Pontificia Universidad Católica de Valparaíso

Facultad de Ingeniería

Escuela de Ingeniería Informática

DESCRIPCIÓN AUTOMÁTICA DE NUBOSIDAD MEDIANTE APRENDIZAJE DE MÁQUINAS

JOEL ESTEBAN PINTO GAMBOA

Tesis para optar al grado de

Magister en Ingeniería Informática

Pontificia Universidad Católica de Valparaíso

Facultad de Ingeniería

Escuela de Ingeniería Informática

DESCRIPCIÓN AUTOMÁTICA DE NUBOSIDAD MEDIANTE APRENDIZAJE DE MÁQUINAS

JOEL ESTEBAN PINTO GAMBOA

Profesor Guía: **Rodrigo Alfaro Arancibia**

Tesis para optar al grado de Magister en Ingeniería Informática

Junio 2014

Dedico esta Tesis a
mi familia y amigos.
En forma especial
agradezco a mi madre
por su constante
apoyo en esta etapa.

Resumen

La descripción de la nubosidad es una de las observaciones de mayor demanda en estudios climatológicos y mientras es deseable un alto grado de calidad en dicha descripción, es muy difícil detectar errores pues se trata de un fenómeno complejo y su observación es realizada por técnicos especializados. En este trabajo se propone implementar un clasificador automático bajo el enfoque *machine learning* sobre imágenes obtenidas con instrumentos de observación del cielo (TSI). Se describe la implementación de un clasificador multi-etiqueta y se presentan los resultados obtenidos al utilizarlo sobre una colección de datos de la Dirección Meteorológica de Chile. El desempeño obtenido demuestra que la clasificación automática es útil para detectar un buen porcentaje de los errores en las observaciones.

Palabras clave: aprendizaje automático, reconocimiento de patrones, Meteorología

Abstract

Cloud observations are one of the most demanded data in climate study and while high data quality is desirable, it is very difficult to assure because clouds are a complex phenomenon, the common process to obtain this data is error prone and it is very difficult to recognize errors in such observations. This work propose the implementation of an automatic cloud image data classification using the machine learning framework over digital images generated by sky observation instruments (TSI). This report describes the implementation of a multi-label classificator and presents the performance results obtained in the classification of a cloud image data-set from the data-bank in the National Weather Agency of Chile. The result shows that the automatic classification is very useful in the detection of errors and in meteorological observations quality assurance.

Keywords: machine learning, pattern recognition, meteorology

Índice

Listado de Figuras

Listado de Tablas

1 Introducción

La Dirección Meteorológica de Chile (DMC) dependiente de la Dirección General de Aeronáutica Civil, es el organismo responsable de los estudios meteorológicos y climatológicos en el país, y cuyo propósito es satisfacer las necesidades de información y previsión meteorológica de todas las actividades nacionales.

Parte del quehacer de esta organización es la mantención del banco nacional de datos meteorológicos y climatológicos el cual es alimentado con observaciones meteorológicas realizadas en una red de estaciones de monitoreo a lo largo de todo el país, además, incorpora los datos aportados por otras organizaciones en un marco de cooperación.

Unas de las tareas más importantes respecto a la mantención del banco de datos nacional es el control de calidad de la información, esto consiste en una serie de procesos y análisis sobre los datos observados para detectar inconsistencias y posibles errores de observación. En la actualidad este análisis se realiza por un especialista que revisa la totalidad de mediciones en un conjunto de observaciones (rango temporal), proceso que es apoyado por un repositorio de conocimiento meteorológico en la forma de reglas de control.

1.1 Calidad en las Observaciones Meteorológicas

Aun cuando se están incorporando un gran número de instrumentos de monitoreo automático para la observación de ciertos elementos, de la gran cantidad de mediciones que se realizan en estas observaciones (de 60 a 250 mediciones en un instante de tiempo) gran parte corresponde a la observación directa de un fenómeno meteorológico complejo que resulta difícil de cualificar sólo a través de instrumentos. Esta observación es realizada por un observador meteorológico, especialista en la observación y registro de estas mediciones que observa de forma directa (a ojo desnudo) o de forma indirecta (a través de algún instrumento para la observación) el fenómeno.

Entre los datos observados, la descripción de la nubosidad es uno de los fenómenos complejos de mayor demanda en la meteorología y consiste en una descripción cualitativa de la nubosidad presente en el cielo visible, aquí un observador inspecciona la bóveda celeste y describe el tipo de nube (cúmulos, cirros, estratos, etc.). Las nubes conforman un fenómeno meteorológico mayor pues se relacionan con el ciclo hídrico, el balance energético en escala global y local a través de la interacción con la radiación solar y terrestre. Esta medición se complementa con otras observaciones cuantitativas de nubosidad como la cobertura nubosa, la capa isobárica donde se encuentra, etc., sin embargo, el tipo de nube aporta mucho más información que cualquier indicador cuantitativo de la misma.

1.2 Problema

El problema con las observaciones directas como la descripción de la nubosidad, es que resulta muy difícil identificar errores que no radican en la codificación del fenómeno sino en la génesis de la observación, es decir, el fenómeno fue mal observado o simplemente no hubo observación y por lo tanto el dato registrado no guarda ninguna relación con la realidad. En el caso de la DMC, el control de calidad actual se limita al hallazgo de errores que se evidencian gracias a la inconsistencia de datos a lo largo de un conjunto de mediciones y el incumplimiento de reglas predefinidas en las relaciones de estos elementos. En lo que respecta a la descripción de la nubosidad, es muy difícil validar esta observación puesto que son pocas las reglas inter-elemento meteorológico que pueden detectar errores en este nivel de detalle.

1.3 Discusión Bibliográfica y Estado del Arte

Diferentes enfoques han sido propuestos para la identificación y clasificación automática de la nubosidad, algunos se sustentan en la utilización de los datos generados por uno o más instrumentos altamente especializados y no convencionales, otros se basan en la explotación de fuentes de datos más accesibles (en algunos casos sólo gracias a programas de cooperación internacional) como imágenes satelitales o imágenes de cielo completo tomadas desde tierra con instrumentos de observación como las *Whole Sky Cameras* o el *Total Sky Imager* de la Figura 1.1.

Figura 1.1: *Total Sky Imager*, utilizado para la observación remota de la nubosidad

El primer enfoque y el más tradicional generalmente implica la construcción de un instrumento altamente especializado cuyos datos se deben complementar con otras fuentes para producir predicciones. Esta clasificación suele ser de muy buena calidad pero el costo asociado a la adquisición, implementación y mantención a gran escala de dicha tecnología alejan el interés de muchas agencias de monitoreo y estudio meteorológico.

El segundo enfoque busca explotar fuentes de datos que comúnmente se encuentran disponibles en la mayoría de los países gracias a programas de cooperación internacional como lo son la fotografía satelital (AVHRR: espectro visible, infrarrojo y otros) y las fotografías tomadas desde tierra como las imágenes de cielo completo WSC (*Whole Sky Cameras)* que se utilizan ampliamente en tareas de análisis y pronóstico. Para esto, se deben utilizar técnicas para el procesamiento de imágenes y el reconocimiento de patrones que permitan rescatar los atributos cualitativos y cuantitativos relevantes de la nubosidad y luego utilizar algún método que permita derivar de dichos atributos la información que se necesita obtener.

Mientras que en algunos trabajos se utilizan técnicas para clasificación de baja complejidad como en [1], también se pueden encontrar algunas propuestas para la implementación de un clasificador basado en técnicas tomadas del campo de la inteligencia artificial y el aprendizaje automático. Por ejemplo, se han utilizado Redes Neuronales Artificiales (ANN) para la clasificación de imágenes satelitales multi-canal [2], máquinas de vectores de soporte (SVM) para la clasificación de datos obtenidos desde imágenes MODIS (*Moderate Resolution Imaging Spectroradiometer)* [3] como las de la Figura 1.2. En [4] se han obtenido buenos resultados con un clasificador basado en instancias (kNN) sobre imágenes de cielo completo (WSI).

En [1], se propone la extracción de un número de *features* o características que permiten describir cuantitativamente un conjunto de aspectos en las imágenes como la de la Figura 1.3 obtenidas con instrumentos de observación del cielo como *Whole Sky Cameras* y *Total Sky Imagers*. Las características propuestas en dicho trabajo podrían ser importantes a la hora de implementar un algoritmo de clasificación automática y en dicho trabajo se mencionan algunas técnicas para el procesamiento de imágenes idóneas para distinguir nubosidad. Además emplean un clasificador basado en paralelepípedos sobre el gráfico de características para evaluar el desempeño del mismo con las características propuestas. Los mismos autores señalan, en sus conclusiones, que las características propuestas aportan un alto grado de discriminación para el tipo de nubosidad y que los resultados de clasificación obtenidos podrían ser mucho mejores con la implementación de un algoritmo de clasificación más complejo.

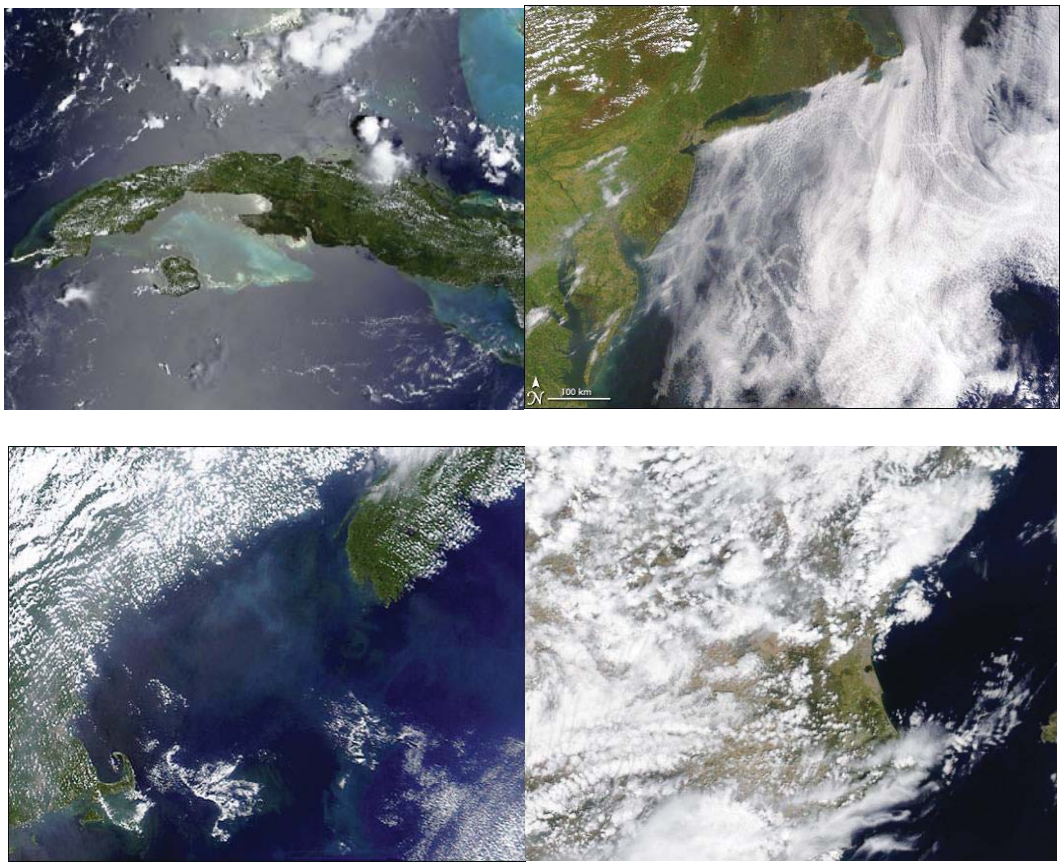


Figura 1.2: Imágenes satelitales obtenidas por un *Moderate Resolution Imaging Spectroradiometer*

Figura 1.3: Imagen obtenida desde una *Whole Sky Camera*

2 Análisis de Objetivos

A continuación se definen los objetivos y el alcance del trabajo a realizar. La justificación de estos objetivos se desprende de los antecedentes que se entregan en esta sección.

Originalmente, el problema al que se propone dar solución es la imposibilidad de detectar y alertar sobre errores en la descripción de la nubosidad que yacen en la génesis de la observación. Respecto a este problema se consideró lo siguiente: si se tuviese la capacidad de determinar de forma automática y con un buen grado de exactitud esta descripción, sería posible cuestionar la decisión humana y así detectar un buen porcentaje de los errores u omisiones en las observaciones. Se considera además lo siguiente: existen antecedentes sobre la utilización de algoritmos de aprendizaje automático para abordar tareas de reconocimiento de patrones y clasificación de imágenes con resultados desde moderados a muy buenos. Se propone entonces el objetivo principal de este trabajo: implementar y evaluar el desempeño de un clasificador automático sobre imágenes meteorológicas de condición nubosa.

Es ampliamente aceptado que en la mayoría de las tareas de clasificación bajo el enfoque *machine learning* la preparación de los datos y la extracción de características, procesos que podrían ser agrupados en una tarea llamada representación, tiene un impacto relevante sobre el desempeño final del algoritmo de clasificación y por esta razón se debe especificar un conjunto de características que permitan describir transversalmente una colección de imágenes en términos de atributos gráficos que faciliten diferenciar distintas clases de nubosidad. Al mismo tiempo, es importante determinar si existe información que evidencie clases de nubosidad más susceptibles a errores al momento de ser identificadas, es decir, que muestren un mayor grado de dificultad en la clasificación tanto humana como automática, si esto es así, sería importante evaluar la incorporación de información adicional a la que puede ser obtenida de las imágenes, considerando que dicho material generalmente se acompaña con otros datos como la temperatura, humedad, etc. Con esta información se podrían implementar meta-clases que apoyen la validación de observaciones en vez de clasificar sobre las clases originales y así verificar el desempeño del clasificador en dichas clases.

Es importante reiterar que este trabajo no pretende automatizar completamente las tareas de observación y clasificación de nubosidad ni tampoco se dispone de los recursos para desarrollar una solución que se adapte a los sistemas de datos específicos de la DMC, sin embargo los resultados obtenidos podrían influenciar la decisión de implementar un sistema bajo esta línea e incluirlo en los flujos de trabajo meteorológico.

2.1 Objetivo General

Implementar un clasificador automático de nubosidad y evaluar su efectividad al usarlo como apoyo al control de calidad de observaciones meteorológicas.

2.2 Objetivos Específicos

- Confeccionar una colección de imágenes meteorológicas clasificadas por expertos y definir un método de representación adecuado para clasificación automática.

- Especificar un clasificador automático bajo el enfoque de *machine learning* y un determinado framework de clasificación, y entrenarlo con el dataset derivado de la colección.

- Evaluar el desempeño del clasificador automático en la asignación de clases y en última instancia, su capacidad para detectar errores en observaciones al comparar la decisión de la máquina y la del humano.

3 Metodología

El dominio del problema que se aborda es muy específico y en consecuencia se reconoce la necesidad de iterar sobre algunas de las etapas de investigación en el contexto de un trabajo apoyado por especialistas de la Dirección Meteorológica de Chile. La experiencia previa que se tiene sobre clasificación automática supervisada de textos bajo el enfoque de *machine learning*, permite tener una estimación sobre las actividades críticas, en función del tiempo que requieren. Estas actividades se han agrupado en ítems de investigación como se muestra en la Figura 3.1 y las actividades asociadas se detallan a continuación.

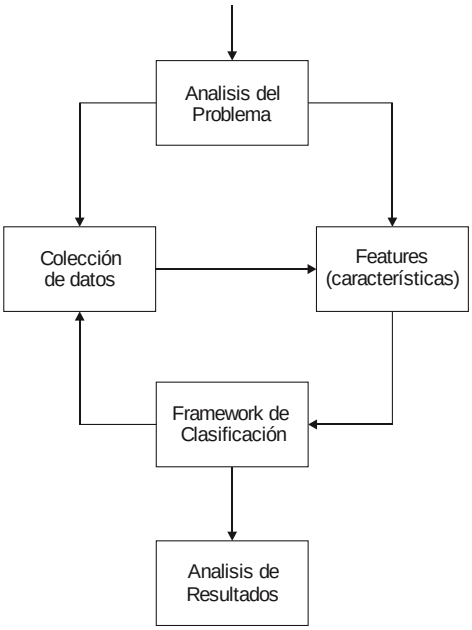


Figura 3.1: Elementos de investigación en el estudio

3.1 Análisis del Problema

Inicialmente se obtienen los antecedentes necesarios para definir el problema y el alcance de lo que se quiere realizar. Estas actividades son apoyadas moderadamente por el aporte de antecedentes que realiza personal de la DMC.

