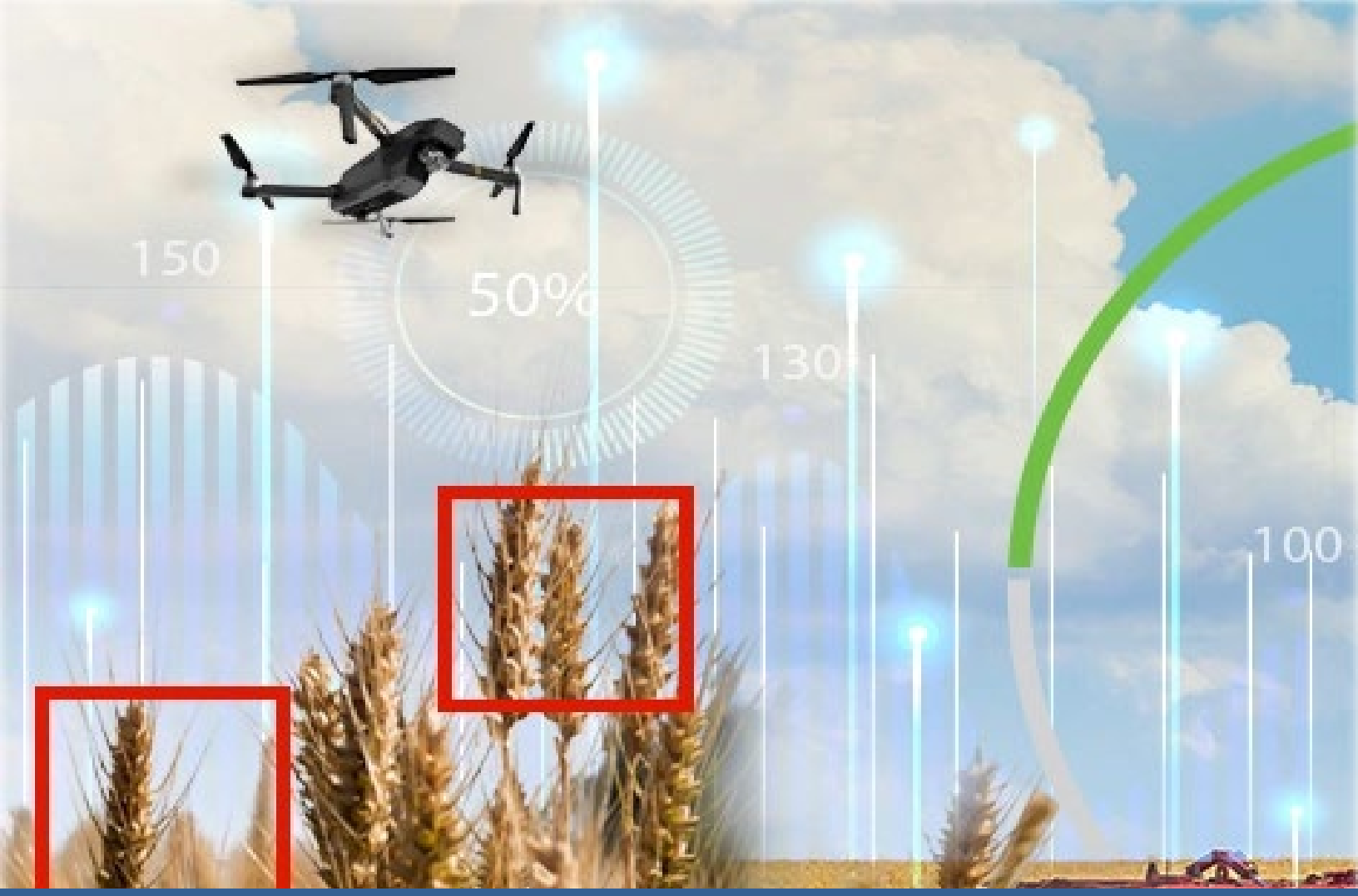

Estudio: Inteligencia artificial para la estimación de superficie de trigo a nivel nacional

ODEPA
Ministerio de Agricultura
Gobierno de Chile

www.odepa.gob.cl

Oficina de Estudios y Políticas Agrarias

Estudio: Inteligencia artificial para la estimación de superficie de trigo a nivel nacional

Noviembre 2023

Publicación de la Oficina de Estudios y Políticas Agrarias del Ministerio de Agricultura, Gobierno de Chile

Artículo producido y editado por la Oficina de Estudios y Políticas Agrarias – Odepa. Ministerio de Agricultura

Directora Nacional y Representante Legal: Andrea García Lizama

Informaciones:

Teatinos #40, piso 7, Santiago Chile.

