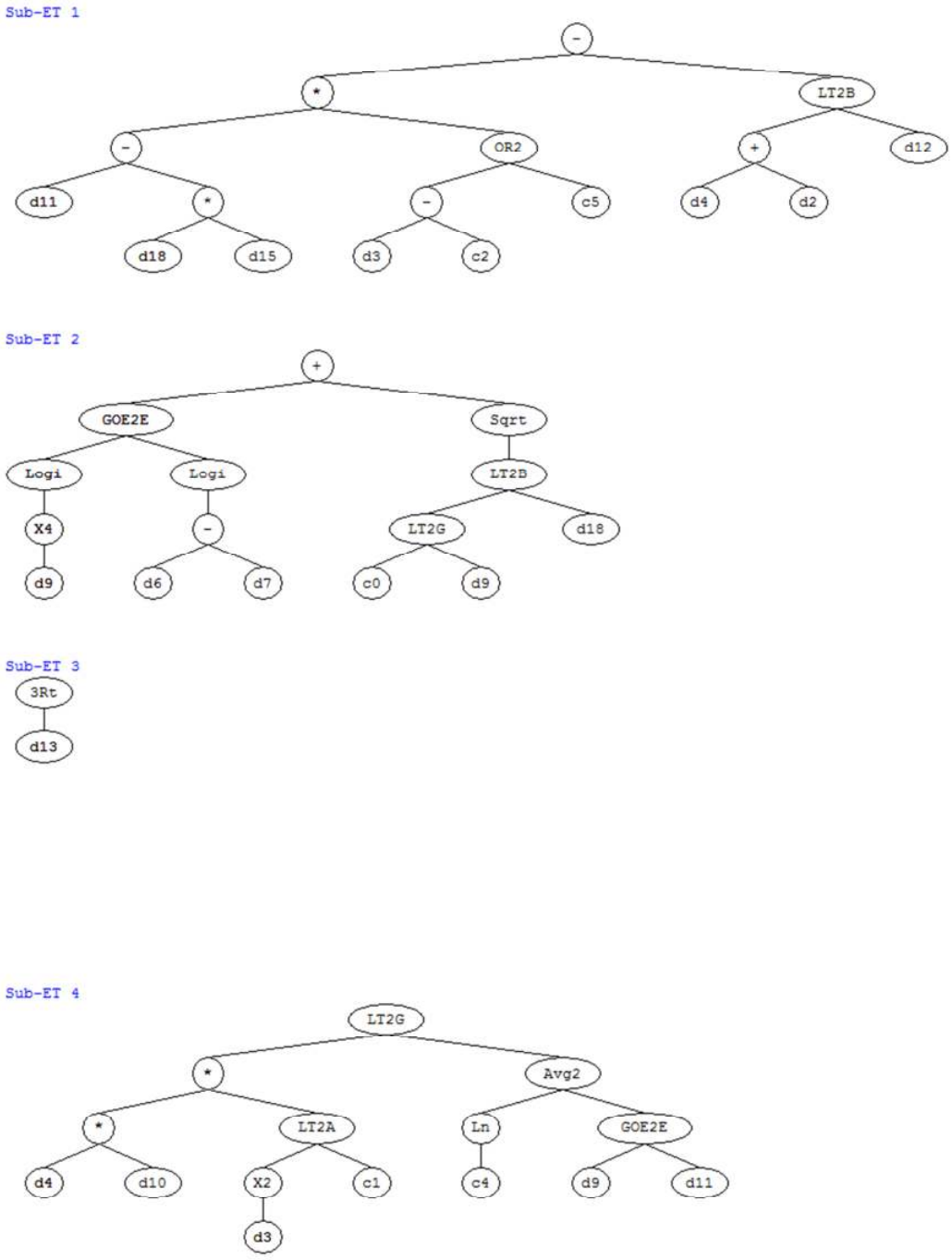


Figura 47. Árbol de expresión para la clase "Severo". Variables: TotdBZ1 = d2,AvdBZ1 = d3, MxRho1 = d4, TotRho1 = d6, AvRho1 = d7, MnZDR1 = d9, TotZDR1 = d10, AvZDR1 = d11, TotEWt = d12, AvEWt = d13, MnEWt = d15. Cultivos = d18. Constantes:G1C5 = -5.42344431897946, G1C2 = 1.00985747856075, G2C0 = -4.00250251777703, G4C1 = 5.19760734885708, G4C4 = 8.8982818079165.