- Recopilación de antecedentes en Dirección Meteorológica de Chile.
- Análisis de la necesidad de estudio.
- Análisis bibliográfico.

El producto de esta etapa se concreta en la definición del problema entregado en la sección 1.2 y el establecimiento de los objetivos definidos en el capítulo 2.

3.2 Colección de Datos

Este ítem se relaciona con la etapa de recolección de datos en una investigación y, particularmente en este trabajo, con las actividades necesarias para la confección de un set de datos útiles para el entrenamiento de un clasificador supervisado bajo el enfoque de *machine learning*. Se reconoce una amplia participación del personal de la DMC en el análisis y control de la calidad de los datos. Se identifica como riesgo la poca disponibilidad de recursos para tareas que requieren un gran esfuerzo (inspección de las imágenes, validación de observaciones, etc.). Las actividades contempladas para este ítem son:

- Identificación de los instrumentos y sus datos.
- Recolección de imágenes desde instrumentos y repositorios.
- Recolección de datos de observación desde banco nacional.
- Análisis de topologías disponibles para la nubosidad.
- Análisis de calidad de los datos.
- Validación de datos.

3.3 Features

Es reconocido que en clasificación automática supervisada, las *features* o características que se utilizan para representar los datos tienen un gran impacto sobre el desempeño de los clasificadores, esto se explica con mayor detalle en la sección 4.5. Por esto se han agrupado las actividades que conforman la investigación realizada sobre este tópico, las cuales se listan a continuación:

- Investigación bibliográfica.
- Definir *features.*
- Herramientas de procesamiento de imágenes.
- Construcción del data-set.
- Estadísticas del data-set.

3.4 Framework de Clasificación

La naturaleza de un problema de clasificación automática se desprende principalmente de la topología empleada para clasificar. En el caso de la nubosidad, dada la existencia de múltiples topologías, se reconoce la aplicabilidad de más de un framework de clasificación, lo que se presentará con mayor detalle en la sección 4.7. Por lo anterior, se ha definido la selección del framework como una de las decisiones que deberán tomarse durante el desarrollo de la investigación. El framework de clasificación define como se transforman los datos que serán

clasificados en términos de categorías y por lo tanto tiene directa incidencia con la confección de la colección y los data-sets derivados de la misma. Además, define como interactúan los datos con el algoritmo de clasificación automática empleado por lo que este último también cae dentro del marco de investigación de este ítem. En resumen, las tareas definidas para este ítem son:

- Investigación bibliográfica.
- Algoritmo de clasificación.
- Definir framework de clasificación.
- Transformaciones al set de datos.
- Investigar herramientas (*data-mining*, *machine learning*).
- Evaluaciones de desempeño.

3.5 Plan de Trabajo

Se desarrolló un plan de trabajo que, además de incluir las tareas propias de investigación, agrupa en actividades todas las tareas relacionadas a un mismo ítem de investigación definido en la metodología. En estas tareas se aplican de manera práctica los conceptos y técnicas analizados en cada ítem y permiten concretar una fase del proceso de clasificación automática. Adicionalmente se incluyeron algunas tareas de apoyo no mencionadas en los ítems de investigación como lo son la verificación de imágenes en conjunto con especialistas de la DMC. El resumen de estas actividades y tareas se muestra en la siguiente lista.

- Inicio
    - Recopilación y análisis bibliográfico.
    - Análisis de los objetivos.
    - Establecimiento de la metodología.
- Confección del Data-Set
    - Recopilación de datos (imágenes).
    - Descripción de los datos.
    - Prototipo para extracción de características.
    - Verificación de la colección de imágenes.
- Desarrollo del clasificador
    - Implementación de prototipos de clasificación.
    - Desarrollo de pruebas.
    - Calibración y validación del clasificador.
- Clasificación Automática de la colección
    - Clasificación del set de imágenes validadas.
    - Obtención de métricas de rendimiento.
    - Análisis de resultados.
- Evaluación en la detección de errores
    - Confeccionar colección de datos con error.
    - Descripción de errores.

- Especificar criterio de detección.
- Clasificación automática de datos con error.
- Análisis de resultados.

4 Solución Propuesta

4.1 Clasificación Automática Supervisada

La solución que se propone para el problema de clasificación de la nubosidad consiste en la aplicación de un algoritmo de clasificación automática configurado en base a alguno de los frameworks de clasificación supervisada conocidos como clasificación de etiqueta simple, multi-etiqueta y multi-instancia, los que se detallarán en la sección 4.7. La clasificación automática bajo el enfoque de aprendizaje automático o *machine learning* como se conoce en inglés, propone la construcción de un proceso inductivo general el cual construye automáticamente un clasificador para las categorías que se desean predecir, observando las características de un conjunto de instancias que han sido previamente categorizados de forma manual por un experto en el dominio; esta observación se realiza durante una etapa de entrenamiento, y en ella el proceso inductivo deduce las características que una instancia debe tener para ser clasificada bajo las distintas clases. En la Figura 4.1 se entrega un modelo conceptual que describe las actividades contempladas en el entrenamiento y evaluación de un clasificador automático bajo el enfoque de aprendizaje automático.

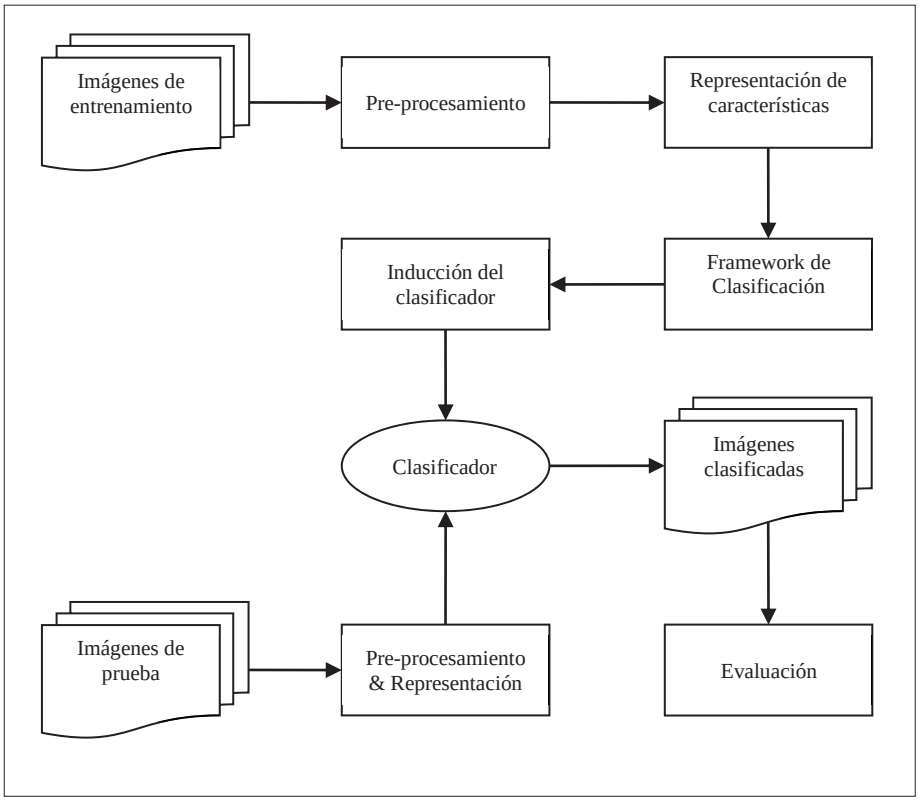


Figura 4.1: Modelo conceptual de clasificación automática bajo el enfoque *Machine Learning*

La solución al problema de reconocimiento de la nubosidad bajo el enfoque propuesto contempla 4 tareas necesarias para la implementación de un clasificador automático semisupervisado. La primera tarea tiene que ver con la especificación y preparación de los datos que serán clasificados, en este caso las imágenes digitales de nubosidad. Lo que se debe especificar para esta etapa son la fuente y la metodología para la obtención de las imágenes. La sección 4.4 detalla esta especificación, la cual es de gran importancia puesto que el tratamiento que se le da a los datos para entrenar el clasificador debe ser el mismo que se aplicará más tarde cuando dicho clasificador ya entrenado sea utilizado para discriminar nuevas instancias (en este caso imágenes), o por decirlo de otra manera el clasificador entre en la etapa operacional.

La segunda tarea consiste en la extracción de características o "*features*" desde estos datos preparados. En machine learning, es común hablar de atributo cuando se quiere indicar una propiedad del objeto o fenómeno real que será representado y *feature* o característica a una métrica (binaria o real valuada) que caracteriza dicho fenómeno en el modelo abstracto que se construye para representarlo. La obtención de características consiste, entonces, en la construcción de variables cuantitativas que puedan representar algún atributo cuantitativo o cualitativo relevante a la hora de discriminar entre las clases de nubosidad que se pretenden discernir. Estas características deben ser susceptibles de ser utilizadas por el algoritmo de clasificación y por lo tanto deben cumplir ciertas exigencias de rango y valor, sin dejar de ser representativas del atributo que quieren caracterizar. En la sección 4.5 se definen un conjunto de características endógenas a la imagen de nubosidad que serán utilizadas en este trabajo, éstas han sido propuestas en [1] y en otros trabajos como características afines para la clasificación de nubosidad. Adicionalmente, en la sección 4.6, se detalla la incorporación de características exógenas provenientes de información meteorológica que normalmente acompaña dichas imágenes como la temperatura y la presión. La inclusión de estas características mejora de manera importante los resultados de clasificación lo que se cuantifica en la sección 4.10.2.

La mayoría de los algoritmos especializados en clasificación trabajan sobre un escenario de clasificación binaria, esto significa que cada instancia se clasifica en una de 2 clases posibles. Generalmente, para poder obtener las predicciones en una tarea de clasificación que involucra múltiples clases (más de 2 clases: clasificación multi-clase), es necesario definir un arreglo de clasificadores de forma tal que las decisiones o la "salida" de todos ellos se conjugan para obtener la clasificación final. Esto también aplica en parte, cuando las clases no son mutuamente excluyentes y funcionan más como una "etiqueta", es decir una misma instancia puede ser categorizada con una o más etiquetas, escenario aludido en la literatura como clasificación multi-etiqueta. Recientemente, se han planteado los escenarios de clasificación recién mencionados (binaria, multi-clase y multi-etiqueta) como una versión degenerada de un escenario llamado clasificación multi-instancia multi-etiqueta en el que cada muestra u objeto que conforma la entrada del clasificador (en nuestro caso la imagen a ser clasificada) está asociada a varias instancias y al mismo tiempo, a varias clases. La tercera tarea entonces consiste en la definición de un clasificador y el framework de clasificación. Esto incluye el algoritmo que será empleado y las transformaciones y estrategias correspondientes al framework de clasificación que se empleará.

Definidos los métodos para representar las imágenes y el clasificador automático, la cuarta tarea consiste en evaluar el desempeño de la clasificación y si corresponde, volver a ajustar algunos parámetros de las 3 etapas anteriores para volver a entrenar y evaluar. El desempeño del clasificador se evaluará en base a mediciones de exactitud y cobertura, estas métricas son descritas en detalle en la sección 4.8.

4.2 Clases de Nubosidad

Pese a los numerosos estudios realizados sobre la nubosidad, se reconoce ampliamente que aún queda mucho por entender respecto a este fenómeno. Se desconocen todos los efectos que pueden tener en conjunto los distintos géneros de nubosidad, lo que se traduce en un gran factor de incertidumbre en modelos climatológicos y predicciones meteorológicas.

Distintas topologías y métodos de clasificación han sido propuestos para describir los tipos de nubosidad y en ese contexto la Organización Mundial de Meteorología (OMM) ha definido diversas topologías de nubes, de las cuales algunas rescatan con mucho detalle la forma, composición y color de las nubes clasificándolas en géneros, especies y variedades con el objeto de alimentar los actuales modelos climatológicos y apoyar los futuros estudios sobre el fenómeno. La Tabla 4.1 describe los 10 géneros de nubes definidos en el Atlas Internacional de Nubosidad (AIN) de la OMM [5].

Tabla 4.1: Clasificación de nubes por familia OMM AIN

Genero	**Cód.**	**Símbolo**	**Descripción**
Cirrus	Ci		Nubes separadas en forma de filamentos blancos y delicados, o de bancos o bandas estrechas, blancas o casi blancas. Estas nubes tienen una apariencia fibrosa, semejante a los cabellos de una persona, o de un brillo sedoso o de ambas características a la vez.
Cirrostratus	Cs		Velo nuboso, transparente y blanquecino, de aspecto fibroso (como cabellos) o completamente liso, que cubre total o parcialmente el cielo y que produce generalmente el fenómeno de halo.
Cirrocumulus	Cc		Banco, capa delgada o sábana de nubes blancas, sin sombras, compuestas por elementos muy pequeños en forma de granos, rizos, grumos, ondulaciones, unidos o separados y distribuidos con mayor o menor regularidad; la mayoría de los elementos tiene una anchura aparente < 1º.
Altocumulus	Ac		Banco, capa delgada o capa de nubes blancas o grises, o a la vez blancas y grises, que tienen sombras compuestas por losetas, masas redondeadas, rodillos, etc., las cuales son a veces parcialmente fibrosas o difusas y que pueden estar unidas o no; la mayoría de los elementos pequeños distribuidos con regularidad tienen una anchura aparente comprendida entre 1º y 5º.

Altostratus	As		Lámina o capa de nubes, grisácea o azulada, de aspecto estriado, fibroso o uniforme, que cubre por entero o parcialmente el cielo, como una gran sábana. Tiene partes suficientemente delgadas que permiten distinguir vagamente el Sol, como a través de un vidrio deslustrado. Los Altostratus, a diferencia de los Cirrostratus, no producen halos.
Nimbostratus	Ns		Capa de nubes gris, a menudo oscura, con un aspecto velado por la precipitación de lluvia o nieve que cae más o menos continuadamente desde ella. El espesor de la nube es lo suficientemente grande como para ocultar el Sol completamente.
Stratus	St		Capa de nubes generalmente gris, con base uniforme, de la que pueden caer llovizna, prismas de hielo o granizo. Cuando el Sol es visible a través de la capa, su contorno se distingue claramente. Los St se presentan a veces en forma de jirones deshilachados (fractus), debajo de otras nubes.
Stratocumulus	Sc		Banco, sábana o capa de nubes grises o blanquecinas, que tienen casi siempre partes oscuras; compuestas por losetas, masas redondeadas, rodillos, etc., no fibrosas, que pueden estar o no unidas.
Cumulus	Cu		Nubes asiladas, en general densas y con contornos bien definidos, que se desarrollan verticalmente en forma de protuberancias, cúpulas o torres, y cuyas partes superiores convexas se parecen con frecuencia a una coliflor. Las partes de estas nubes iluminadas por el Sol son blancas brillantes; su base es oscura y horizontal. A veces, aparecen desgarrados por el viento.
Cumulonimbus	Cb		Nube amazacotada y densa, con un desarrollo vertical considerable, en forma de montaña o de enormes torres. Parte, al menos de su cima es normalmente lisa, fibrosa o estriada, y casi siempre aplastada; esta parte se extiende a menudo en forma de un yunque o de un vasto penacho. Por debajo de la base, muy oscura, aparecen nubes bajas desgarradas y precipitaciones o chubascos.

Estos géneros, a excepción de las nubes con desarrollo vertical (Cb), están asociados típicamente a una altura específica y las observaciones realizadas en base a los sistemas derivados del AIN pueden describir para una observación, cero (*sky-clear*), uno o más géneros de nube simultáneamente. El desarrollo de un clasificador automático supervisado que identifique estos géneros para una observación debería considerar entonces el framework de clasificación Multi-Etiqueta, en donde una observación puede tener una o más "etiquetas" asignadas, lo que se verá con más detalle en la sección 4.7.1.

También se han definido clasificaciones fenomenológicas que buscan identificar las condiciones nubosas con efectos significativos en algún dominio de estudio específico. Un ejemplo de estas topologías es la clasificación entregada en [4] la cual podría tener importancia en términos de balance energético. En dicha topología, se definen 7 clases fenomenológicas de la condición nubosa en el área de observación de una estación de monitoreo, estas condiciones se presentan en la Tabla 4.2.