Casilla 13.320 – correo 21

Código postal 8340700

Teléfono: 800 630 990

www.odepa.gob.cl e-mail: odepa@odepa.gob.cl

Contenido

Resumen ejecutivo

Disponer de información actualizada sobre las estadísticas de producción agrícola es fundamental para el diseño de programas y políticas públicas relacionadas con la seguridad y soberanía alimentaria. Este informe entrega los resultados del estudio “Inteligencia artificial para la estimación de superficie de trigo a nivel nacional”, en el cual se desarrolló una metodología para estimar, a partir de imágenes satelitales, el número de parcelas, la superficie cultivada y la producción de trigo a nivel nacional. El modelo de clasificación del tipo de cultivo se entrenó con un conjunto limitado de datos de la región de Ñuble, para luego realizar la inferencia en las regiones del Maule, Ñuble, Bio-Bio y Araucanía. Comparando con datos del INE (2023), los resultados muestran una subestimación de la superficie y, por consiguiente, de la producción de trigo para todas las regiones analizadas. La región de Ñuble es la que presenta una estimación más cercana a los datos estadísticos oficiales, lo que indicaría que el modelo no presenta un buen nivel de generalización a otras zonas geográficas diferentes a las empleadas durante su entrenamiento. En este informe se analizan las posibles causas de estas discrepancias y se proponen estrategias para mejorar el desempeño de los distintos modelos utilizados, en particular aquellos de delineación automática de parcelas, clasificación de tipos de cultivos, y estimación del rendimiento. Finalmente, se generan manuales de uso de los modelos, incluyendo las etapas de pre y post procesamiento de los datos, y se pone a disposición los archivos con los resultados del estudio.

1. Introducción

Según los objetivos y resultados esperados, establecidos en el contrato de este estudio, este informe final incluye los siguientes puntos:

a) Una revisión de la literatura acerca de las prácticas internacionales y/o nacionales que incorporen, o lo harán, el uso de Inteligencia Artificial y Machine Learning en el sistema estadístico (capítulo 2).
b) Estimación de la superficie del trigo nacional y subnacional:
    i. Algoritmos y/o programación utilizada para la estimación (Secciones 3.2, 3.3 y 3.4).
    ii. Conjunto de imágenes y/o conjunto de datos utilizados para la clasificación y estimación (Sección 3.1).
    iii. Resultados entregados de forma visual y en planilla (Capítulo 4 y Anexo).
c) Evaluación de la estimación de la superficie de trigo y la factibilidad para avanzar hacia otros rubros del sector agrícola (Capítulo 5).
d) Entrega de materiales utilizados para la transferencia tecnológica (Anexo).

2. Revisión bibliográfica

2.1. Iniciativas internacionales relacionadas a sistemas estadísticos

A continuación, se describen las principales iniciativas internacionales relacionadas a sistemas estadísticos, donde se utilizan imágenes satelitales para el monitoreo y estimación de la producción agrícola a nivel nacional o regional.

Monitoring Agricultural ResourceS (MARS)[1]: Es el sistema de monitoreo agrícola del JRC (European Joint Research Centre). Comenzó en 1988, y fue inicialmente diseñada para aplicar tecnologías espaciales emergentes y proporcionar información independiente y oportuna sobre las áreas y rendimientos de los cultivos. Desde 1993, esta actividad ha contribuido a una gestión más efectiva y eficiente de la Política Agrícola Común (CAP)[2] a través de la prestación de una gama más amplia de servicios de apoyo técnico a la Dirección General de Agricultura y Desarrollo Rural de la Comisión Europea y las Administraciones de los Estados miembros. Desde el año 2000, la experiencia en rendimientos de cultivos se ha aplicado fuera de la UE. Se han desarrollado servicios para respaldar las políticas y la asistencia de ayuda de la UE y proporcionar elementos fundamentales para una capacidad europea de monitoreo agrícola global y de evaluación de la seguridad alimentaria. El trabajo de monitoreo de recursos agrícolas utiliza una variedad de fuentes de datos, incluyendo datos meteorológicos y pronósticos, mapas y estadísticas existentes, información de posición y datos de teledetección (de satélites y fuentes aéreas). En este último caso, el trabajo de monitoreo de recursos