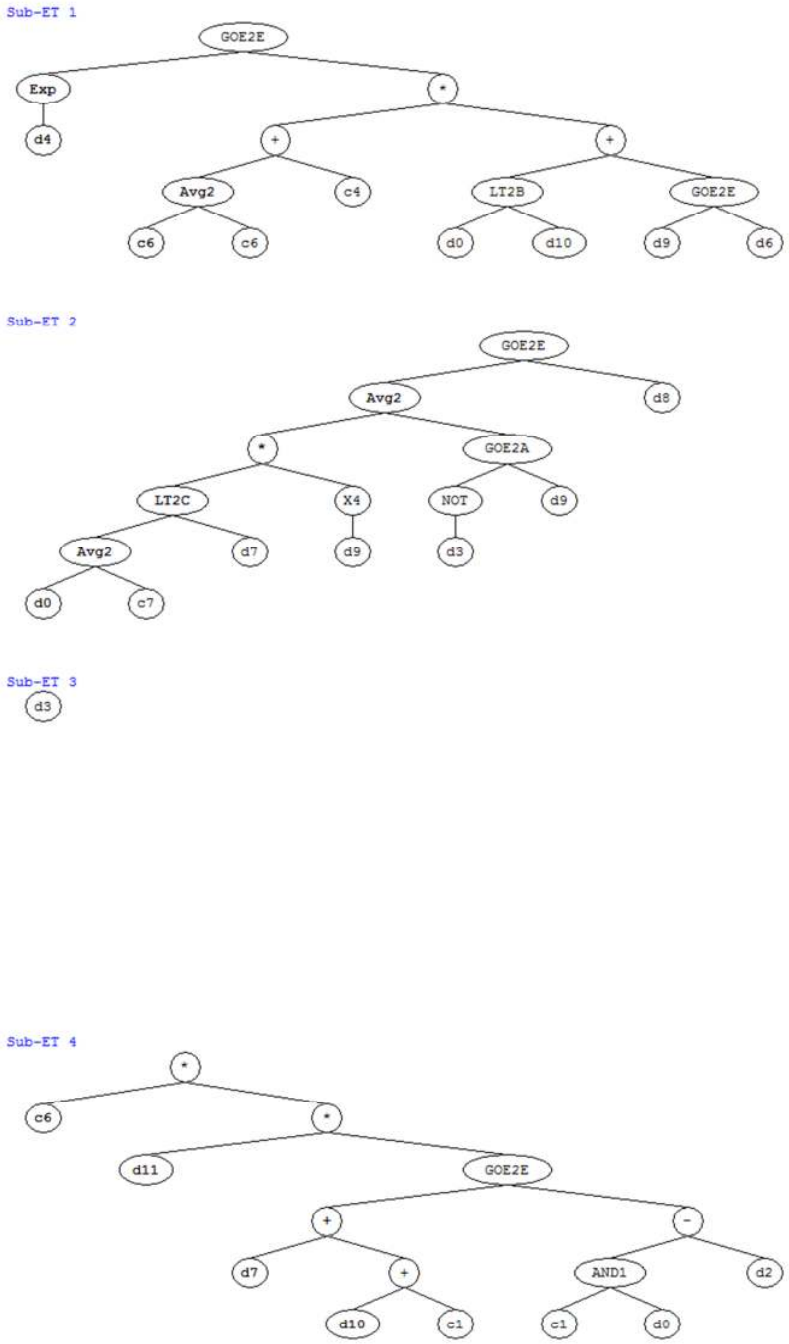


Figura 48. Árbol de expresión para la clase "Grave" (se presenta el modelo que menos variables utiliza). Variables: MxdBZ = d0, TotdBZ = d2, AvdBZ = d3, MxRho = d4, TotRho = d6, AvRho = d7, MxZDR = d8, MnZDR = d9, TotZDR = d10, AvZDR = d11. Constantes: G1C4 = 8.38312936796167, G1C6 = 1.70493789483322, G2C7 = -3.88936124759667, G4C6 = -7.61889400921659, G4C1 = -7.89422284615619.

Target con 3 clases

Tabla 43. Función logística para la probabilidad positivo de la clase "Sin Daño" y detalles de su construcción.