Observaciones meteorológicas realizadas sobre la condición atmosférica en base a este método entregarían una de las 7 categorías mutuamente excluyentes definidas en esta tabla, por lo que la definición de un clasificador automático supervisado que discrimine entre estas categorías podría trabajar en función de algún framework de clasificación de etiqueta simple.

Tabla 4.2: Clasificación fenomenológica de la condición nubosa

Cód.	Condición Nubosa	Géneros de Nube OMM
1	Sky-Clear (cielo despejado)	Sin Nubes, o nubosidad menor a 1 octa.
2	Nubes bajas y esponjosas con bordes bien definidos, de color blanco o gris claro	Cumulus
3	Bancos de nubes altas y delgadas, que pueden cubrir todo o casi todo el cielo, de color blanquecino	Cirrus & Cirrostratus
4	Nubosidad alta fragmentada en nubes pequeñas, tipo mosaico, blancas	Cirrocumulus & Altocumulus
5	Nubosidad baja o media, distribuida en bultos. Condición casi nublada de tono blanco o gris.	Stratocumulus
6	Capa uniforme de nubes bajas o medias, nublado generalmente gris	Stratus & Altostratus
7	Nubes gruesas y oscuras. Mayormente nublado y gris	Cumulonimbus & Nimbostratus

4.3 Colección de Imágenes Clasificadas

En Climatología, la nubosidad es uno de los fenómenos más importantes pues influyen en el balance energético del planeta, el ciclo hidrológico, la distribución del calor, etc. Los efectos más importantes están dados por la reflexión o absorción de radiación solar y esta propiedad estará definida principalmente por el tipo de nube. Puesto que los efectos más importantes tienen que ver con la interacción entre el sol y las nubes, se le da especial importancia a la descripción de la nubosidad en periodos de Hora de Sol, es decir en el intervalo de tiempo durante el día donde el ángulo cenital solar supera los 20º. El Periodo de Hora de Sol varía según el momento del año en el que se mide por lo que para cada Mes se consideraran las imágenes que están dentro del intervalo de hora sol promedio de dicho mes menos 1 periodo de observación horario.

Mientras el instrumento captura imágenes cada 10 minutos, la observación que se pretende validar es representativa de la hora inmediatamente anterior al momento de la observación. En la práctica, la descripción de la nubosidad se realiza en algún momento dentro de un intervalo de 10 minutos antes del instante de observación, sin embargo un cambio significativo se puede dar en promedio dentro de un intervalo no menor a 15 minutos.

4.3.1 Fuente de la Colección

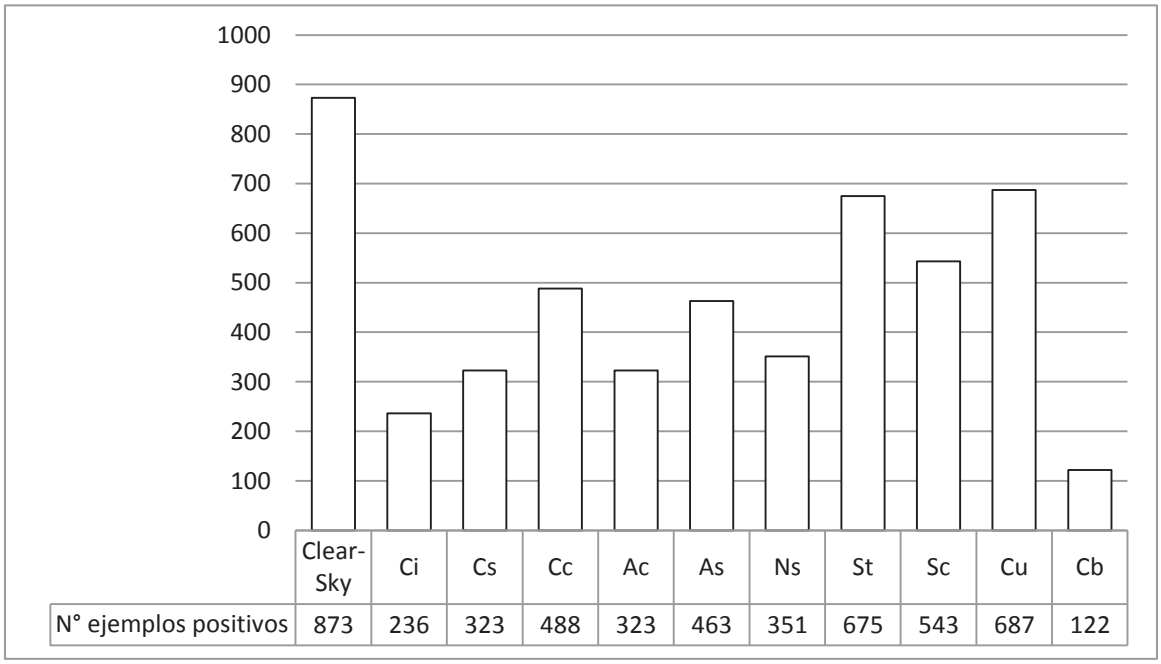
Las imágenes de nubosidad serán aportadas por la Dirección Meteorológica de Chile, éstas consisten en archivos de imagen digital con instantáneas de cielo completo capturadas por TSI instalados en la estación de monitoreo Sinóptico Quinta Normal. El momento de captura de estas imágenes se corresponde con observaciones horarias desde las 12.00 UTC a las 20.00 UTC

durante el año 2011 (febrero a diciembre) por lo que se obtendrá la descripción de nubosidad realizada por Técnicos observadores en dichas observaciones desde el Sistema de Administración de Datos Climatológicos SACLIM. Adicionalmente se incorporaron imágenes TSI desde la estación sinóptica aeronáutica, El Tepual, Puerto Montt con el fin de disponer de una mayor variedad de condiciones de nubosidad en la colección.

4.3.2 Estadísticas de la Colección

La colección de datos clasificados contiene 2922 imágenes de las cuales 873 corresponden a condición cielo despejado con distinto ángulo cenital solar. La distribución de ejemplos positivos según género de nube y para la condición sky-clear se muestra en el gráfico de la Figura 4.2.

La superposición de clases en la colección puede ser cuantificada por la cardinalidad, la cual se define como el número promedio de etiquetas presentes en cada imagen a lo largo de la colección En la colección de imágenes clasificadas que se está construyendo, la etiqueta se refiere a un tipo de nube informada en la observación que es representativa de la imagen capturada por el instrumento (TSI). Los tipos de nubes que pueden informarse en cada observación, corresponden a alguno de los géneros del atlas internacional para la clasificación de nubes definido por la OMM explicado en la Tabla 4.1.



	Clear-Sky	Ci	Cs	Cc	Ac	As	Ns	St	Sc	Cu	Cb
N° ejemplos positivos	873	236	323	488	323	463	351	675	543	687	122

Figura 4.2: Ejemplos positivos para cada clase en la colección original

De la colección completa de imágenes, se generó un subconjunto con aquellas imágenes que presentan al menos un tipo de nube, es decir el conjunto complemento de aquellas imágenes

clasificadas como sky-clear (SKC). En la Tabla 4.3 se entregan algunas estadísticas de la colección completa y del subconjunto que considera solo las imágenes con nubosidad presente (no sky-clear). Para el cálculo de la cardinalidad y de las clases distintas en la colección completa, se consideró la condición sky-clear como una etiqueta más que puede salir en combinación con los otros tipos de nube, esto se ajusta para la clasificación que considera todos los pixeles de la imagen incluyendo los que corresponden a regiones de cielo sin aerosoles (sin nubes), la variable 'distinto' indica el número de combinaciones de etiqueta distintas que pueden encontrase en la colección. En el caso de entrenar un clasificador por separado para las imágenes con presencia de nubosidad, el promedio de categorías asignadas es cercano a 2 con una desviación estándar de 0.764.

Tabla 4.3: Estadísticas de la colección de imágenes de nubosidad

	Colección Completa	Nubosidad Presente
Total de imágenes	2922	2049
Cardinalidad	2,852	2,136
Distinto	35	34
SKC	873	0

4.3.3 Validación de la Colección

Se validaron las imágenes junto con la clasificación obtenida desde el sistema SACLIM para las observaciones que presentaban la condición sky-clear, en base a una inspección visual realizada por un técnico validador del centro de análisis de la DMC. Cabe mencionar que en esta inspección se detectaron 132 observaciones de sky-clear que no eran representativas de ninguna de las imágenes obtenidas en los intervalos de 10 minutos correspondientes al periodo horario de la observación. Esto dio cuenta de los problemas de calidad del dato aludidos en la sección 1.1 que justifican el investigar formas de apoyar el proceso de control de calidad y al mismo tiempo alertó sobre la necesidad de verificar los datos de la colección para descartar problemas de clasificación sistemáticos en el algoritmo que se está implementando. En consecuencia, se definió una tarea de validación de los datos en la colección descrita, junto a personal de la DMC, que consistió en una inspección visual sobre las mismas imágenes por dos técnicos validadores. El objetivo de esta validación fue usar la decisión de los técnicos validadores para corregir la clasificación presente en el sistema y al mismo tiempo, analizar las posibles diferencias entre la clasificación de ambos especialistas para de este modo tener una idea de los conjuntos de clases con un alto grado de dificultad para diferenciar. En un problema de clasificación, cuando 2 expertos difieren reiteradamente en la clasificación de instancias sobre una u otra clase, se habla de un problema de inconsistencia inter-indizador y podría requerirse un análisis mayor sobre la consistencia de la topología sobre la cual se quiere clasificar. En la sección 5.1 se profundiza más sobre ésta problemática la cual toma mayor importancia cuando se quiere implementar y evaluar un proceso de clasificación automática para comparar la decisión de la máquina y la del humano.

En el gráfico de la Figura 4.3 se entrega la distribución de ejemplos positivos en la colección tras las revisiones y validaciones realizadas, mientras que en la Tabla 4.4, se entregan algunas estadísticas para 2 subconjuntos derivados de la colección, el primero considera todas las imágenes incluyendo aquellas que presentan la condición sky-clear (SKC) y el segundo solo las que presentan nubosidad. Al igual que en la sección anterior, para el cálculo de la cardinalidad, se considera la condición SKC como una etiqueta adicional para aquellas imágenes con regiones de cielo despejado.

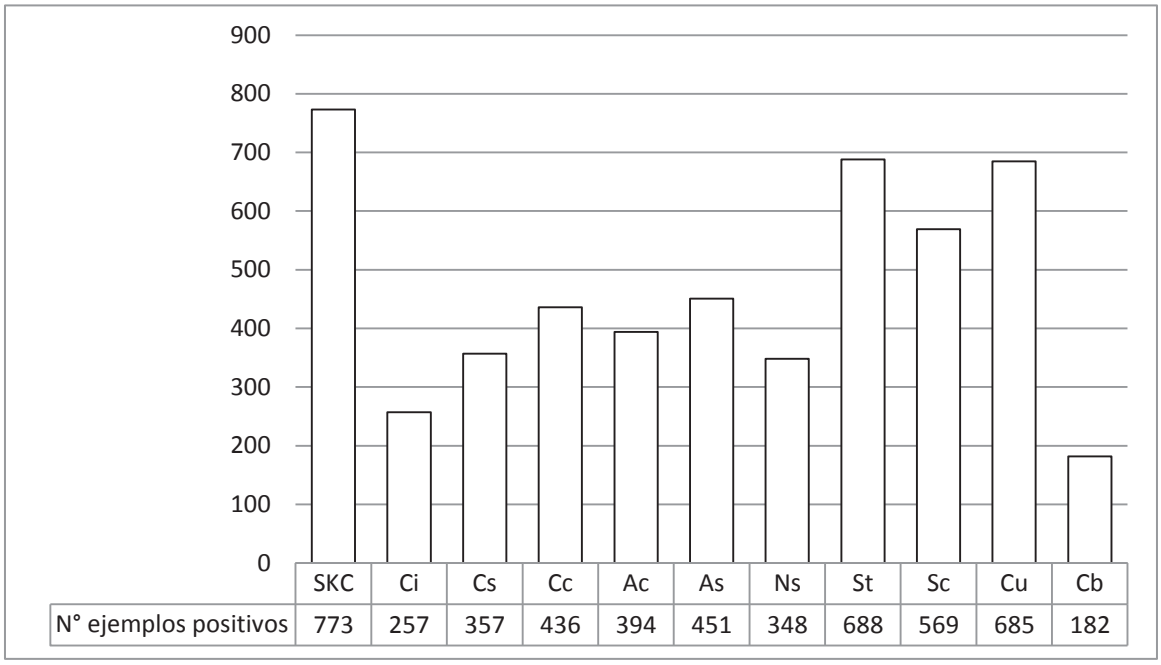


Figura 4.3: Ejemplos positivos para cada clase en la colección validada

Tabla 4.4: Estadísticas de la colección validada

	Colección Completa	Nubosidad Presente
Instancias	2400	1627
Cardinalidad	2.14	2.63
Distinto	28	27
SKC	773	0

Cabe mencionar que por un lado, se han incorporado alrededor de 60 imágenes con ejemplos positivos para el género Cumulonimbus el cual presentaba un déficit en la colección original, en

un esfuerzo por balancear esta categoría. Por otro lado se han eliminado aproximadamente un mes de observaciones para las cuales existió un problema con los datos extraídos del banco nacional según indicaciones de la DMC.

4.4 Procesamiento de las Imágenes

Antes de abordar los métodos para extracción de características desde las imágenes digitales que se utilizarán, es preciso definir qué es una imagen digital, cuáles son sus componentes, las principales vías de obtención de las imágenes de nubosidad y en que consiste la segmentación de imágenes digitales. Esto último tiene importancia a la hora de querer diferenciar las regiones (vecindades de pixeles) en la imagen que corresponden a nubosidad de las que presentan cielo despejado (sky-clear).

4.4.1 Imagen digital

Una imagen digital de tonalidades de grises, puede ser vista como una matriz de dimensiones *h* x *v* donde *h* y *v* representan la cantidad de filas y de columnas de la matriz, respectivamente (Figura 4.4). Los índices de las filas y las columnas sirven para identificar un punto de la imagen, mientras que el valor numérico de un elemento de la matriz está asociado a alguna magnitud física medida. Cada punto de la imagen se denomina también *píxel* (*picture element*).

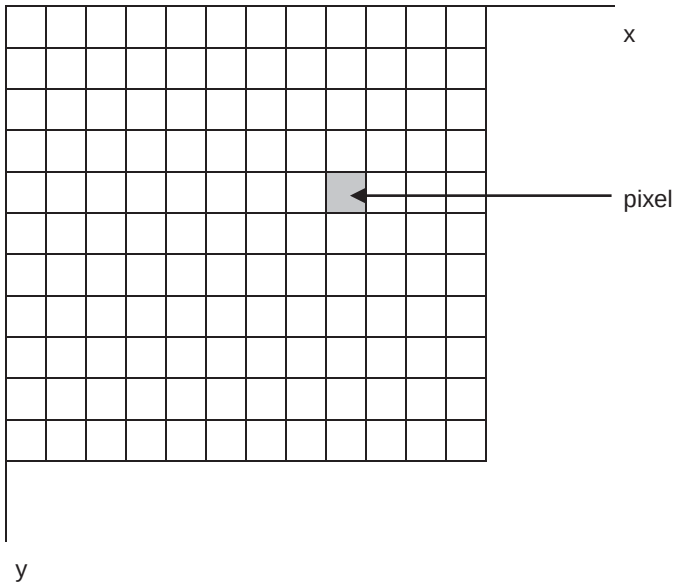


Figura 4.4: Representación matricial de una imagen digital monocromática

Para generar una imagen digital se realiza un muestreo a lo largo de los ejes de coordenadas, por conveniencia orientados según se muestra en la Figura 4.4. De esta manera se obtiene una matriz

de $h \times v$ muestras, cada una de las cuales representa un píxel. Se dice entonces que la imagen tiene una dimensión de h píxeles por filas, y v píxeles por columnas.