[1] https://joint-research-centre.ec.europa.eu/monitoring-agricultural-resources-mars_en
[2] https://agriculture.ec.europa.eu/common-agricultural-policy/cap-overview/cap-glance_en

agrícolas ha desarrollado con éxito técnicas operativas relacionadas con la observación de la Tierra. Las actividades de monitoreo se basan en la experiencia en modelamiento de cultivos, agrometeorología, métodos de muestreo, análisis geoespacial ambiental, econometría y el uso de infraestructuras de datos europeas y globales. Adicionalmente, se han desarrollado sistemas de control para la gestión de parcelas de tierra y verificaciones de teledetección, para la implementación eficiente de la política agrícola común, incluyendo los aspectos de sustentabilidad.

El Bulletin MARS es un ejemplo concreto de uso de IA y Machine learning para extraer resúmenes estadísticos de producción agrícola. La figura 1 es un ejemplo de los datos presentados en el informe MARS de agosto de 2023. El bulletin MARS es accesible a todos a través de la página del JRC[3].

Crop	Yield t/ha				
	Avg 5yrs	July Bulletin	MARS 2023 forecasts	%23/5yrs	% Diff July
Total cereals	5.44	5.46	**5.44**	+ 0	- 0
Total wheat	5.58	5.59	**5.58**	+ 0	- 0
Soft wheat	5.79	5.80	**5.78**	- 0	- 0
Durum wheat	3.50	3.39	**3.41**	- 3	+ 1
Total barley	4.89	4.74	**4.74**	- 3	+ 0
Spring barley	4.19	3.62	**3.60**	- 14	- 1
Winter barley	5.77	5.91	**5.92**	+ 3	+ 0
Grain maize	7.48	7.53	**7.45**	- 0	- 1
Rye	3.98	4.12	**4.12**	+ 4	+ 0
Triticale	4.22	4.29	**4.31**	+ 2	+ 0
Rape and turnip rape	3.10	3.20	**3.19**	+ 3	- 0
Potatoes	34.1	34.4	**34.4**	+ 1	+ 0
Sugar beet	72.0	73.3	**73.7**	+ 2	+ 0
Sunflower	2.21	2.12	**2.18**	- 2	+ 3
Soybeans	2.76	2.86	**2.82**	+ 2	- 1
Green maize	40.7	40.6	**40.3**	- 1	- 1

Issued: 21 August 2023

Figura 1. Ejemplo de datos en el informe MARS 2023.

Sistema de Pronóstico de Rendimiento de Cultivos MARS (MCYFS)[4]: Creado en 1992 por el Centro Común de Investigación (JRC) como parte del programa MARS (descrito anteriormente), para satisfacer la necesidad de estimaciones operativas de superficie, rendimiento y producción a nivel paneuropeo para los Estados miembros de la UE. Se opera bajo el mandato del Reglamento Europeo No 1306/2013 (Art. 6 y 22). Este reglamento establece un sistema de monitoreo agrícola y pronósticos de producción y rendimiento para

[3] https://publications.jrc.ec.europa.eu/repository/handle/JRC133188
[4] https://esdac.jrc.ec.europa.eu/Projects/SINFO/index_en.htm

gestionar los mercados agrícolas. Como sistema de apoyo a la toma de decisiones, el MCYFS proporciona información independiente y basada en evidencia sobre el estado de los cultivos anuales en la UE y los países vecinos mediante el monitoreo del crecimiento de los cultivos y la predicción de los rendimientos de los cultivos. El MCYFS se basa en la adquisición y procesamiento casi en tiempo real de tres fuentes principales de datos: datos meteorológicos (observaciones y pronósticos), simulaciones de modelos de cultivos y parámetros biofísicos derivados de la teledetección por satélite para monitorear el estado de los cultivos. Todos estos datos, junto con una serie temporal de estadísticas históricas de superficie y rendimiento, se utilizan en un proceso estadístico de pronóstico de rendimiento. Se publican boletines mensuales de MARS que proporcionan una visión general sobre el desarrollo de los principales cultivos y áreas de preocupación, incluidos pronósticos de rendimiento para cereales, oleaginosas y cultivos de tubérculos, un análisis de pastos y análisis específicos por país. La información casi en tiempo real e histórica sobre las condiciones meteorológicas y el progreso del crecimiento de los cultivos se puede visualizar a través del JRC MARS Explorer[5]. Se encuentran disponibles mapas de varios indicadores meteorológicos y de cultivos, y la información se actualiza tres veces al mes.