<table>
<tr><th colspan="4">Función logística</th></tr>
<tr><td colspan="4">$P(Y=1) = \frac{1}{1 + \exp(0.682238413415047*y + 4.30616235143977)}$</td></tr>
<tr><th colspan="4">Armado de y</th></tr>
<tr><td colspan="4">y = sqrt(gepLogi(pow(gepGOE2E(gepLT2A(d[MxRho],((gepAND1(G1C7,d[MnRho])+d[MnZDR])/2.0)),atan(gepAND1(G1C8,G1C1))),3.0)))

y = y + (pow(atan(d[AvdBZ]),2.0)*d[MnRho])

y = y + pow(gepAND2(((atan(atan(d[TotZDR]))+gep3Rt((d[MnZDR]*d[TotRho])))/2.0),(gepOR2(d[AvdBZ],d[MnRho])+pow(d[TotZDR],4.0))),4.0)

y = y + gepGOE2G(gepGOE2G(gepOR1(gepGOE2C(G4C1,G4C1),d[MnRho]),gepGOE2A((G4C4*d[AvdBZ]),d[AvdBZ])),((G4C9+(1.0-d[MxHDR]))/2.0))</td></tr>
<tr><th colspan="4">Funciones lógicas para el armado de y</th></tr>
<tr><th>gep3Rt</th><th>gepOR1</th><th>gepOR2</th><th>gepAND1</th></tr>
<tr><td>Recibe x:
if (x < 0.0):
return -pow(-x,(1.0/3.0)), else:
return pow(x,(1.0/3.0))</td><td>Recibe(x, y):
if ((x < 0.0) or (y < 0.0)): return 1.0,
else: return 0.0</td><td>Recibe(x, y):
if ((x >= 0.0) or (y >= 0.0)): return 1.0,
else: return 0.0</td><td>Recibe (x, y):
if ((x < 0.0) and (y < 0.0)):
return 1.0
else:
return 0.0</td></tr>
<tr><th>gepAND2</th><th>gepLT2A</th><th>gepGOE2A</th><th>gepGOE2C</th></tr>
<tr><td>Recibe (x, y):
if ((x >= 0.0) and (y >= 0.0)):
return 1.0
else:
return 0.0</td><td>Recibe (x, y):
if (x < y):
return x
else:
return y</td><td>Recibe (x, y):
if (x >= y):
return x
else:
return y</td><td>Recibe (x, y):
if (x >= y):
return (x+y)
else:
return (x-y)</td></tr>
<tr><th>gepGOE2E</th><th>gepGOE2G</th><th>gepLogi</th><th></th></tr>
<tr><td>Recibe (x, y):
if (x >= y):
return (x+y)
else:
return (x*y)</td><td>Recibe (x, y):
if (x >= y):
return (x+y)
else:
return atan(x*y)</td><td>Recibe (x):
if (abs(x) > 709.0):
return 1.0 / (1.0 + exp(abs(x) / x * 709.0))
else:
return 1.0 / (1.0 + exp(-x))</td><td></td></tr>
</table>

Tabla 44. Función logística para la probabilidad positivo de la clase "Menos de 50%" y detalles de su construcción.

<table>
<tr><th colspan="4">Función logística</th></tr>
<tr><td colspan="4">$P(Y=1) = \frac{1}{1 + \exp(-(2.69480869158104E\text{-}03 * y + -4.93561195211844))}$</td></tr>
<tr><th colspan="4">Armado de y</th></tr>
<tr><td colspan="4">y = (gepAND2(gepOR1(exp(d[AvHDR]),pow(G1C6,3.0)),(1.0-gepLT2E(d[MndBZ],G1C4)))*gepAND2((d[AvdBZ]-d[AvdBZ]),gepGOE2G(d[MndBZ],G1C9)))

y = y + gepLT2E(gepLT2E(atan(gepOR1(d[MnHDR],G2C2)),gep3Rt((d[AvdBZ]*G2C8))),(1.0-gepGOE2G((d[AvHDR]*G2C5),(G2C1*d[TotZDR]))))

y = y + (gepLogi(gepLogi(exp(G3C8)))*gepLT2A(gepLogi((G3C7-d[MndBZ])),((G3C7*G3C8)*(d[MndBZ]*d[MnHDR]))))