4.4.2 Segmentación de imágenes

Para la extracción de algunas características que conformaran el espacio vectorial de la representación de las imágenes se utilizarán algunos métodos de segmentación de imágenes. La segmentación involucra subdividir una imagen en partes constituyentes o aislar ciertos aspectos presentes en la imagen, como detectar líneas, círculos, formas o regiones específicas. Se puede plantear que la segmentación es el proceso en el que se extraen los objetos de interés presentes en las imágenes. El resultado de la segmentación será la frontera de un objeto respecto a su exterior, o los puntos que forman su interior. La forma de representación del resultado de la segmentación está asociada a la manera de representar ya sea la frontera o el interior del objeto. Las fronteras de los objetos delimitan las regiones que estos ocupan. Una vez que se hayan obtenido las fronteras es posible, en principio, reconstruir la regiones. De manera similar, si se tiene una partición de una imagen en regiones es posible hacer un seguimiento de la frontera de cada región.

Para el caso particular de las imágenes de nubosidad una técnica frecuente para segmentar la imagen en áreas despejadas y áreas nubosas yace en la comparación de los niveles de intensidad en los canales rojo (R) y azul (B). En una atmosfera despejada, sin presencia de aerosoles, las moléculas de gas dispersan una mayor cantidad de luz azul en comparación a la luz roja, es por este motivo que el cielo despejado se aprecia de color azul. Al contrario, la nubosidad, que contiene partículas como aerosoles, gotas de agua y cristales de hielo, dispersan luz azul y luz roja en una proporción similar, motivo por el cual se aprecian de color blanco o gris. Dada esta propiedad del cielo, las regiones de una imagen que corresponden a cielo despejado mostraran valores de rojo relativamente bajos en sus pixeles, comparados con aquellos que pertenezcan a regiones nubosas y por lo tanto es posible utilizar alguna razón entre los valores de rojo y azul para diferenciar ambas regiones, en donde se debe determinar algún umbral separador el cual dependerá tanto de la cámara utilizada como de las condiciones de luminosidad del emplazamiento del instrumento.

En [6] proponen la determinación de un umbral utilizando la relación R / B evaluando el resultado vía inspección visual de la segmentación resultante, en dicho trabajado, valores cercanos a R/B = 8 serían apropiados para separar las regiones con pequeñas variaciones según el tipo de CCD (diodo digital para captura de imagen) incorporado en el instrumento. En este trabajo se ha definido utilizar la relación R – B, empleada en [4], en este último trabajo, los autores obtuvieron mejores resultados para diferenciar zonas de la imagen aledañas a la circunferencia solar de zonas pertenecientes a nubosidad gruesas con partes translucidas. Para el TSI desplegado en la estación de Quinta Normal, un valor de R – B = 26 entregó los mejores resultados, los que pueden verse en la Figura 4.5.

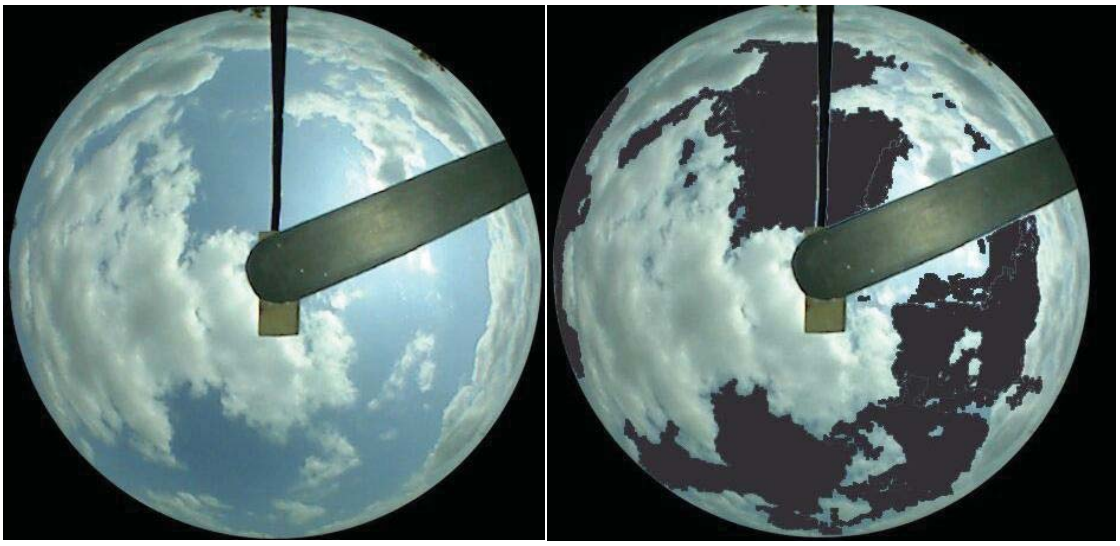


Figura 4.5: Derecha: imagen TSI RGB, izquierda: pixeles del segmento Cp (opacados)

4.5 Obtención de características

4.5.1 Textura

Un enfoque frecuente para el análisis de textura de las imágenes se basa en las propiedades estadísticas del histograma de valores de gris. Estos métodos no son aplicables a imágenes a color ya que presentan múltiples valores para cada pixel, en cambio para una imagen en escala de grises cada pixel se describe con un único valor numérico. Si se dispone de imágenes a color, éstas deben ser transformadas a escala de grises, y para esto se puede ocupar la taza de composición entre canales de color, por ejemplo, taza de composición rojo-azul (R/B), u obtener la imagen de valores de intensidad de color que se define como un tercio de la suma de valores de cada canal de color RGB (rojo, verde, azul).

Del histograma de valores de gris obtenido, se pueden cuantificar una variedad de estadísticas las que serán utilizadas en este trabajo para construir los atributos o *'features'* endógenos de la imagen, las secciones siguientes detallan cada una de estas estadísticas.

4.5.1.1 Característica: Promedio (ME)

$$\mathrm{ME} = \sum_{i=0}^{L-1} z_i p(z_i)$$

Donde z es la variable que señala los valores en la imagen en escala de grises (R/B, o intensidad), p(z) es la distribución de frecuencia de estos valores en la imagen, y *L* es el número de posibles valores de z (256 en una imagen de escala de grises de 8 bits).

4.5.1.2 Característica: Desviación Estándar (SD)

$$\mathrm{SD} = \sqrt{\sum_{i=0}^{L-1} (z_i - \mathrm{ME})^2 p(z_i)}.$$

La desviación estándar entrega una medida del contraste en la imagen.

4.5.1.3 Característica: Fundido (SM: *Smoothness*)

$$\mathrm{SM} = 1 - \frac{1}{(1 + \sigma^2)}$$

Donde $\sigma^2 = \mathrm{SD}^2/(L-1)^2$. Se refiere al fundido, suavizado o difuminado de la imagen, en la ecuación Los valores de SM caen dentro del rango [0, 1]: SM es 0 para una imagen de valores constantes (valor de grises en cada pixel) y 1 para una imagen de gran variabilidad.

4.5.1.4 Característica: Tercer Momento (TM)

$$\mathrm{TM} = \sum_{i=0}^{L-1} (z_i - \mathrm{ME})^3 p(z_i)$$

El tercer momento entrega una medida del sesgo en el histograma.

4.5.1.5 Característica: Uniformidad (UF)

$$\mathrm{UF} = \sum_{i=0}^{L-1} p^2(z_i)$$

El valor de la uniformidad alcanza un máximo cuando todos los pixeles en la imagen presentan el mismo nivel de gris, y un mínimo cuando todos presentas distintos niveles de gris.

4.5.1.6 Característica: Entropía (EY)

$$\mathrm{EY} = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$$

La entropía entrega una medida de la aleatoriedad en las diferencias de niveles de gris presentes en la imagen.

4.5.2 Patrones y Espectro

Muchos de los atributos propuestos en la literatura para describir imágenes entregan alguna característica global, pero son pocos los métodos que se enfrentan al reconocimiento de algún tipo de patrón o formas presentes en las mismas. Para disponer de un atributo que considere la forma de las nubes, se ha propuesto el uso de algoritmos de Transformación Rápida de Fourier FFT (*Fast Fourier Transform*), un tipo de algoritmo especializado en calcular eficientemente la transformación de Fourier Discreta (DFT), cuyo cómputo en base a la definición directa sería demasiado lento y poco práctico. Una transformación de Fourier discreta descompone una secuencia de valores en componentes diferenciados por su frecuencia, esta operación ha sido útil en una variedad de aplicaciones incluyendo el procesamiento de señales y de imágenes. En [7] se afirma que, en general, el análisis de las imágenes en términos de frecuencia en los canales, (como lo es el resultado de algunos FFT como puede verse en la Figura 4.6), puede ser una herramienta importante a la hora de obtener información útil para diferenciar imágenes con distinto patrón. Ésta idea es apoyada en la literatura referente a procesamiento de imágenes y precisamente en algunos trabajos referentes a clasificación de nubosidad [1], donde se propone el uso de FFT para obtener una descripción de la distribución de la potencia espectral de una imagen del cielo (utilizan la rutina 2D FFT disponible en MATLAB). Previo a la utilización de una operación FFT, se requiere un pre-procesamiento de imagen para normalizar niveles de contraste y otros aspectos, luego, la operación FFT obtiene las amplitudes complejas de los armónicos que se corresponden con cada *wavenumber* (en las 2 direcciones: vertical y horizontal). Un *wavenumber* o frecuencia espacial es un recíproco en lo espacial de lo que es la frecuencia en lo temporal, se define comúnmente como el número de ciclos por unidad de longitud dada una dirección y corresponde a la frecuencia dividida por la velocidad de onda, el módulo de estas amplitudes complejas es conocido como la función de energía espectral.

Normalizando esta función por el tamaño de la imagen N (tamaño en términos de elementos discretos: *picture elements* o pixeles que la componen) se obtiene la función de potencia espectral que será la base para análisis posteriores. La Figura 4.6 gráfica de mejor forma la correspondencia entre los *wavenumbers* y la distribución de ondas que puede identificarse en ambas direcciones para 2 imágenes con distinto patrón tras aplicar una rutina FFT.

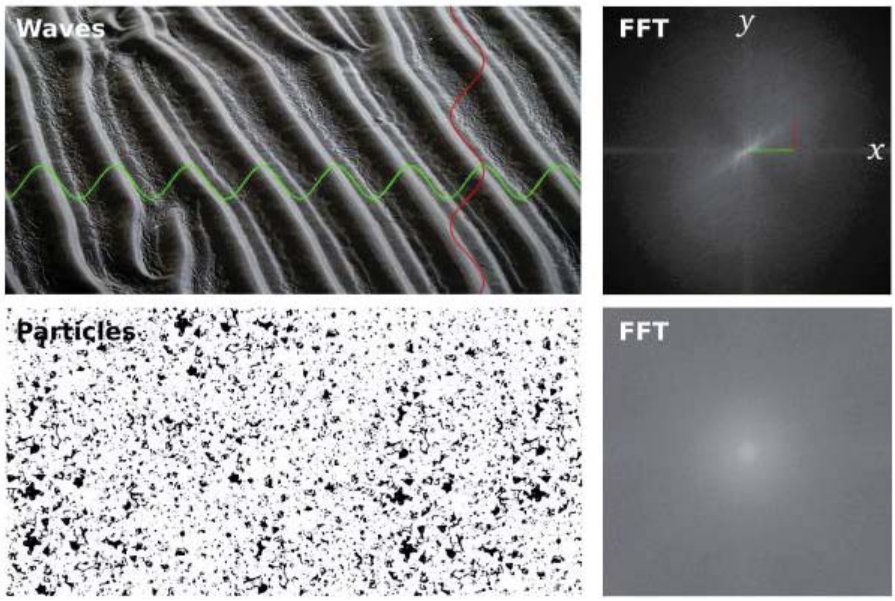


Figura 4.6: Función de potencia espectral (derecha) obtenida para 2 imágenes con distinto patrón (izquierda)

La función de potencia espectral es distinta para diferentes tipos de nubosidad, sin embargo es necesario extraer alguna característica simple que permita clasificar el tipo de nubes en la imagen. En [1] y basándose en estudios realizados en trabajos previos, proponen el uso de 2 características: *Correlation whit Clear* (CC) y la intensidad Espectral (SI).

4.5.2.1 Característica: Correlación con Despejado (CC)

La característica CC cuantifica la similitud entre la función de poder espectral de cualquier imagen de nubosidad respecto de la función de poder espectral de una imagen correspondiente a cielo despejado (*sky-clear)* tomada como referencia. En específico, el valor de CC es el coeficiente de correlación lineal entre los logaritmos de dos funciones de poder espectral y por lo tanto es un valor que puede ir desde 0 a 1, donde entre más cercano a 1 el valor de esta característica, más uniforme es el aspecto del cielo.

4.5.2.2 Característica: Intensidad Espectral (SI)

La intensidad espectral considera la distribución de la potencia espectral en un rango determinado de frecuencia espacial. Esto significa que dependiendo del patrón que describen las nubes

presentes en una imagen, habrá más o menos potencia espectral en frecuencias espaciales específicas. Para poder cuantificar este efecto en un único valor se puede seguir un procedimiento como se detalla a continuación: Primero se define la potencia espectral acumulada $E^*(k_1, k_2)$ entre 2 frecuencias espaciales k_1 y k_2 como:

$$E_*(k_1, k_2) = \sum_{k_x=k_1}^{k_2} \sum_{k_y=k_1}^{k_2} S(k_x,k_y)$$

Donde $S(k_x, k_y)$ es la función de potencia espectral y la dependencia sobre la frecuencia espacial en ambas direcciones se ha hecho explicita. Luego se define y se calcula la razón de potencia espectral R como:

$$R(k_2) = \frac{E_*(k_{\min},k_2)}{E_*(k_{\min},k_{\max})}$$

Donde $k_{min} = 1/ N$, donde N es el número de pixeles en la imagen y $k_{max} = ½$. Si se superponen los valores del logaritmo de R y de la longitud de onda λ se hace evidente la existencia de una relación lineal aproximada. La característica SI corresponde al valor absoluto de la pendiente de la recta de regresión correspondiente. La Figura 4.7 muestra una gráfica de la razón de potencia espectral versus la longitud de onda λ de 2 imágenes del cielo, una correspondiente a cielo despejado y la otra a cierta clase de nubosidad, donde puede notarse la diferencia en la pendiente para ambas imágenes.

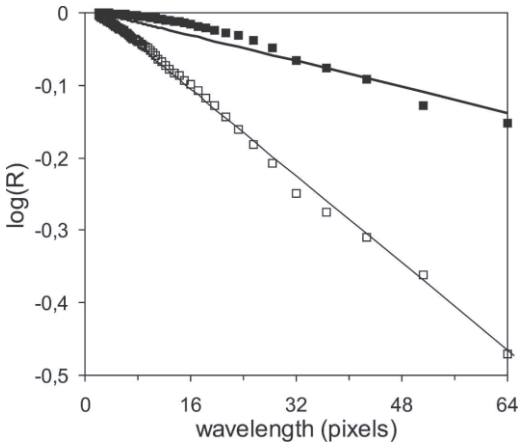


Figura 4.7: Gráfica de la intensidad espectral para 2 imágenes de nubosidad: cielo despejado y nubosidad dispersa

4.5.3 Segmentos

La división de una imagen en varios componentes o regiones con algún significado asociado, y la asignación a cada píxel de la imagen de una etiqueta correspondiente al componente al que pertenece, se denomina segmentación de imágenes. Una imagen segmentada es el resultado de aplicar alguna transformación que resalta la diferencia entre determinadas regiones de la imagen original en donde el criterio para definir la región se conoce como el umbral. Las técnicas para la determinación de este umbral son muy variadas en complejidad, siendo populares métodos de agrupamiento o *clustering.* Así, en una fotografía aérea podría definirse un determinado umbral para diferenciar las regiones terrestres de las cubiertas por agua, y es muy común también la segmentación para la detección de bordes (*edge detection*) en una imagen. En [1], los autores proponen que es recomendable utilizar, además de las características que consideran la totalidad de pixeles, características que generen una clara distinción entre pixeles correspondientes a nubosidad de los que corresponden a regiones de cielo despejado. En dicho trabajo, los autores utilizan una rutina del software MATLAB que "adivina" el umbral para definir las regiones basándose en el histograma de la imagen en escala de grises R/B.