Sistema Global de Información y Alerta Temprana sobre Alimentación y Agricultura (GIEWS)[6]: Fue establecido a principios de la década de los 70 y fue una de las primeras fuentes clave a nivel global de información sobre la producción de alimentos y la seguridad alimentaria dentro de la FAO. El GIEWS cuenta con una red de 115 gobiernos, 61 organizaciones no gubernamentales (ONG) y numerosas organizaciones comerciales, de investigación y medios de comunicación. Esta plataforma monitorea de manera continua el suministro y la demanda de alimentos, así como otros indicadores clave para evaluar la situación general de la seguridad alimentaria en todos los países del mundo. Publica informes analíticos y objetivos periódicos sobre las condiciones prevalecientes y emite alertas tempranas sobre posibles crisis alimentarias a nivel nacional o regional. A solicitud de las autoridades nacionales, el GIEWS apoya a los países en la recopilación de evidencia para la toma de decisiones políticas o la planificación de socios de desarrollo a través de sus Misiones de Evaluación de Cultivos y Seguridad Alimentaria (CFSAMs), que se realizan en colaboración con el Programa Mundial de Alimentos (WFP). Al aplicar herramientas de observación de la Tierra y monitoreo de precios a nivel nacional, el GIEWS también fortalece las capacidades nacionales en la gestión de información relacionada con la seguridad alimentaria. El GIEWS monitorea las condiciones de crecimiento de los principales cultivos alimentarios en todo el mundo para evaluar las perspectivas de producción. Para respaldar el análisis y complementar la información basada en el terreno, el GIEWS utiliza datos de teledetección que pueden proporcionar información valiosa sobre la disponibilidad de agua y la salud de la vegetación durante las temporadas de cultivo. En 2014, el GIEWS desarrolló el Índice de Estrés Agrícola (ASI), un indicador de rápida

[5] https://marsop.info/en/web/mars-explorer/home
[6] https://www.fao.org/giews/background/en/

consulta para la identificación temprana de áreas agrícolas afectadas por déficits de agua o, en casos extremos, sequías.

Sistema de Red de Alerta contra la Hambruna (FEWS NET)[7]: Fue establecido en 1985 por parte de USAID[8] con el propósito de brindar apoyo a programas de asistencia alimentaria y agencias de ayuda humanitaria. FEWS NET intenta cuantificar tanto los cambios en la superficie cultivada como el rendimiento de los cultivos, pero no monitorea directamente la producción. Actualmente, cubre 36 de los países más vulnerables en términos de seguridad alimentaria en el mundo y no solo publica informes mensuales especializados sobre la seguridad alimentaria actual y proyectada, sino que también emite alertas oportunas sobre crisis emergentes. FEWS NET utiliza diversas fuentes de datos, como imágenes de satélite y análisis de datos en el terreno, para proporcionar información crítica que contribuye a prevenir y responder a crisis alimentarias, mejorando la seguridad alimentaria en comunidades en riesgo.