y = y + pow(gepLT2G(G4C8,d[MxdBZ]),2.0)</td></tr>
<tr><th colspan="4">Funciones lógicas para el armado de y</th></tr>
<tr><th>gep3Rt</th><th>gepOR1</th><th>gepAND2</th><th>gepLT2A</th></tr>
<tr><td>Recibe (x):
if (x < 0.0): return -pow(-x,(1.0/3.0)),
else: return pow(x,(1.0/3.0))</td><td>Recibe (x, y):
if ((x < 0.0) or (y < 0.0)): return 1.0, else: return 0.0</td><td>Recibe (x, y):
if ((x >= 0.0) and (y >= 0.0)): return 1.0,
else: return 0.0</td><td>Recibe (x, y):
if (x < y): return x,
else: return y</td></tr>
<tr><th>gepLT2E</th><th>gepLT2G</th><th>gepGOE2G</th><th>gepLogi</th></tr>
<tr><td>Recibe (x, y):
if (x < y): return (x+y), else: return (x*y)</td><td>Recibe (x, y):
if (x < y): return (x+y), else: return atan(x*y)</td><td>Recibe (x, y): if (x >= y): return (x+y),
else: return atan(x*y)</td><td>Recibe (x):
if (abs(x) > 709.0):return 1.0 / (1.0 + exp(abs(x) / x * 709.0)),else: return 1.0 / (1.0 + exp(-x))</td></tr>
</table>

Tabla 45. Función logística para la probabilidad positivo de la clase "Más de 50%" y detalles de su construcción.

<table>
<tr><th colspan="4">Función logística</th></tr>
<tr><td colspan="4">$P(Y=1) = \frac{1}{1 + \exp(-(3.84883671107244E-08 * y + -7.00046478207421))}$</td></tr>
<tr><th colspan="4">Armado de y</th></tr>
<tr><td colspan="4">y = (((d[TotdBZ]+exp(d[MnRho]))/2.0)*((gepAND1(d[TotdBZ],G1C8)*gepGOE2E(d[MxHDR],d[MxZDR]))+d[MnHDR]))

y = y + gepLT2A(gepLT2B(d[AvdBZ],gepLT2E(gepLogi(gepLT2C(d[TotZDR],G2C9)),(G2C9-G2C0))),atan(G2C8))

y = y + pow((((exp(d[MnRho])*((d[MxHDR]+d[MxdBZ])/2.0))-d[AvdBZ])-((d[MndBZ]+(d[AvdBZ]*G3C4))/2.0)),4.0)

y = y + (((gepLogi(gepAND2(exp(G4C6),pow(d[AvRho],4.0)))+(((d[MndBZ]+d[AvZDR])/2.0)-(d[MnHDR]*d[MxdBZ])))/2.0)*d[AvdBZ])</td></tr>
<tr><th colspan="4">Funciones lógicas para el armado de y</th></tr>
<tr><th>gepAND1</th><th>gepAND2</th><th>gepLT2A</th><th>gepLT2B</th></tr>
<tr><td>Recibe (x, y):
if ((x < 0.0) and (y < 0.0)): return 1.0,
else: return 0.0</td><td>Recibe (x, y):
if ((x >= 0.0) and (y >= 0.0)): return 1.0,
else: return 0.0</td><td>Recibe (x, y):
if (x < y): return x,
else: return y</td><td>Recibe (x, y):
if (x < y): return 1.0,
else: return 0.0</td></tr>
<tr><th>gepLT2C</th><th>gepLT2E</th><th>gepGOE2E</th><th>gepLogi</th></tr>
<tr><td>Recibe (x, y):
if (x < y):
return (x+y)
else:
return (x-y)</td><td>Recibe (x, y):
if (x < y):
return (x+y)
else:
return (x*y)</td><td>Recibe (x, y):
if (x >= y):
return (x+y)
else:
return (x*y)</td><td>Recibe (x):
if (abs(x) > 709.0):
return 1.0 / (1.0 + exp(abs(x) / x * 709.0))
else:
return 1.0 / (1.0 + exp(-x))</td></tr>
</table>