En éste trabajo, para la segmentación de la imagen en regiones despejadas y regiones nubosas se ha utilizado el método de segmentación descrito en la sección 4.4.2 obteniéndose 2 subconjuntos de pixeles: pixeles Nubosos "Np" y pixeles despejados "Cp". Sobre el conjunto de pixeles nubosos es posible obtener algunas de las características espectrales y de textura descritas en las secciones 4.5.1 y 4.5.2. Mientras las características definidas en dichas secciones entregan una descripción global de la imagen incluyendo zonas despejadas y nubladas, las características sobre pixeles nublados permiten incorporar una descripción específica de la nubosidad presente en la imagen. La Tabla 4.5, resume las características obtenidas desde el conjunto de pixeles nublados Np. La siguiente sección explica características derivadas de la relación entre Np y Cp.

Tabla 4.5: Características de regiones nubosas

Clave	**Nombre**
CME	Nubosidad: Promedio
CSD	Nubosidad: Desviación estándar
CSM	Nubosidad: Fundido
CTM	Nubosidad: Tercer momento
CUF	Nubosidad: Uniformidad
CSI	Nubosidad: Intensidad espectral

4.5.3.1 Característica: Cobertura de Nubosidad (CN)

Tras aplicar el umbral para segmentación explicado en la sección 4.4.2 y hacerse explicitas las regiones de nubosidad y de cielo cubierto, se puede obtener la proporción de cielo cubierto en la imagen dividiendo el número de píxeles nublados Np por el número total de píxeles *N*.

4.5.3.2 Característica: Quebrantamiento de Nubosidad (CB: Cloud Brokenness)

Tras aplicar el umbral y conocer el número de pixeles "nubosos" Np y el número de pixeles "despejados" Cp se puede determinar el grado de fraccionamiento de las nubes dividiendo Np por Cp. Puesto que los bordes de una imagen totalmente nublada, se consideran los bordes de la región nublada, CB será siempre mayor a 4/*N*. Así, el quebrantamiento de la nubosidad CB será mínimo para imágenes de cielo totalmente cubierto por las nubes y será máximo para imágenes de nubosidad intermitente o fraccionada.

4.5.3.3 Característica: Espesor de Nube (CT: Cloud Tickness)

Conociendo los pixeles de la imagen que corresponden a nubosidad Np, es posible obtener una medida del espesor de las nubes calculando el promedio de valores de gris en la imagen en estos pixeles desde la imagen R/B. Es razonable pensar que entre más delgada la nube, mayor será la cantidad de azul en los pixeles y por lo tanto, valores pequeños de CT serán indicador de nubes delgadas.

4.6 Inclusión de Características Exógenas

Tradicionalmente un técnico observador que se dispone a registrar la descripción de la nubosidad para un momento sinóptico determinado tiene a su disposición un conjunto de variables meteorológicas tomadas por instrumentos de medición automáticos. Es común que datos como la temperatura del aire seco, la temperatura de rocío, la humedad relativa, la presión QNH y otros, ya estén disponibles en un formulario de ingreso de datos como el que se muestra en la Figura 4.8. Según especialistas de la dirección meteorológica, algunos de estos datos tienen una estrecha relación con el tipo de nubosidad presente en el momento de la medición y es normal que sean considerados para apoyar la tarea y complementar lo observado por el personal de las estaciones de monitoreo.

Tomando en consideración esta situación, se propone integrar al vector de características algunos datos exógenos a la imagen, que puedan complementar la información extraída desde ésta (complementar las características endógenas a la imagen digital), así por ejemplo, la temperatura del aire seco junto con la presión QNH, podrían ser un buen discriminador para identificar bancos de nubes altas y delgadas y diferenciarlas de condiciones sky-clear. Por otra parte, la temperatura

de rocío, la humedad relativa y la presión QFE podrían ayudar a distinguir bancos de nubes bajas de otras con desarrollo vertical, cuya base también se encuentra a poca altura.

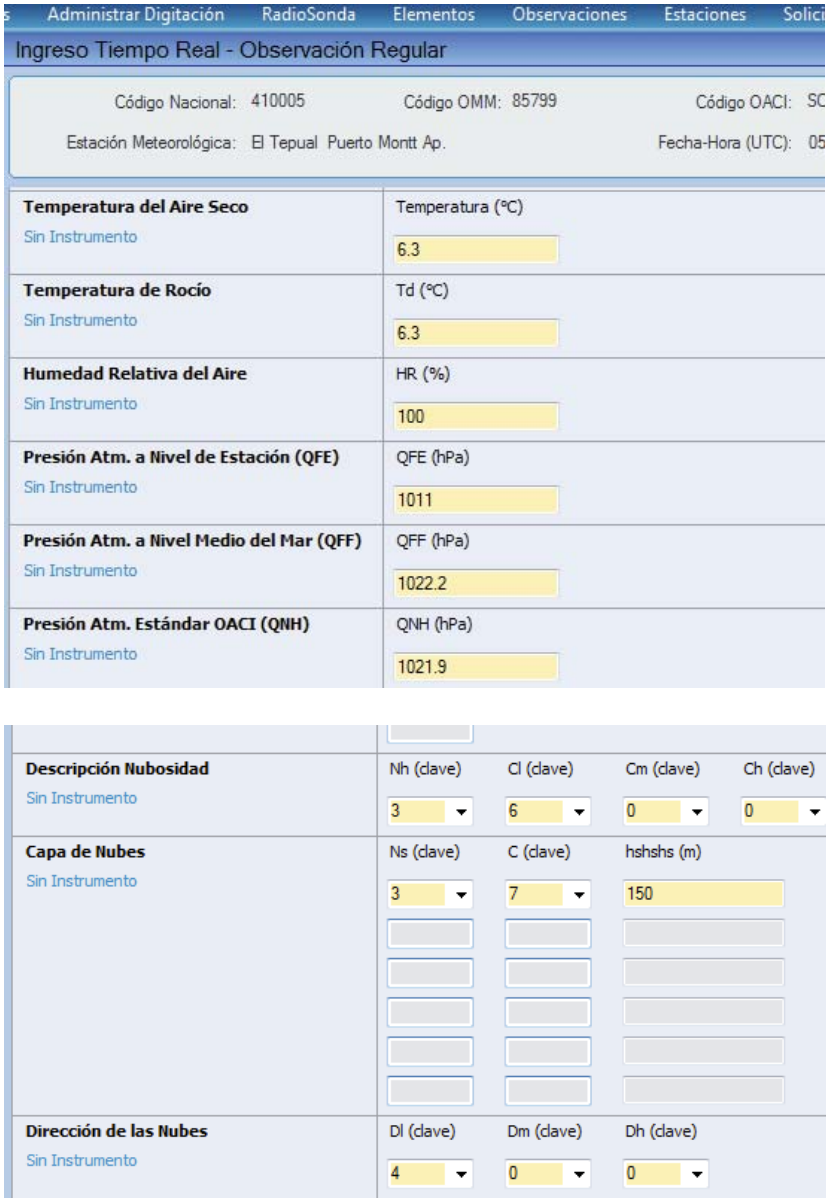


Figura 4.8: Formulario de ingreso de observación

Pese a que existe una gran cantidad de mediciones atmosféricas que se relacionan estrechamente con el tipo de nubosidad, la existencia de estos datos en el momento del ingreso y registro de observaciones depende de la disponibilidad de instrumentos de medición en la estación de monitoreo. A nivel nacional, la distribución de instrumentos de monitoreo es bastante heterogénea y esto se debe principalmente a dos razones: primero, la escasez de instrumentos de

medición para ciertos parámetros, en donde muchos de los instrumentos están disponibles sólo gracias a donaciones de organismos extranjeros, y segundo, a lo largo del territorio nacional se manifiestan diversas necesidades meteorológicas asociadas a la actividad de la zona, por lo que se pueden encontrar distintos grupos de instrumentos según la ubicación de la estación de monitoreo. Es así como por ejemplo, en el sur del país hay estaciones orientadas a la agro meteorología, con disponibilidad de instrumentos de medición para la depresión de rocío, la temperatura del bulbo húmedo y la razón de mezcla saturada, que no es posible encontrar en estaciones de monitoreo emplazadas desde la 4º región hacia el extremo norte.

Del mismo modo, instrumentos de medición especializados para obtener componentes relacionados con el índice de radiación (UV, Gm, etc.) son escasos y están ubicados principalmente en estaciones de monitoreo emplazadas en zonas urbanas o cercanas a balnearios.

Considerando el propósito del estudio, se ha definido como alcance para la inclusión de características exógenas a las imágenes tomadas por WSC aquellos datos meteorológicos considerados relevantes a la nubosidad, que tengan alguna relación con el fenómeno pero que al mismo tiempo su disponibilidad sea transversal a la gran mayoría de las estaciones de monitoreo del país. De esta forma existen datos que estarán siempre disponibles como lo son las mediciones sinópticas en horario principal, que obedecen a necesidades aeronáuticas. Aquí se encuentran algunas mediciones como por ejemplo: la temperatura del aire seco (Ts), la presión atmosférica corregida a nivel medio del mar mediante atmosfera estándar OACI (QNH), la temperatura de rocío (Td), la humedad relativa del aire (HR), la presión atmosférica corregida a nivel de elevación de la estación o aeródromo y la altura geopotencial de la superficie isobárica (G4ahhh).

Figura 4.9: Características endógenas y exógenas del vector

ME	SD	SM	TM	UF	EY	CC	SI	CME	CSD	CSM	CTM	CUF	CSI	CEY	CN	CB	CT	Ts	Td	HR	QNH	QFE	G4a

De esta forma el nuevo vector de características estará formado por una sección de características endógenas a la imagen y otra sección con conocimiento exógeno. En la sección endógena, estarán aquellas características que se desprenden exclusivamente de la imagen obtenida desde la WSC y sus propiedades, las cuales han sido descritas en la sección 4.5, en la sección exógena estarán los datos correspondientes a mediciones meteorológicas que acompañan la observación en la que se registra la descripción de la nubosidad, ambas secciones pueden verse en la Figura 4.9 y una descripción es incluida en la Tabla 4.6. El impacto de incluir estas características se discutirá en la sección 4.10.2.

Tabla 4.6: Componentes del vector de características

	componente	Clave	Nombre
endógenas	e01	ME	Promedio
	e02	SD	Desviación estándar
	e03	SM	Fundido
	e04	TM	Tercer momento
	e05	UF	Uniformidad
	e06	EY	Entropía
	e07	CC	Correlación con despejado
	e08	SI	Intensidad espectral
	e09	CME	Nubosidad: Promedio
	e10	CSD	Nubosidad: Desviación estándar
	e11	CSM	Nubosidad: Fundido
	e12	CTM	Nubosidad: Tercer momento
	e13	CUF	Nubosidad: Uniformidad
	e14	CSI	Nubosidad: Intensidad espectral
	e15	CEY	Entropía
	e16	CN	Cobertura nubosa
	e17	CB	Quebrantamiento de nubosidad
	e18	CT	Espesor de nubosidad
exógenas	x01	Ts	Temperatura del aire seco
	x02	Td	Temperatura de rocío
	x03	HR	Humedad relativa del aire
	x04	QNH	Presión Atmosférica estándar OACI
	x05	QFE	Presión Atmosférica a nivel de estación
	x06	G4a	Altura Geopotencial de Nivel Isobárico

4.7 Frameworks para Clasificación Automática

En consideración a los antecedentes entregados en las secciones anteriores respecto a la naturaleza del problema y las topologías utilizadas para la clasificación, se propone abordar la clasificación de las imágenes de nubosidad bajo alguno de los frameworks de etiqueta múltiple: el framework multi-etiqueta o el framework multi-instancia multi-etiqueta. En el framework multi-instancia multi-etiqueta, existe ambigüedad en el espacio de entrada (que elementos de la imagen corresponden a instancias) y ambigüedad en el espacio de salida (que clases o etiquetas corresponden a dichas instancias). En [8] se proponen 2 métodos para abordar este tipo de problemas, uno de ellos involucra un mapeo del problema multi-instancia multi-etiqueta a un framework de clasificación multi-etiqueta, lo que se explica con más detalle en la sección 4.7.2. Es importante entonces mencionar las características y los métodos de solución propuestos para la clasificación multi-etiqueta antes de abordar las particularidades del enfoque multi-instancia multi-etiqueta. Estas características deberán considerarse para decidir cuál de los dos frameworks se adapta de mejor forma al problema de clasificación que se pretende solucionar.

4.7.1 Clasificación Multi-Etiqueta

En la clasificación tradicional, un objeto es representado por una instancia (un vector de características). Estas instancias pertenecen al espacio de instancias *X*. Al mismo tiempo se tiene un espacio de categorías o etiquetas C. La tarea de clasificación automática se define como el aprendizaje de una función $f : X \rightarrow C$, es decir se toma una instancia de *X* y se le asigna alguna etiqueta de *C*. Como se vio en la sección 4.1, esta función se infiere desde un conjunto de instancias ya etiquetadas. *C*uando las instancias o ejemplos se asocian a una sola etiqueta $\lambda \in C$ como en la Figura 4.10, se habla de clasificación de datos de etiqueta simple (*single-label*), si de otra forma las instancias se asocian a un subconjunto de etiquetas $Y \subseteq C$, corresponde a una clasificación de datos multi-etiqueta (*multi-label*) como se grafíca en la Figura 4.11.

Figura 4.10: Esquema de un ejemplo de etiqueta-simple

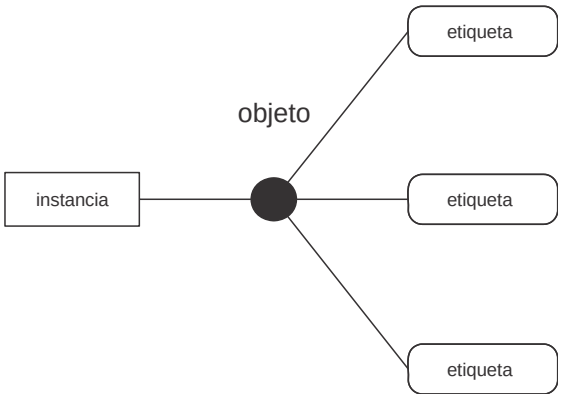


Figura 4.11: Esquema de un ejemplo multi-etiqueta

Diferentes métodos para la resolución de problemas bajo el framework multi-etiqueta han sido propuestos, y en [9] se entrega un buen resumen de las consideraciones a tomar respecto a la preparación de los datos, las estrategias para lidiar con los datos multi-etiquetados, la evaluación de clasificadores y la caracterización de las colecciones de datos multi-etiqueta. En [10] se establece que se han propuesto básicamente 2 tipos de estrategias para abordar la clasificación multi-etiqueta:

i. **métodos de transformación de problema:** En estas metodologías, el problema de clasificación multi-etiqueta es transformado en múltiples problemas de clasificación de etiqueta-simple manipulando el data-set de instancias. Esta transformación es independiente del algoritmo que produce la función f , referido comúnmente como '*learner'*.
ii. **métodos de adaptación de algoritmos:** Estos métodos extienden los algoritmos de aprendizaje, en orden de poder procesar los datos multi-etiquetados directamente por lo tanto son *'learner'*-dependientes.

4.7.2 Clasificación Multi-Instancia

Muchos problemas de clasificación no se adaptan lo suficiente a los frameworks de clasificación tradicionales (etiqueta-simple, multi-etiqueta) y esto es especialmente cierto en algunos problemas de reconocimiento de objetos y clasificación de imágenes [8]. Una imagen constituye el objeto del mundo real y puede estar asociada a múltiples instancias, por ejemplo, múltiples regiones o segmentos en la imagen, y simultáneamente estar asociada a una clase, como la imagen representada en la Figura 4.12, o a múltiples categorías como en la Figura 4.13. Cada imagen entonces es como un "saco de instancias" que se envía al clasificador.

La tarea de clasificación en este caso consiste en obtener una función $f: 2^X \rightarrow 2^C$ desde un set de datos: $\{(X_1,Y_1), (X_2,Y_2), \ldots, (X_m,Y_m)\}$ donde $X_i \subseteq X$ es un conjunto de instancias $\{x_1^i, x_2^i, \ldots, x_n^i\}$ e Y_i es un conjunto de etiquetas $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$.