CropWatch4[9], dirigido por el Instituto de Teledetección y Tierra Digital de la Academia China de Ciencias, se encarga de evaluar la producción de cultivos a nivel nacional y global. Iniciado en 1998, el propósito de este sistema es ofrecer pronósticos confiables, oportunos e imparciales sobre las condiciones y producción de cultivos, tanto en China como en todo el mundo, con el fin de planificar importaciones, exportaciones y precios de cultivos, y asegurar la seguridad alimentaria nacional. Desde 2013, CropWatch ha estado publicando boletines a nivel internacional. Este sistema considera cuatro niveles espaciales: global, regional, nacional (que incluye a treinta y un países clave, incluyendo China) y subnacional (para los nueve países más grandes). Estos treinta y un países representan más del 80% de la producción y exportación de maíz, arroz, soja y trigo. Los patrones globales de las condiciones de crecimiento se analizan utilizando indicadores como la precipitación, temperatura, radiación fotosintéticamente activa (PAR) y biomasa potencial. A nivel regional, se emplean otros indicadores como el Índice de Salud de la Vegetación (VHI) y el Índice de Condición de la Vegetación (VCI) para caracterizar la situación de los cultivos, la intensidad agrícola y el estrés. CropWatch también realiza análisis detallados de las condiciones de los cultivos a nivel nacional y subnacional, utilizando una amplia gama de variables e indicadores para obtener estimaciones de la producción de alimentos. CropWatch4 utiliza datos de teledetección de MODIS (Moderate Resolution Imaging Spectroradiometer)[10] para monitorear y evaluar las condiciones de los cultivos, analizar factores ambientales que afectan los cultivos y proporcionar información oportuna para la gestión de cultivos y la seguridad alimentaria a nivel nacional y global. Los datos de MODIS, que incluyen índices de salud vegetal (como el NDVI y el EVI), parámetros ambientales (como temperatura y precipitación) y cobertura global, son herramientas clave para seguir la salud de los cultivos, identificar anomalías y

[7] https://fews.net/
[8] https://www.usaid.gov/
[9] http://www.cropwatch.com.cn/
[10] https://modis.gsfc.nasa.gov/

analizar tendencias a largo plazo. Esto ayuda a tomar decisiones informadas relacionadas con la agricultura, estrategias de importación/exportación y planificación de la seguridad alimentaria.

2.2. Revisión bibliográfica de métodos de delineación, clasificación de parcelas, y estimación de producción.

En esta revisión bibliográfica abordamos de manera secuencial las prácticas actuales relacionadas con las cuatro técnicas en las que se basa la metodología descrita en el punto 3, a saber: recolección de datos, delineación de parcelas, clasificación de tipos de cultivos, y estimación de rendimientos de cultivos.

2.2.1. Recopilación de datos

La interacción entre la teledetección (de aquí en adelante usaremos la forma inglesa, i.e., "remote sensing" (RS)) y la inteligencia artificial (IA), en particular el aprendizaje profundo, puede ofrecer soluciones innovadoras a los problemas relacionados con los recursos terrestres y el medio ambiente. Una limitación de estos métodos de aprendizaje profundo es la necesidad de contar con una gran cantidad de datos etiquetados. Para maximizar la capacidad de obtener estos datos, se han implementado métodos de colaboración en RS con usos específicos en varios dominios (consultar [1] para una revisión exhaustiva). En agricultura, la colaboración ha sido utilizada para recopilar distintos tipos de datos [2]. En el marco de este proyecto, nos centraremos en los datos de uso/cobertura de tierras agrícolas (en inglés, land use/land cover (LULC)) [3], [4]. Los datos LULC representan parcelas en puntos únicos o geolocalizados (es decir, la delineación de los límites de las parcelas), así como en otras características como el tipo de cultivo o producto que se está cultivando. Proyectos de colaboración como Geo-Wiki [5], DIY-landcover [6] o Collect earth [7] han demostrado la utilidad de la recopilación de datos realizada por la colaboración, a través de las acciones voluntarias de actores como agricultores. Los proyectos Geo-Wiki y DIY-landcover han generado con éxito mapas de áreas de cultivo y tamaños de parcelas. Geo-Wiki se centró a escala global, mientras que DIY-landcover se focalizó específicamente en Sudáfrica.

Sin embargo, es importante tener en cuenta que, aunque estos proyectos proporcionan información valiosa, en su gran mayoría, no ofrecen detalles sobre los tipos de cultivos cultivados dentro de las parcelas delimitadas. En contraste, OpenStreetMap (OSM) ha hecho esfuerzos para mapear áreas agrícolas definiendo varios tipos de cultivos. Estos incluyen tierras de cultivo, praderas, huertos, viñedos y horticultura. No obstante, OSM permite principalmente una diferenciación básica entre tierras de cultivo y pastizales, pero carece de delimitaciones precisas de parcelas individuales. Por lo tanto, además de las limitaciones lingüísticas obvias, aún falta una plataforma de colaboración que permita a los agricultores o partes interesadas chilenas recopilar simultáneamente polígonos de campo y etiquetas de cultivos.