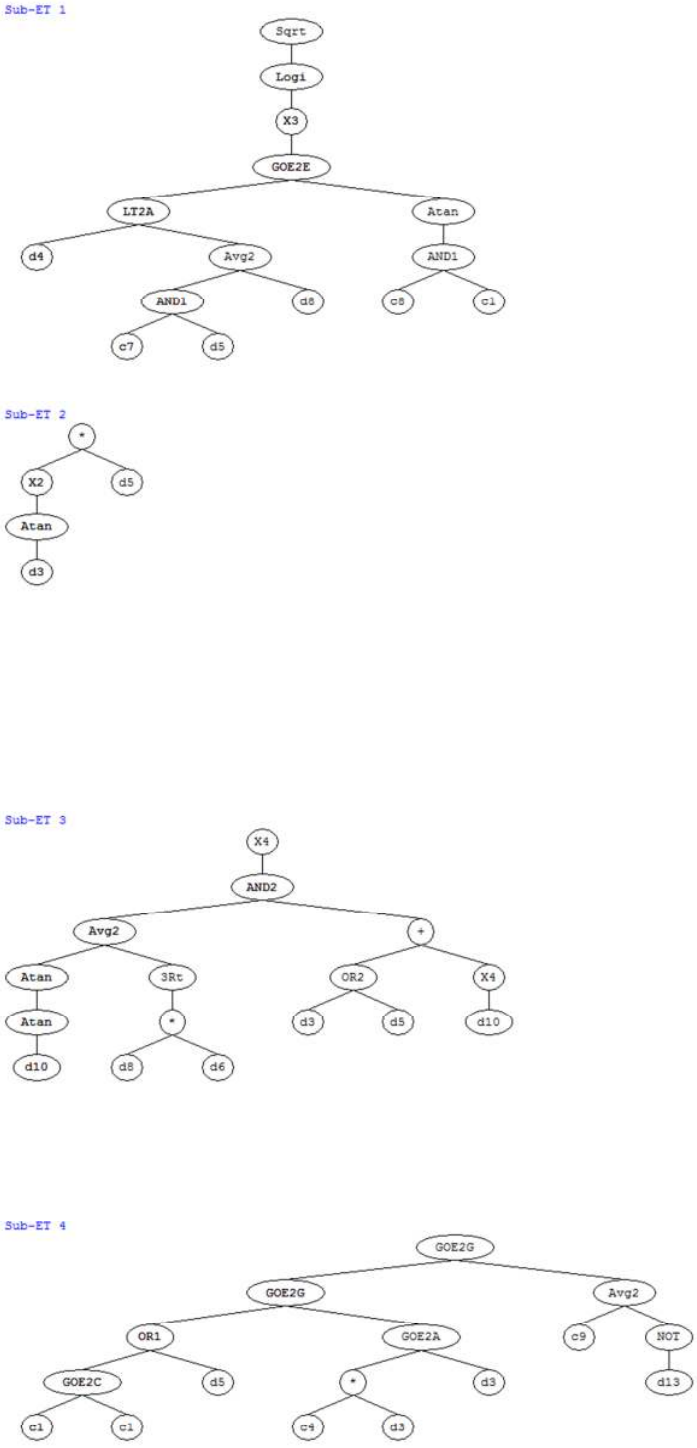


Figura 49. Árbol de expresión para la clase "Sin Daño". Variables: AvdBZ = d3, MxRho = d4, MnRho = d5, TotRho = d6, MnZDR = d8, TotZDR = d10, MxHDR = d13, Constantes: G1C8 = -5.04562517166662, G1C1 = 0.33051545762505, G1C7 = -5.06332590716269, G4C9 = 4.93575853755303, G4C1 = -9.5501571703238, G4C4 = -7.68059327982421.