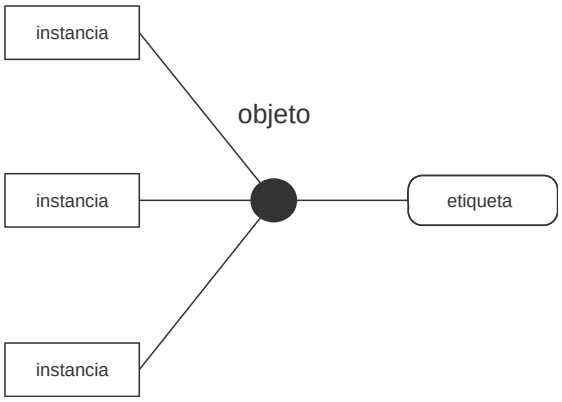


Figura 4.12: Esquema de un ejemplo multi-instancia

Para abordar este tipo de problemas se han propuesto varios enfoques [8], [11], los más recientes orientados a problemas con una gran número de etiquetas [12]. La mayoría de estas propuestas se relacionan con alguna de las siguientes estrategias:

I. Transformar cada "saco de instancias" en instancias simples y tras esto clasificar bajo el framework de clasificación multi-etiqueta.
II. Transformar el problema multi-instancia multi-label a uno multi-instancia, generando un clasificador multi-instancia etiqueta-simple para cada asociación (X_i,λ_l) encontrada en la colección.

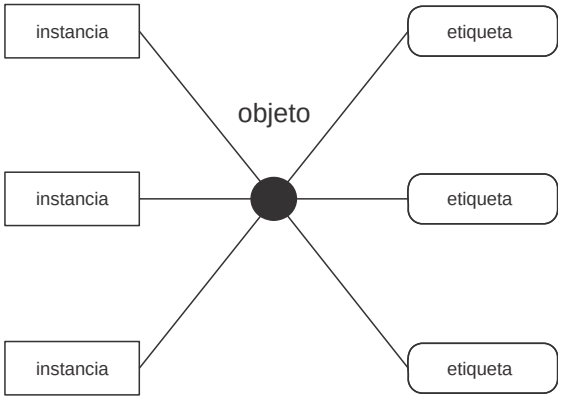


Figura 4.13: Esquema de un ejemplo multi-instancia multi-etiqueta

4.8 Medidas de Evaluación

Para precisar las medidas que se emplearan para evaluar el desempeño del clasificador de nubosidad es importante explicar algunas medidas ampliamente utilizadas en los frameworks de clasificación tradicionales (clasificación de etiqueta simple o single-class), puesto que en la mayoría de los trabajos existentes los resultados se presentan en base a estas medidas. Sin perjuicio de lo anterior, se obtendrán las medidas asociadas al framework de clasificación multi-etiqueta que son el *haming-loss* y algunas medidas *single-class* aplicables bajo el modelo de *micro-promedio.*

4.8.1 Medidas de Evaluación de Etiqueta simple

Estas medidas están orientadas al desempeño en tareas de clasificación donde la categoría asignada a cada instancia es una sola del conjunto de categorías disponibles. En este contexto para una imagen sólo existen 2 posibilidades cuando se trata de evaluar.

4.8.1.1 Precisión y Recuerdo

La precisión y recuerdo (*precision and recall*), son medidas comunes en el área de recuperación de información. La precisión es la probabilidad de que un elemento clasificado en la clase *c* corresponda realmente a esa clase. El recuerdo es la probabilidad de que un documento que pertenece a la clase *c* sea clasificado dentro de esa clase. Así, la precisión se puede ver como una medida de la corrección del sistema, mientras que el recuerdo da una medida de cobertura o completitud.

Para calcular estas medidas en un conjunto de prueba, se considera el problema de forma binaria. En este contexto, la Tabla 4.7 resume el comportamiento de un sistema según los casos de aciertos y errores, ordenamiento de resultados conocido como matriz de confusión de un predictor.

Tabla 4.7: Matriz de confusión

	Predicción positiva	Predicción negativa	Total de predicciones
Clase positiva	a	b	a+b
Clase negativa	c	d	b+c
	a+c	b+d	a+b+c+d = *n*

En la Tabla 4.7, cada celda representa el número de predicciones positivas y negativas. Así, *a + d* son los aciertos del sistema y c + b son los errores, y la suma de las cuatro celdas *(a + b + c + d)* equivale al número total de predicciones binarias. Los valores de esta tabla permiten estimar las medidas de precisión y recuerdo según las siguientes expresiones:

La precisión expresa en qué medida el clasificador toma una decisión correcta al ubicar cualquier documento en la clase que le corresponde.

$$precisión = \frac{a}{a+c}$$

El recuerdo refleja cuantos de todos los documentos de una clase son clasificados en ella.

$$recuerdo = \frac{a}{a+b}$$

4.8.1.2 Medida *F*

Describir el comportamiento de un clasificador de textos con dos medidas no es práctico para comparar sistemas. Para ello es común utilizar la medida F_β que se define como:

$$F_\beta = \frac{\left(1+\beta^2\right) precisión * recuerdo}{\beta^2 * precisión + recuerdo}$$

En la medida *F*, β es un parámetro que controla la importancia relativa entre la precisión y el recuerdo. Es común usar $\beta = 1$, conocido como la medida armónica de la precisión y el recuerdo que da igual importancia a ambas mediciones.

Otra medida que es empleada en algunas evaluaciones es la exactitud (*accuracy*), la cual entrega el porcentaje de predicciones correctas versus el total de predicciones realizadas.

$$exactitud = \frac{a+d}{a+b+c+d}$$

Cuando se requiere medir la efectividad en la clasificación de múltiples clases, se pueden adoptar 2 enfoques:

i. Micropromedio (Microaverage): consiste en calcular la efectividad considerando el conjunto completo de predicciones n como un sólo grupo de muestras.
ii. Macropromedio (Macroaverage): consiste en calcular un promedio de efectividad considerando cada clase como un grupo de muestra distinto.

4.8.2 Medidas de Evaluación Multi-etiqueta

A la hora de evaluar las predicciones de un clasificador del tipo bipartición, es decir predicciones el tipo verdadero o falso para cada etiqueta, se utilizan 2 tipos de medidas:

i. Basadas en ejemplos (example based)
ii. Basadas en etiqueta (label based)

Las primeras evalúan el rendimiento de clasificación a lo largo de todo el conjunto de ejemplos para finalmente obtener un promedio de las evaluaciones, las segundas primero obtienen las evaluaciones de clasificación por cada categoría, estas evaluaciones puede ser cualquiera de las utilizadas en la clasificación single-label (precisión, recuerdo etc.).

4.8.2.1 Hamming Loss

Es una medida basada en ejemplos y se define como el promedio de todas las diferencias simétricas entre el conjunto real de etiquetas asignadas a una instancia (etiquetado real) y el propuesto por el clasificador multi-etiqueta (predicción) para cada uno de los ejemplos. Si se considera *m* como el número de instancias, Y_i el conjunto de etiquetas reales de una instancia, Z_i el conjunto de etiquetas otorgadas a una instancia por un clasificador y M la suma $\#Z_i + \#Y_i$ se tiene:

$$hamming\ loss = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \Delta Z_i|}{M}$$

donde el operador Δ obtiene la diferencia simétrica entre 2 conjuntos, en este caso, los conjuntos Y_i y Z_i.

4.8.2.2 Sub-Set Accuracy

También corresponde a una medida basada en ejemplos y evalúa la razón de predicciones con total acierto, es decir, el porcentaje de predicciones donde el conjunto real de etiquetas y el resultado del clasificador son exactamente iguales. Se puede decir que esta evaluación es mucho más estricta que el *hamming loss* y es más apropiada para sistemas en que el impacto de clasificar mal al menos una de las etiquetas es alto.

4.9 Clasificación de Nubosidad bajo el Framework Multi-Etiqueta

Una revisión en profundidad sobre la naturaleza de las imágenes que se obtienen desde los instrumentos TSI permitió observar que existe una elevada complejidad a la hora de identificar fronteras que puedan separar instancias de uno u otro tipo de nubosidad en una misma imagen. Tras indagar con especialistas en la observación del fenómeno que realizan descripciones en campo sobre la estrategia empleada por los humanos, se llegó a la conclusión de que, en general, los aerosoles que se presentan en una imagen se pueden encontrar superpuestos o adyacentes y la determinación sobre la presencia o ausencia de cada género de nube se obtiene, en mayor medida, con la observación de condiciones generales en la imagen: patrones en las formas, textura, contraste y luminosidad.

En comparación al framework de clasificación multi-instancia, el framework multi-etiqueta evita la complejidad que conlleva el extraer instancias en base a criterios de agrupamiento y trata cada ejemplo (imagen) como una única instancia. En consideración a lo anterior, se tomó la decisión de implementar un clasificador bajo el framework multi-etiqueta.

Se instrumentó una clasificación de la colección completa considerando la topología de nubosidad del atlas OMM AIN según género la que se describió en la Tabla 4.1. Para esta topología, las imágenes de la colección corresponden a instancias multi-etiqueta donde una imagen puede tener asignada una o más clases de nube. Se utilizaron las características descritas en la sección 4.5 las cuales se obtuvieron principalmente con la herramienta IMMI (*IMaging MIning extension*), una extensión para el software de data-mining de Rapid-I. De este software se obtuvo un dataset el cual se convirtió al formato ARFF (csv2arff) de tipo *sparse-data* [13], formato utilizado por las librerías para data-mining de WEKA. Éste dataset se complementó con los metadatos requeridos por el framework para clasificación MULAN [14].

MULAN es una librería de clases que se construye como una extensión de las clases de WEKA y que incorpora interfaces y clases abstractas para implementar clasificadores capaces de manejar datos multi-etiquetados. Además, permite utilizar los clasificadores presentes en WEKA como algoritmos base para tareas de clasificación multi-etiqueta en base a métodos de transformación de problema. Como se vio en la sección 4.7.1, los métodos de transformación mapean un problema de clasificación multi-etiqueta en múltiples problemas de clasificación de etiqueta simple para los cuales es posible utilizar los clasificadores presentes en la suite WEKA.

Con el propósito de obtener información sobre las clases de nubes que podrían presentar un mayor grado de dificultad en su identificación respecto de otras clases, se implementó un clasificador en base a transformación binaria. Esta transformación convierte todos los ejemplos multi-etiqueta en ejemplos de etiqueta simple (clase y su complemento), siguiendo el ejemplo que se muestra en la Tabla 4.8.

Tabla 4.8: Transformación relevancia binaria

clases / img	Ci	Cc	St	Sc
1	X			X
2			X	X
3	X			
4		X	X	

	Ci	¬Ci
1	X	
2		X
3	X	
4		X

	Cc	¬Cc
1	X	
2	X	
3		X
4		X

	St	¬St
1	X	
2	X	
3		X
4		X

	Sc	¬Sc
1		X
2	X	
3		X
4	X	

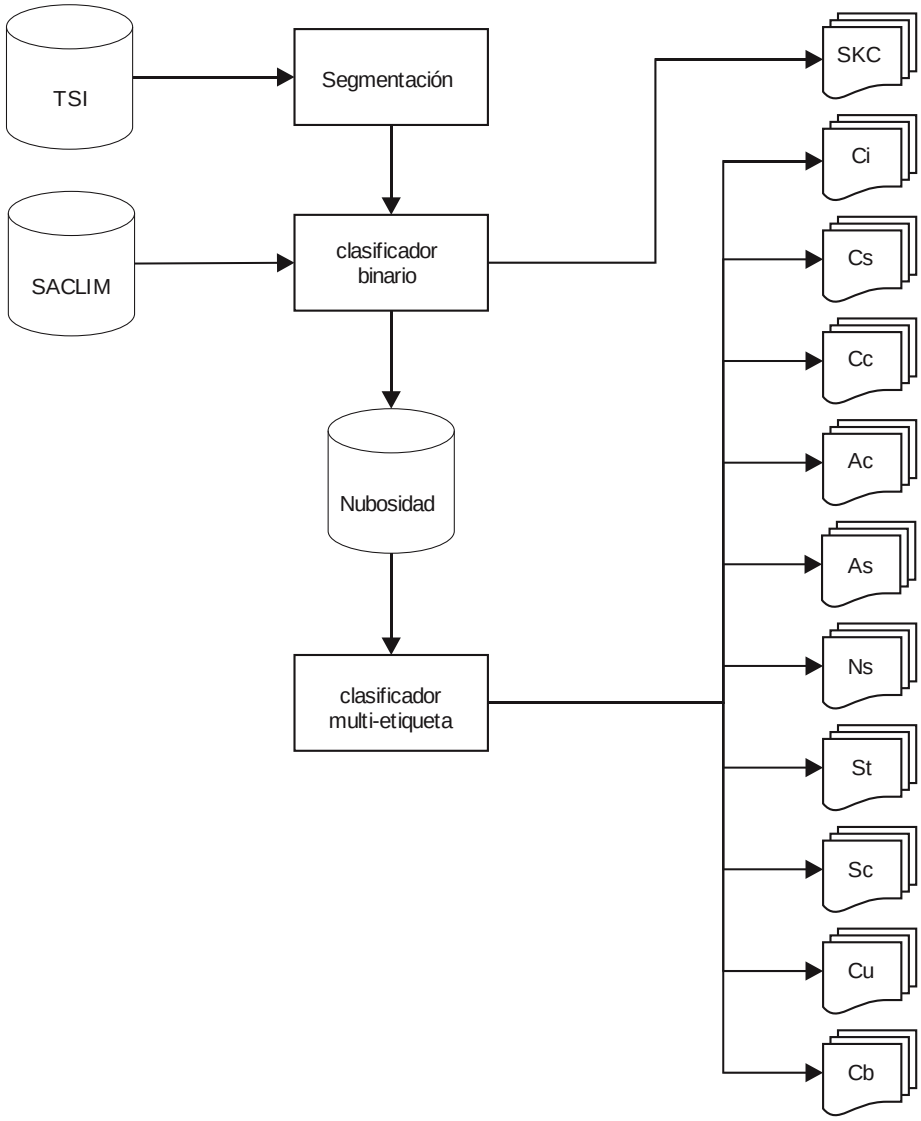
De esta forma se obtiene un dataset por cada una de las clases presentes en el problema con el cual se puede entrenar un clasificador binario que decide si el género está o no presente en la imagen. Luego para cada clase de nube, se pueden obtener algunas medidas de desempeño de etiqueta simple como las presentadas en la sección 4.8.1, este tipo de medición basada en etiqueta contrasta con las medidas de evaluación de desempeño basadas en ejemplos como el *hamming-loss* en las que se evalúan las diferencias entre el conjunto de nubes reales respecto del asignado a lo largo de todos los ejemplos de la colección.

El proceso de clasificación etiqueta las imágenes con los géneros de nube presentes y califica como sky-clear (SKC) aquellas imágenes sin presencia de nubosidad. Como puede verse en el esquema de la Figura 4.14, se incluyen tareas de pre-procesamiento para eliminar los pixeles que no deben ser considerados en la decisión, así como la obtención de los segmentos de nubosidad y cielo despejado según el criterio de segmentación descrito en la sección 4.4.2. En este punto se implementa un clasificador binario para la separación preliminar de imágenes correspondientes a condición sky-clear de aquellas que presentan algún grado de nubosidad, utilizando un umbral de 2% en la característica de cobertura nubosa (CN < 0.02). A continuación se obtienen las características espectrales y texturales de la imagen y se incorporan las mediciones meteorológicas en la forma de características exógenas. Es importante aclarar que en el contexto de una clasificación bajo el framework multi-etiqueta, una imagen con la condición sky-clear corresponderá a una que no presente ninguno de los géneros de nubosidad del atlas.

Una codificación alternativa que se evalúo, incluía la condición sky-clear como una etiqueta especial de nubosidad que puede aparecer junto con regiones de cielo nublado y en cuyo caso una imagen con total cobertura, correspondería a un ejemplo negativo para dicha clase. Esta

codificación se descartó ya que desde el punto de vista meteorológico es muy confuso hablar de una etiqueta sky-clear como un clase que puede estar superpuesta con otros géneros de nubosidad y habitualmente se utiliza para describir una 'condición' especial en la que no existen nubes.

Figura 4.14: Proceso de clasificación de géneros y condición Sky-Clear



Finalmente, sobre las imágenes que presentan nubosidad se implementa un clasificador multi-etiqueta bajo la transformación de relevancia binaria. La unión de las decisiones del clasificador multi-etiqueta sobre nubosidad y del clasificador binario para sky-clear entrega la categorización final.