Un ejemplo concreto de plataforma, en el caso de la experiencia francesa, es el Registro Gráfico de Parcelas: una base de datos geográfica utilizada como referencia para la evaluación de las ayudas de la Política Agrícola Común (PAC). La versión anonimizada publicada es parte del servicio público para la provisión de datos de referencia y contiene los datos gráficos de las parcelas (unidad básica de tierra en la declaración de los agricultores) junto con su cultivo principal. Estos datos han sido producidos por la Agencia de Servicios y Pagos (ASP) desde el 2007.

Los datos anónimos del RPG están marcados con el año de producción y contienen las parcelas correspondientes a las declaradas para la campaña N en su situación conocida y establecida por la administración, generalmente el 1 de enero del año N+1. Estos datos cubren todo el territorio de Francia, incluyendo Mayotte y Saint-Martin, pero excluyendo Saint-Barthélemy.

2.2.2. Delineación de parcelas

La delimitación de los bordes de las parcelas (en inglés, parcel boundary delineation (PBD)) basada en RS desempeña un papel crucial en varias aplicaciones, incluida la agricultura de precisión, la gestión de tierras y el monitoreo ambiental. En este proyecto, la PBD es crucial para la clasificación de tipos de cultivos y la estimación de rendimientos de cultivos. Se han aplicado varios métodos para la PBD utilizando datos de RS. Métodos tradicionales de aprendizaje automático, por ejemplo, han implementado “Vector Support Machines” [8]–[11] o “Random Forest” [12], [13]. Más recientemente, los métodos de aprendizaje profundo han liderado el frente para producir mapas de delineación de parcelas. Por ejemplo, varios estudios han mostrado las impresionantes capacidades de las redes neuronales convolucionales (CNN), y en particular el uso de la arquitectura U-Net [14], para general delineación de parcelas agrícolas [15], [16]. Nuestros colaboradores en el JRC (Joint Research Center, de la comisión europea) han desarrollado un modelo de vanguardia de PBD basado en la arquitectura ResUNet-a [42], que se basa sobre la arquitectura U-Net y agrega bloques residuales [17]. Un trabajo similar ha propuesto el modelo U-TAE (U-Net Temporal Attention Encoder), que aprovecha aún más la información temporal en los datos de series temporales satelitales para realizar PBD [16]. Interesantemente, se han publicado varios conjuntos de datos que permiten entrenar y probar estos modelos. Un ejemplo de ello es el conjunto de datos AI4Boundaries desarrollado por nuestro colaborador en el JRC (Dr. Raphaël d’Andrimont) [18]. AI4Boundaries se divide en dos conjuntos de datos distintos. En primer lugar, incluye una colección de compuestos mensuales Sentinel-2 de 10 metros, que son adecuados para análisis retrospectivos a gran escala. En segundo lugar, incorpora un conjunto de ortofotos de 1 metro que facilita el análisis a escala regional. Las etiquetas para ambos conjuntos de datos se han obtenido a partir de datos accesibles públicamente que abarcan Austria, Cataluña, Francia, Luxemburgo, los Países Bajos, Eslovenia y Suecia. Este conjunto de datos incorpora un área significativa, con 2.5 millones de parcelas que suman un total de

47,105 km². Sin embargo, no está claro si los modelos de vanguardia de PBD entrenados con estos datos se generalizarán bien al territorio chileno. La causa principal de esta incertidumbre sobre la capacidad a generalizar de estos modelos emerge de las diferencias en terreno y clima entre Europa y Chile. Dicho esto, modelos como el FracTAL ResUNet [19] muestran buenos rendimientos de adaptación de dominio entre diferentes continentes.