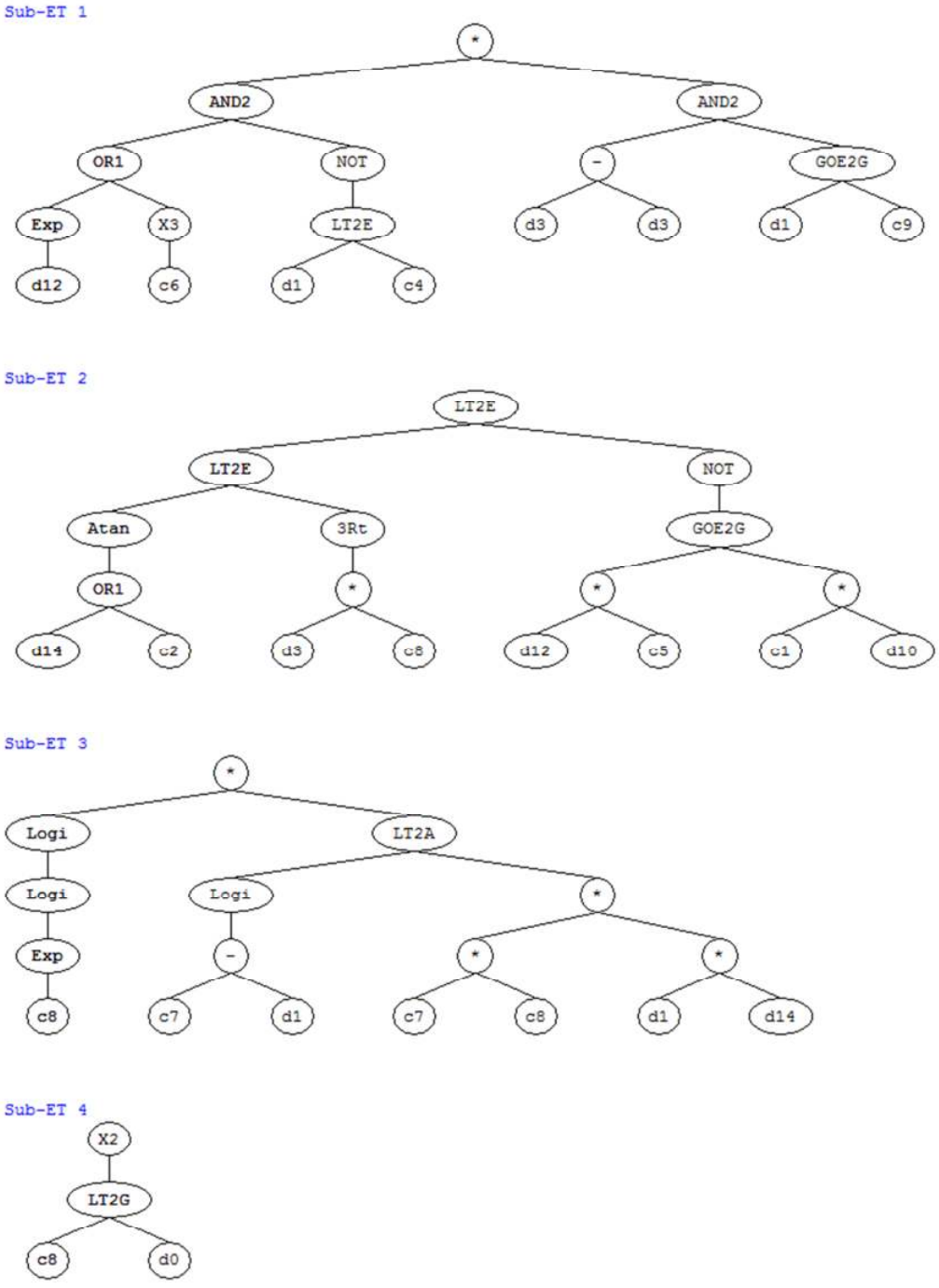


Figura 50. Árbol de expresión para la clase "Menos 50%". Variables: MxdBZ = d0, MndBZ = d1, AvdBZ = d3, TotZDR = d10, AvHDR = d12, MnHDR = d14. Constantes: G1C9 = -2.05908383434553, G1C6 = -3.34879604480117, G1C4 = -3.62407300027467, G2C2 = -0.135196996978668, G2C8 = -0.356761375774407, G2C5 = 7.6281014435255, G2C1 = 5.92883083590197, G3C8 = 0.924405652027955, G3C7 = 2.61452070680868, G4C8 = -6.3365581225013.