En base a los buenos resultados obtenidos en otros trabajos de clasificación de imágenes [15], se utilizó como algoritmo base el clasificador kNN, este calcula para cada nueva imagen que se va a clasificar, la similitud de su vector representativo respecto a cada uno de los vectores representativos de las imágenes que conforman la colección de entrenamiento. El cálculo de la distancia entre vectores se realiza en base al método *Linear Nearest Neighbor Search* descrito en [16]. Luego, para decidir la clase, el algoritmo escoge la clase más ocurrente dentro del conjunto de los k vectores más similares. El parámetro k se configuro en 8, valor con el que se obtuvieron los mejores resultados.

Las medidas de evaluación basadas en categoría obtenidas para los distintos tipos de nubes se muestran en la Tabla 4.9 y el gráfico de la Figura 4.15.

Tabla 4.9: Medida *F* armónica obtenida para los géneros de nubosidad

	Ci	Cs	Cc	Ac	As	Ns	St	Sc	Cu	Cb	sky-clear
Medida F (B=1)	0,742	0,621	0,41	0,66	0,826	0,652	0,785	0,844	0,923	0,523	0,982

4.10 Evaluación del Clasificador Implementado

4.10.1 Clasificación con Características Endógenas

El *hamming-loss* obtenido para la prueba de clasificación sobre la representación que no incorpora características exógenas a la imagen digital, fue de 0,0942 que no es un mal resultado si se considera que se trata de una prueba preliminar de clasificación y que la colección de datos aún está siendo validada. No ocurre lo mismo con el *sub-set accuracy* que no supero el 60% (0.562), porcentaje moderado que puede explicarse principalmente por el número de aciertos en la categorización que tuvo la condición *sky-clear*. En este caso, la condición *sky-clear* está dada por la no presencia de nubosidad y, por lo tanto, se suman a esta categoría todas aquellas instancias cuya predicción fue negativa en todos los clasificadores binarios para cada clase de nubosidad. De alguna forma, el algoritmo puede identificar las características que definen la presencia o no presencia de nubosidad en una imagen y esto se puede explicar por las *features* de homogeneidad y textura, pero principalmente por las características correlación con despejado (CC) y cobertura nubosa (CN) explicadas en las secciones 4.5.2.1 y 4.5.3.1 respectivamente.

Existe un mayor grado de confusión a la hora de distinguir los distintos tipos de nubes y sus combinaciones, esto es especialmente cierto para tipos de nubes que tienen una relación de parentesco con alguno de los otros géneros descritos en la topología OMM AIN, por ejemplo los altocúmulos (Ac) y cirrocúmulos (Cc), los cirros (Ci) y los cirrostratus (Cs) estuvieron entre las clases con menor eficacia en la clasificación. Por otro lado en proporción al resto de las clases, los cumulonimbus tenían una menor cantidad de ejemplos disponibles para entrenar el

clasificador, sin embargo las propiedades visuales de estas nubes gruesas y oscuras al parecer compensaron dicha falencia de la colección.

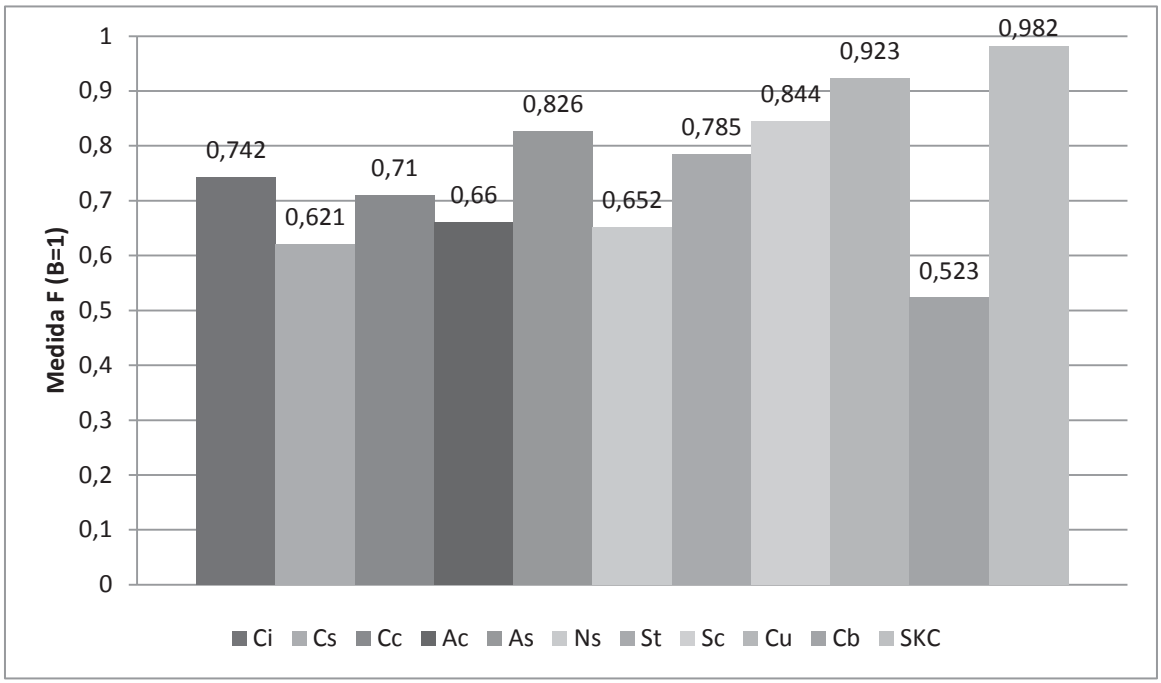


Figura 4.15: Desempeño en la clasificación de cada tipo de nube.

4.10.2 Impacto de las Características Exógenas

En la sección 4.6 se describió el proceso de incorporar variables externas a las propiedades de las imágenes obtenidas por instrumentos TSI (características exógenas), estas variables corresponden principalmente a mediciones meteorológicas que forman parte de la observación en la que está incluida la descripción de la nubosidad. La justificación para esto yace en la relación que pueden tener determinados valores de estas variables con el tipo de nubosidad presente en la atmosfera y por ende, en la práctica, aportan conocimiento útil para validar la descripción cualitativa registrada por un técnico observador. Los resultados de clasificar incorporando estas características se listan en la Tabla 4.10.

Tabla 4.10: Desempeño al incorporar características exógenas

	Ci	Cs	Cc	Ac	As	Ns	St	Sc	Cu	Cb	sky-clear
Medida F (B=1)	0,764	0,646	0,723	0,69	0,874	0,751	0,789	0,866	0,942	0,626	0,995

El gráfico de la Figura 4.16 muestra la variación porcentual del desempeño de clasificación para cada clase, cuando se utiliza el vector que incorpora estas características versus el vector que solo utiliza características endógenas a la imagen digital.

Se aprecia una mejora significativa en la clasificación de imágenes que contienen nubes Cb y Ns (cumulonimbus y nimbostratus), mejora que podría estar siendo influenciada fuertemente por la incorporación de mediciones meteorológicas relativas a la presión. Ambas categorías corresponden a nubes de tipo nimbo (lluvia) y por lo tanto presentan condiciones de presión distintivas respecto a otras categorías de nubes. Las nubes de desarrollo vertical (cumulonimbus) son difíciles de diferenciar de bancos de nubes bajas (estratos) cuando la observación se realiza bajo su base, especialmente para grandes formaciones de nubes de tormenta y en estos casos generalmente el observador se apoya en la información de observaciones anteriores, la tendencia barométrica, mediciones de temperatura y otras para realizar la descripción.

Extrapolando esta idea, es lógico pensar que la incorporación de otro tipo de datos externos a las imágenes WSC o TSI pueda ayudar a discriminar entre ciertos géneros de nubosidad visualmente similares. Por ejemplo, si fuese posible incorporar la clasificación de imágenes satelitales de radio espectrómetro en la forma de una característica del vector, posiblemente mejoraría la clasificación de nubosidad al incorporar datos como la temperatura superior de la nube, detectando por ejemplo las nubes de desarrollo vertical.

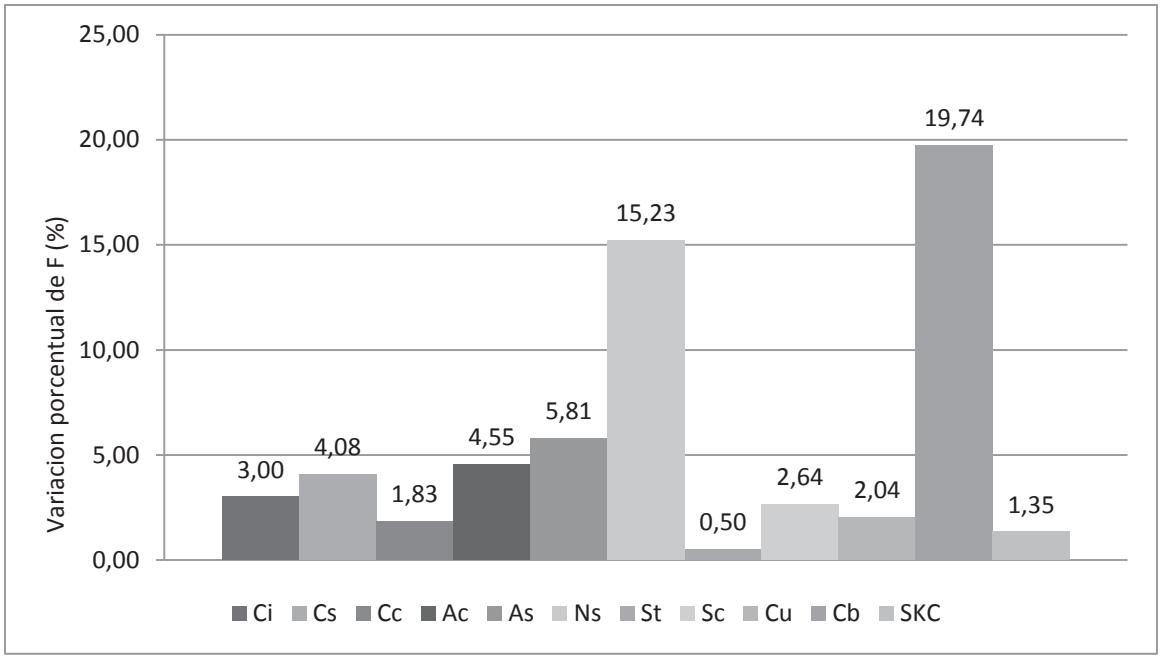


Figura 4.16: Mejora porcentual de la clasificación incorporando características exógenas

4.10.3 Resultados en Otros Trabajos

A diferencia de clasificar la nubosidad en base a la condición nubosa general, identificar el género de las nubes presentes en la bóveda celeste como lo realizan en las estaciones meteorológicas de la DMC es particularmente difícil y no se conocen otros trabajos que propongan automatizar la descripción en base a dichas clases, al menos en la revisión bibliográfica realizada. En cambio se han propuesto clasificadores automáticos sobre otras colecciones de datos confeccionados en base a clases fenomenológicas las cuales en última instancia pueden resultar de una combinación de los géneros descritos en el atlas mundial de nubosidad, pese a esto, dichas combinaciones no son estrictas, permitiéndose en algunas clases la presencia o ausencia de algunos de los géneros [6].

Sumado a esto, dichos trabajos han empleado datasets los cuales no están disponibles, imposibilitando realizar una comparación directa de desempeño sobre los mismos datos. En algunos casos los autores han identificado la procedencia de las imágenes pero no se han mencionado con detalle los criterios de asignación de sus clases, ni el período de la observación subyacente.

Cabe mencionar que especialistas de la DMC han identificado cierta similitud en las topologías de algunos de estos trabajos [1] con formularios para el ingreso de la condición nubosa especializados para la actividad aeronáutica que se ingresan en estaciones ubicadas en aeródromos. La información de dichos formularios debe ser consistente con la ingresada con fines meteorológicos y climatológicos y se almacenan de igual forma en el banco de datos nacional, no obstante, su uso no es tan amplio como los formularios MET empleados en estaciones sinópticas, pluviométricas, etc.

Se evaluó la posibilidad de confeccionar un dataset utilizando la colección validada agrupando los géneros de forma que las clases resultantes se correspondan con las empleadas en algunos de estos trabajos, sin embargo dada la incertidumbre sobre el criterio de agrupamiento para algunos casos donde la presencia o ausencia de un género es opcional, no es posible realizar un mapeo completo de la colección. Una forma de enfrentar este problema sería seleccionar sólo aquellos ejemplos que pueden ser mapeados directamente a la clase equivalente, sin embargo el hecho de estar dejando fuera parte de los ejemplos para los cuales más tarde se requerirá una predicción hace que dicha evaluación se aleje del propósito original del clasificador automático.

5 Detección de Errores en Observación

El desempeño obtenido mediante la clasificación automática descrita en las secciones anteriores puede considerarse bastante bueno en términos generales, obteniéndose una medida-F armónica sobre 0.75 en 7 de las 10 clases del Atlas de nubosidad, y en 4 de estas F se situó por encima de 0.85. No obstante, este desempeño no satisface la necesidad de exactitud que se precisa para utilizar estos datos con cierto grado de confianza en modelos climatológicos y aplicaciones de investigación. En este trabajo se propuso utilizar dicha clasificación para mejorar los criterios de control sobre la descripción realizada por técnicos observadores, contrastando la descripción entregada por el clasificador y la realizada por el humano.

Si se considera la observación y descripción de la nubosidad como un proceso, la descripción humana y la descripción automática descrita son 2 implementaciones distintas de dicho proceso. Mientras ambas implementaciones son propensas al error, la naturaleza de estos errores es muy distinta pues se derivan principalmente de la naturaleza de su implementación. Así, mientras los errores encontrados en la clasificación automática pueden derivar de similitudes o tendencias estadísticas en ciertos atributos del vector, coocurrencia fortuita de alguna propiedad a lo largo de las clases [17] o simplemente errores en las instancias de la colección utilizada para entrenar el clasificador, los errores introducidos por un observador humano pueden incluir razones de un origen totalmente ajeno a las propiedades del fenómeno, por ejemplo: omisiones injustificadas, intencionalidad, falta de probidad o desconocimiento.

En el contexto de identificación de errores, la capacidad de discriminación de un clasificador automático no se puede comparar a la del clasificador humano en términos generales si no que debe ser estudiada considerando el error al que es propenso y como se relaciona este error con las observaciones erradas de los especialistas.

5.1 Criterio para la Detección de Errores

Con el objeto de obtener información respecto a lo señalado previamente, la DMC confeccionó una colección de 125 imágenes de nubosidad para las cuales existía la sospecha de errores en la observación registrada en el banco de observaciones del sistema SACLIM, y se utilizó el clasificador entrenado según la configuración descrita en la sección 4.9 para contrastar las decisiones de éste con las tomadas por el técnico observador que realizó la descripción. En resumen se define un proceso de control de calidad que incluye las siguientes actividades:

i. Identificar todas las instancias para las que el clasificador entregó una descripción distinta a la ingresada por el técnico observador.
ii. Someter las instancias identificadas a inspección visual e identificar cuáles de estas diferencias corresponden a errores de observación por parte del técnico observador.

Además, con el fin de validar los errores detectados, identificar los errores no detectados mediante el proceso y al mismo tiempo describir la naturaleza de los mismos se solicitó una inspección para la totalidad de las imágenes del set, 125 en total realizando el siguiente procedimiento:

i. Someter las instancias a inspección visual por 2 especialistas donde cada uno entrega la descripción de manera independiente.
ii. Comparar la decisión de ambos especialistas para identificar problemas de inconsistencia inter-indizador (*inter-indexer inconsistency*) [18].
iii. Comparar la decisión de los especialistas a la del técnico observador que realizó la descripción errada y determinar si es posible explicar la causa del error en términos de las propiedades endógenas de la condición nubosa del momento, por ejemplo: dificultad para distinguir algún género de otro visualmente similar.

La naturaleza del error se definió en base a un conjunto de atributos los cuales pueden ser obtenidos tras el procedimiento de inspección descrito anteriormente, estos atributos permiten diferenciar el tipo de error cometido por el clasificador y por el humano y se describen en la Tabla 5.1.

Tabla 5.1: Atributos de los errores identificados

atributo	**valor**	**descripción**
Dificultad	Si	Ambos expertos difieren en la asignación inicial de clase.
	No	La decisión inicial de ambos expertos coincide y es considerada correcta.
Inconsistencia inter-indizador	Si	Si ambos expertos no llegan a acuerdo respecto a la asignación correcta de la clase.
	No	Uno de los expertos reconoce correcto el criterio del otro y ajusta su decisión. En el caso de que ambos especialistas decidan cambiar su decisión inicial, se considerará inconsistencia y se debe indicar SI.
Error explicado	Si	El error se considera explicado si se constata dificultad para identificar la nubosidad desde la imagen. Por ejemplo se ha confundido el género con otro similar, o se ha omitido la nube por tener una presencia despreciable.
	No	El error se considera no explicado cuando no es posible identificar una dificultad endógena a la imagen capaz de inducir al error del observador, es decir, es un error evidente y fácil de detectar vía inspección visual.

5.2 Eficacia de la Detección de Errores

Tras aplicar el procedimiento descrito en la sección anterior, fue posible identificar gran parte de los errores en la descripción de nubosidad ingresada al sistema SACLIM por parte de técnicos observadores. La Tabla 5.2 entrega un resumen respecto a las diferencias entre la descripción realizada automáticamente y la registrada vía observación directa. En la primera fila, se entrega el conteo de imágenes de nubosidades para las que el clasificador entregó una asignación distinta a la registrada manualmente. Para estas diferencias, en la columna "descripción observador" se presentan cuales imágenes resultaron tener errores y cuáles no, logrando observar que del total de imágenes controladas en base a este criterio (31), 24 correspondían a errores reales (74% aproximadamente) y sólo un 26% implicó revisar una imagen que no contenía errores. En este caso, las 7 imágenes sin errores conformarían los falsos positivos del criterio de control y significan el "overhead" del esfuerzo invertido en mejorar la calidad del dato.

Tabla 5.2: Decisión clasificador versus observador

Observador vs Clasificador		Descripción observador		Error observador		Clasificación Automática	
Desacuerdo	31	errores observador	24	explicado	3	sin errores	17
				No explicado	21	con errores	7
		No errores	7				
Acuerdo	94	errores observador	2	explicado	2		
				No explicado	0		
		No errores	92				

Es importante señalar en este punto, que la asignación entregada por el clasificador automático no necesariamente es la correcta, encontrándose en los errores reales detectados 7 descripciones automáticas que también presentaban errores en la asignación de alguna de las clases (29%) y que por lo tanto no pueden ser utilizadas en reemplazo de lo ingresado por el observador, sino que deben ser corregidas por otro especialista. Otro aspecto que se evidencia, es que la naturaleza de los errores humanos detectados mediante el procedimiento es mayoritariamente ajena a los atributos de la imagen (las propiedades del fenómeno observado), siendo 21 de los 24 errores del tipo No explicado.

Por otra parte, la capacidad de detección de errores en base al procedimiento se debe medir considerando además el número de errores no detectados, que en este caso corresponde a aquellas observaciones en donde ambas clasificaciones coincidieron y estaban erradas, es decir, el clasificador automático comete el mismo error que el observador. En esta colección de 125 imágenes, sólo 2 quedaron en esta categoría, lo que corresponde a un 7% del total de errores reales presentes y que pueden explicarse por una dificultad visual en la distinción de un género (en este caso particular, se confunde un género con otro).

Tabla 5.3: Errores detectados y sus atributos

		Observador Humano		**Clasificador Automático**	
descripción	Imágenes	125	**%**	125	**%**
	Con error	26	20,8%	31	24,8%
	Sin error	99	79,2%	94	75,2%
asignaciones (10 * imagen)	Total	1250	**%**	1250	**%**
	Correctas	1138	91,04%	1041	83,28%
	Erradas	112	8,96%	209	16,72%
errores	Dificultad	8	**7,14%**	36	**17,22%**
	Inconsistencia	3	**2,68%**	16	**7,66%**
	Explicado	25	**22,32%**	195	**93,30%**
	No explicado	87	**77,68%**	14	**6,70%**

Es interesante observar además en la Tabla 5.3, que tras inspeccionar la totalidad de la colección y clasificar sus errores, queda en evidencia que la naturaleza del error humano en estas observaciones es en gran medida no explicada, es decir no obedecen a una dificultad intrínseca del proceso de reconocimiento y descripción sino más bien a problemas ajenos a éste, como los descritos previamente, mientras que el error del clasificador automático puede ser mayoritariamente explicado por dificultades visuales, encontrándose incluso un gran número de imágenes que presentan inconsistencia inter-indizador, lo que puede estar hablando de un problema en la misma topología empleada para clasificar.

Otro resultado inesperado es que en suma, en esta colección y a nivel de instancia (imagen u observación), el porcentaje de asignaciones correctas y erradas son similares, entre el clasificador

automático y la observación directa, lo que habla de los problemas asociados al proceso, sin perjuicio de que la colección se confeccionó precisamente por la sospecha justificada sobre la presencia de errores en la observación directa.

Finalmente se puede medir el desempeño del clasificador en la detección de errores de observación estableciendo 2 clases: observación consistente y observación inconsistente, en donde se consideran como aciertos, las descripciones humanas erróneas detectadas en base al procedimiento descrito, y como errores de detección, tanto los falsos positivos (diferencias en clasificación que no corresponden a un error) como las omisiones (clasificador comete el mismo error que el observador humano).

Utilizando los datos obtenidos se determinó una matriz de confusión en base a la cual se calcularon las medidas de desempeño descritas en la sección 4.8.1, las que se muestran en la Tabla 5.4. Se puede observar un recuerdo de 0.923 lo que habla de una buena capacidad para anunciar los errores presentes en la colección, que en definitiva es el objetivo del proceso de control de calidad. La precisión es menor (0.774), lo que se traduce en la inspección de imágenes que no contienen errores, sin embargo para este caso en particular, el interés en evitar revisar imágenes sin error (falsos positivos) es mucho menor que el interés de identificar la mayor cantidad posible de errores.

Tabla 5.4: Desempeño de la detección de errores

		Detección de error		
		Error	No error	Σ
Observación	Observación incorrecta	24	2	26
	Observación correcta	7	92	99
	Σ	31	94	125

precisión	0,774
recuerdo	**0,923**
F	0,842
accuracy	**0,928**

6 Conclusiones

6.1 Nuevo Data-Set para Clasificación Automática

Se ha confeccionado una colección de imágenes de nubosidad con su correspondiente clasificación multi-etiqueta sobre los géneros del atlas mundial, extrayendo observaciones desde el sistema de administración del banco de datos climatológicos SACLIM y corrigiendo vía inspección visual un gran número de errores.

Gran parte de la verificación, realizada por especialistas de la Dirección Meteorológica de Chile, detectó un número importante de errores en la descripción realizada por técnicos observadores, lo que confirma la necesidad de estudiar la efectividad del aprendizaje automático en la clasificación de imágenes que posee la DMC, pues un buen grado de efectividad es útil para la implementación de controles bajo este enfoque, que apoyen el aseguramiento de calidad de dichos datos.

Esta colección de imágenes podrá seguir siendo mejorada y enriquecida, siguiendo el método establecido en este trabajo, pasando a ser parte del patrimonio informático de la organización. El método desarrollado, involucra tareas para detectar y describir problemas inter-indizador y de esta forma, facilitar el desarrollo y evaluación de procesos que se sustenten con datasets derivados de la colección.

6.2 Reconocimiento de Nubosidad en Imágenes TSI

La revisión bibliográfica ha permitido identificar diferentes propuestas para la clasificación de la nubosidad, incluidos algunos trabajos que abordan el problema con el enfoque del aprendizaje automático (machine learning). Sin embargo, ninguno de los trabajos ha enfrentado la topología entregada por el atlas mundial de nubosidad de la OMM, que corresponde a la clasificación asignada en las estaciones de propósito meteorológico y climatológico, a diferencia de las observaciones de propósito aeronáutico.

En éste trabajo, se especificó un clasificador automático bajo la combinación del framework multi-etiqueta y el de etiqueta simple el cual fue entrenado con un data-set generado a partir de una colección de imágenes y métodos de representación basados en características del histograma de grises, textura e intensidad espectral, además de una etapa de segmentación para identificar regiones nubosas y obtener características específicas de las mismas. Para la predicción del conjunto de clases de nubosidad presentes en las imágenes, se obtuvo un haming-loss de 0,0942 que podría considerarse un buen resultado.

Se está evaluando la posibilidad de publicar la colección y/o el data-set generado para incentivar el desarrollo de otras técnicas o la aplicación de métodos ya existentes y así poder obtener comparaciones de desempeño consistentes.

Respecto a la obtención de características, se ha constatado una mejora importante de la clasificación al incorporar un conjunto de mediciones meteorológicas complementarias a la imagen digital, en la forma de características exógenas, alcanzando un aumento aproximado de 19% y 15% en la medida *F* para las clases stratus y cumulonimbus respectivamente, y promediando en todas las clases una mejora en *F* de 5.94% respecto a la clasificación en base sólo a características endógenas. Esto se explica por la incorporación de conocimiento relevante a la hora de discriminar entre tipos de nubosidad que en determinadas condiciones de filmación son visualmente similares.

6.3 Identificación de Errores en la Descripción de la Nubosidad

El desempeño del clasificador implementado en la entrega de una descripción de la nubosidad resultó ser desde moderado en la identificación de algunas especies a muy bueno en la identificación de otras, sin embargo se ha demostrado que al emplear la decisión del mismo para identificar posibles errores en la observación del fenómeno realizada por humanos se obtienen muy buenos resultados, alcanzando un recuerdo de 0.923 en la identificación de observaciones erróneas. Este buen grado de exhaustividad es independiente de los resultados al clasificar todos los géneros, debido a que la naturaleza de la mayoría de los errores cometidos en la observación directa no se relacionan con dificultades intrínsecas en la observación e identificación, sino que, como se ha evidenciado en una colección de validación obtenida desde el mismo banco de datos, se deben principalmente a factores no explicables y, por lo tanto, difícilmente el clasificador automático entregue una clasificación idéntica y errada.

6.4 Finalización y Trabajo Futuro.

En este trabajo se ha demostrado que el aprendizaje automático (*machinne learning*), no sólo es útil en procesos que requieren una completa automatización de la tarea realizada por un experto, sino que también pueden ser una herramienta importante a la hora de implementar procesos de apoyo que no exigen una potencia predictiva cercana al 100%. En la metodología para la detección de errores propuesta en este trabajo, los errores más probables son del tipo falso positivo, que en la práctica se traducirían en la inspección manual de una imagen para verificar una descripción que se realizó correctamente. Este escenario sigue siendo mejor al escenario actual, en el cual para lograr la mayor exhaustividad en el control de calidad, se deberían inspeccionar la totalidad de las descripciones realizadas a lo largo de todo el país.

También es importante mencionar que, dependiendo del problema, algunas de las típicas consideraciones en la implementación de estas técnicas, como por ejemplo la complejidad computacional, pierden importancia considerando el beneficio otorgado por la entrega de una predicción. En el caso de este trabajo, se utilizó la implementación de un algoritmo de la familia kNN, el cual es asociado habitualmente con un alto costo computacional, ya que en su fase de entrenamiento almacena en memoria todo el conjunto de vectores que representan las instancias de ejemplo. Esto, en la fase de desarrollo de prototipos, se tradujo en la utilización de un máximo de ~42 MegaBytes de memoria por uno de los prototipos desarrollados, al entrenarlo con la

colección inicial de imágenes, la cual contenía el mayor número de ejemplos utilizado durante toda la fase de evaluación. Además, mientras las imágenes que estarían involucradas en la implementación de un flujo de control se originan con una frecuencia máxima de 2 minutos, el tiempo tomado por el clasificador para la predicción de clases en una instancia sin clasificar siempre fue de una fracción de segundo al correr predicciones en un equipo portátil.

Sin llegar a cuantificar el costo versus el beneficio, es evidente que el costo recién descrito es totalmente irrelevante en este caso particular de aplicación, ya que el beneficio que conlleva el disponer de una sugerencia automática sobre la presencia de errores, para el personal responsable del aseguramiento de la calidad en las observaciones es mucho mayor a la problemática técnica de implementar un proceso que efectúe el procesamiento necesario.

Si bien el desempeño en la categorización automática de clases de nubosidad no requiere una potencia predictiva excepcional para ser útil en la búsqueda de errores humanos, una mayor exactitud siempre traerá beneficios. Durante el desarrollo de este trabajo se dejaron notar etapas en el proceso de aprendizaje automático sobre las que se podría seguir investigando y un ejemplo de esto es la obtención de características endógenas a la imagen digital. Para el dataset implementado, estas características se escogieron fundamentalmente en base a conocimiento experto sobre el dominio del problema (nubosidad), sin embargo, el espectro de técnicas que pueden ayudar a describir cuantitativamente las propiedades de una imagen es muy alto y esto se refleja en la abundante literatura que es posible encontrar, abordando específicamente éste tema.

Para finalizar, se debe indicar que con el trabajo realizado se han entregado antecedentes suficientes para justificar la implementación de técnicas de aprendizaje automático (*machine learning*) en apoyo a las tareas de control de calidad de observaciones meteorológicas, despertando interés en los especialistas de la Dirección Meteorológica de Chile que han contribuido a este trabajo. Por lo anterior, no se descartan futuras investigaciones que busquen mejorar la implementación propuesta, proponer soluciones distintas bajo la línea del aprendizaje automático, o la utilización de estas técnicas en otras problemáticas de estudio meteorológico similares.

7 Referencias

[1] Josep Calbó and Jeff Sabburg, "Feature Extraction from Whole-Sky Ground-Based Images for Cloud-Type Recognition," *Journal of Atmospheric and Oceanic Technology*, vol. 25, no. 1, pp. 3-14, Jan. 2008.

[2] Yu Liu, Jun Xia, Chun-Xiang Shi, and Yang Hong, "An Improved Cloud Classification Algorithm for China's FY-2C Multi-Channel Images Using Artificial Neural Network," *Open Access: Sensors*, vol. 9, pp. 5558-5579, 2009.

[3] Yoonkyung Lee, Grace Wahba, and Steven A. Ackerman, "Cloud Classification of Satellite Radiance Data by Multicategory SVM," *Journal of Atmospheric and Oceanic Technology*, vol. 21, no. 2, pp. 159-169, Feb. 2004.

[4] Heinle. A., Macke. A., and A Srivastav, "Automatic cloud classification of whole sky images," *Atmospheric Measurement Techniques*, vol. 3, no. 3, pp. 557-567, 2010.

[5] WMO, *International Cloud Atlas, Vol 2, Am. Meteorol. Soc.*, 1987.

[6] C. N. Long, J. Sabburg, J. Calbó, and D. Pagès, "Retrieving cloud characteristics from ground-based daytime colorall-sky images.," *Journal of Atmospheric and Oceanic Technology*, vol. 23, no. 5, pp. 633-652, 2006.

[7] Gonzales, Woods, and Eddins, *Digital Image Processing Using MATLAB, 2nd edition*.: Gatesmark Publishing, 2009.

[8] Zhi-Hua Zhou and Min-Ling Zhang, "Multi-Instance Multi-Label Learning with Application to Scene Classification," in *Advances in Neural Information Processing Systems 19*, Vancouver, Canada, 2007.

[9] Grigorios Tsoumakas, Katakis Ioannis, and Vlahav Ioannis, "Mining Multi-label Data," in *Data mining and knowledge discovery handbook*.: Springer US, 2010, pp. 667-685.

[10] Grigorios Tsoumakas and Ioannis Katakis, "Multi-Label Classification: An Overview," in *International Journal of Data Warehousing and Mining*, 2007.

[11] Zheng-Jun Zha et al., "Joint multi-label multi-instance learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[12] Oksana Yakhnenko and Vasant Honavar, "Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies," in *Proceedings of the British Machine Vision Conference*, 2011.

[13] University of Waikato. (2010) WEKA Knowledge Base.

[14] E. Spyromitros-Xioufis, K. Sechidis, G. Tsoumakas, and I. Vlahavas, "Mulan: A Java Library for Multi-Label Learning," *Journal of Machine Learning Research*, vol. 12, pp. 2411-2414, 2011.

[15] Zhi-Hua Zhou and Min-Ling Zhang, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038-2048, 2007.

[16] Aha W., D. Kibler, and Albert K., "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37-66, 1991.

[17] Imran Sarwar Bajwa and Syed Irfan Hyder, "PCA based Image Classification of Single-layered Cloud Types," in *Proceedings of the IEEE Symposium on Emerging Technologies*, 2005.

[18] Fabricio Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.