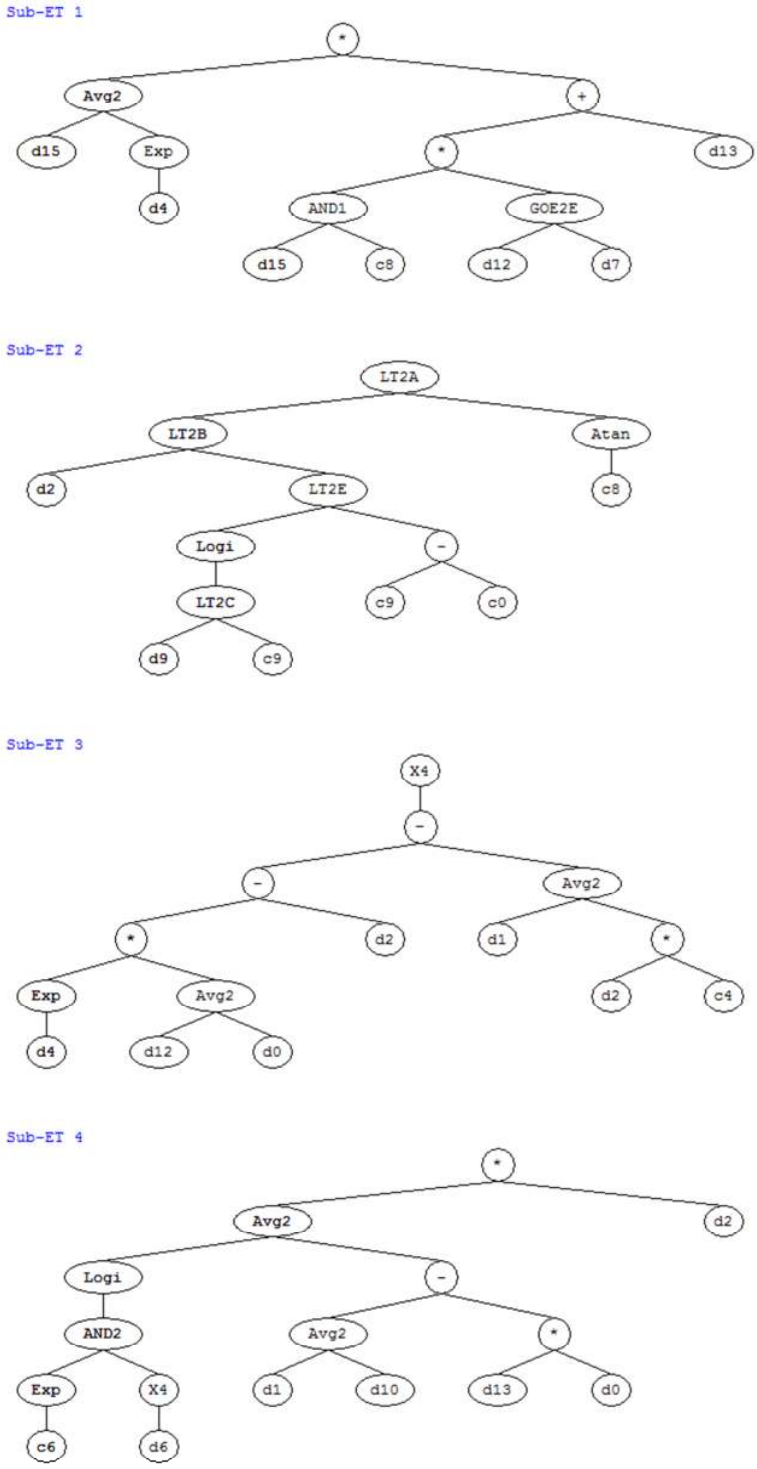


Figura 51. Árbol de expresión para la clase "Menos 50%". Variables: MxdBZ = d0,MndBZ = 1, AvdBZ = d2,MnRho = d4, AvRho = d6,MxZDR = d7,TotZDR = d9, AvZDR = d10, MxHDR = d12, MnHDR = d13, TotdBZ = d15. Constantes: G1C8 = -6.88711203344829, G2C8 = -1.01255836664937, G2C9 = 5.52171391949217, G2C0 = 5.27695547349467, G3C4 = -5.55345316934721, G4C6 = -4.31440168462172

Anexo 4. Productos obtenidos en el marco de la tesis

Publicaciones.

1. **Clasificación de granizo en superficie usando técnicas de minería de datos y datos de radar meteorológico.** Bellini Saibene, Yanina; Volpacchio, Martín. 16º Argentine Symposium on Artificial Intelligence. ISSN: 1850-2784. Argentina 31 de Agosto de 2015. Rosario, Santa Fe, Argentina. Trabajo completo.
2. **Desarrollo de software para el procesamiento y análisis de datos de la red de radares meteorológicos del INTA.** Bellini Saibene, Yanina; Banchero, Santiago; Mezher, Romina; Volpacchio, Martín. CongreMet XII. Congreso Argentino de Meteorología. 26 al 29 de mayo de 2015. Mar del Plata, Argentina. Poster.
3. **HAR (Hail-ARchive): desarrollo de un sistema de información y base de datos sobre granizo en la región semiárida pampeana central.** Bellini Saibene, Yanina; Caldera, Juan Marcelo; Volpacchio, Martín. CongreMet XII. Congreso Argentino de Meteorología. 26 al 29 de mayo de 2015. Mar del Plata, Argentina. Poster.
4. **Red de radares de INTA.** Bellini Saibene, Yanina; Belmonte María Laura. Horizonte Agropecuario Pampeano-Puntano Nº 103. Pág. 07. Octubre 2014. ISSN 0327-3180. Artículo de Divulgación.
5. **Desarrollo y uso de herramientas libres para la explotación de datos de los radares meteorológicos del INTA.** Bellini Saibene, Yanina; Volpacchio, Martín., Banchero, Santiago., Mezher, Romina. Anales del 6º Congreso Argentino de AgroInformática. ISSN: 1852-4850. 1 al 5 de Septiembre de 2014. Buenos Aires, Argentina. Trabajo completo.
6. **Radar Meteorológico INTA.** Bellini Yanina, Casagrande Guillermo. Revista 365. Pág. 18 y 19. Abril del 2012. Artículo de Divulgación.
7. **Identificación de granizo con la utilización de variables polarimétricas de los radares de Paraná y Anguil, el radar de Pergamino y daño en cultivos.** Romina Nahir Mezher, Santiago Banchero y Yanina Bellini. CongreMet XI. XI Congreso Argentino de Meteorología. 28 de Mayo al 1 de Junio de 2012. Mendoza, Argentina. Poster.
8. **Minería de datos: el que busca, encuentra.** Bellini Saibene, Yanina. Horizonte Agropecuario Pampeano-Puntano Nº 90. Pág. 06. Julio 2011. ISSN 0327-3180. Artículo de Divulgación.
9. **"RADAR Meteorológico en la EEA Anguil"**. Lic. Yanina Bellini Saibene, Ing. Agr. Guillermo Casagrande. Horizonte Agropecuario Nº 85. Abril 2010. ISSN: 0327-3180. Artículo de Divulgación.

Convenios.

1. **Convenio de Cooperación Técnica entre el INTA y La Facultad de Ingeniería de la Universidad Nacional de La Pampa.** Firmado el 16 Septiembre de 2015. General Pico, La Pampa. *Objetivo:* desarrollo e implementación de herramientas para el almacenamiento, acceso libre y procesamiento de los datos generados por el radar meteorológico situado en la EEA Anguil. La cooperación surge del abordaje multidisciplinario que requiere

la problemática, específicamente del aporte agronómico, agrometeorológico y de explotación de datos por parte del INTA y del aporte de LA FACULTAD en lo relacionado con sistemas de información, procesamiento de flujos de datos, bases de datos distribuidas y no estructuradas. Comité coordinador: *Lic. Yanina Bellini Saibene (por INTA)* y Dr. Mario Diván (por la Facultad).

2. **Convenio de comisión de estudios entre la Facultad De Ingeniería de la Universidad Nacional De La Pampa y el INTA.** Firmado el 16 Septiembre de 2015. General Pico, La Pampa. Objetivo: establecer un "Programa de Prácticas Educativas" destinado a estudiantes avanzados de la facultad, en instalaciones de INTA y con temas propuestos por INTA, para la realización por parte de los alumnos de grado y postgrado, de prácticas relacionadas con la educación y formación, de acuerdo a la especialización que reciben. *Responsable por INTA: Lic. Yanina Bellini Saibene.*

Premios.

1. **Premio Nacional de Gobierno Electrónico. Categoría Proyecto.** Septiembre de 2015. Rosario, Santa Fe. Otorgado por SADIO. Premio obtenido por el trabajo: *"Hacia una Arquitectura de Procesamiento de Datos del RADAR Meteorológico de INTA Anguil"*. Mario Diván, Yanina Bellini Saibene, María de los Ángeles Martín, María Laura Belmonte, Guillermo Lafuente y Juan Marcelo Caldera. *Resultado del trabajo conjunto en el convenio INTA – Fac.Ing. UNLPam. El software generado en esta tesis, para el procesamiento de los datos del radar y los modelos obtenidos sobre granizo, son parte de esta arquitectura.*

2. **Mención en el 1er Hackatón de AgroDatos.** Octubre de 2014. Buenos Aires. Organizado por: Fundación Sadosky – Maestría en Minería de Datos de la UBA. Desarrollo del proyecto *"YBYTU: Identificación de eventos climáticos con impacto agropecuario en noticias locales":* se generaron mapas para registrar granizadas utilizando como fuente de datos el relevamiento automatizado de medios periodísticos locales con técnicas de análisis de textos. *Se utilizaron los datos relevados en esta tesis para la generación del modelo.* Código del proyecto: https://github.com/DrDub/yvytu. Dr. Pablo Duboe, Lic. Sist. Yanina Bellini Saibene.