2.2.3. Clasificación de tipos de cultivos

La clasificación de tipos de cultivos (CTC) es el proceso de identificar el tipo de cultivo en un área específica [20]. Uno de los enfoques más populares para la clasificación de tipos de cultivos es la clasificación supervisada, donde se entrena a un algoritmo de aprendizaje automático con un conjunto de datos etiquetados para clasificar los cultivos [21]. Sin embargo, vale la pena señalar que también se han desarrollado métodos no supervisados [22], basados, por ejemplo, en el agrupamiento isométrico de componentes principales [23]. Un enfoque popular para la clasificación de tipos de cultivos es el análisis de imágenes basado en objetos (OBIA), que implica segmentar la imagen en objetos significativos y luego clasificarlos en función de sus características espectrales, espaciales y texturales [24]. En particular, se ha demostrado que OBIA es efectivo para la clasificación de tipos de cultivos en áreas con un uso fragmentado de la tierra agrícola

Sin embargo, estudios recientes han mostrado que las redes neuronales convolucionales (CNN) pueden mejorar significativamente la precisión de la clasificación de tipos de cultivos en comparación con algoritmos tradicionales de aprendizaje automático [25]. Por ejemplo, un modelo desarrollado recientemente por el Dr. Valentin Barriere (miembro del equipo CENIA) ha demostrado producir resultados de alto nivel en la clasificación de tipos de cultivos fuera de temporada utilizando un modelo jerárquico con una arquitectura compuesta por una RNN-LSTM (red neuronal recurrente - memoria a corto y largo plazo) que implementa además mecanismos de atención bidireccionales [26]. Este modelo presenta dos ventajas. En primer lugar, utiliza la atención para aprovechar la información temporal. En segundo lugar, es un modelo multimodal que incorpora datos de rotación de cultivos y datos basados sobre atributos extraídos de los tiempos series de bandas espectrales, fusionando estos datos de manera jerárquica. Estas dos características importantes permiten que este modelo de CTC alcance una precisión del 80% 6 meses antes de la cosecha en un entorno de 28 clases.

2.2.4. Estimación de rendimiento

La estimación de rendimientos de cultivos (en inglés, crop yield estimation (CYE)) implica predecir la cantidad de cultivos que se cosecharán en un área específica [27]. El enfoque empírico más común para la estimación de rendimientos de cultivos implica el uso de una combinación de datos de RS, datos meteorológicos y observaciones basadas al nivel del terreno para predecir los rendimientos de cultivos a través de métodos estadísticos (es decir, modelos de regresión; [28]). Los algoritmos de aprendizaje automático, como árboles de

regresión [29] y “support vector machines” [30], también se han utilizado para la estimación de rendimientos de cultivos. Recientemente, se han aplicado algoritmos de aprendizaje profundo, como CNN, a la estimación de rendimientos de cultivos con resultados prometedores [31]–[34]. Estos algoritmos pueden aprender relaciones complejas entre los datos de RS y los rendimientos de cultivos, y son capaces de capturar relaciones no lineales que los algoritmos tradicionales de aprendizaje automático pueden pasar por alto. Uno de los desafíos en la estimación de rendimientos de cultivos es la variabilidad espacial y temporal del crecimiento de los cultivos dentro de un campo. Para abordar este desafío, los estudios han propuesto el uso de datos de RS de alta resolución, como imágenes de vehículos aéreos no tripulados (en inglés, unmanned aerial vehicle (UAV)), para capturar la variabilidad espacial del crecimiento de los cultivos dentro de un campo [35]. El uso de series temporales de RS, como Landsat o Sentinel-2, puede capturar la variabilidad temporal del crecimiento de los cultivos a lo largo de la temporada. Además, para abordar este problema, se ha demostrado que la integración de múltiples fuentes de datos, como datos de RS, datos meteorológicos y modelos de crecimiento de cultivos, mejora la precisión de la estimación de rendimientos de cultivos [36], [37]. Específicamente, se han desarrollado modelos de rendimiento para el trigo de invierno combinando datos tanto satelitales como climatológicos [38].

Más recientemente se ha propuesto un modelo multimodal de CYE [36]. CropYieldNet toma en cuenta simultáneamente datos satelitales y meteorológicos (CYN; figura 2). Este modelo aprovecha múltiples fuentes de datos y modalidades para mejorar su precisión predictiva, lo que lo hace adaptable para una amplia variedad de escenarios agrícolas. En términos de arquitectura, CYN consta de cuatro módulos: el Codificador de Reflectancia Superficial (SRE), que aprende los patrones espaciales en los datos de reflectancia superficial sin alterar los patrones temporales; el Codificador de Datos del Suelo (SDE), que aprende la información de intensidad de píxeles en cada atributo del suelo; y el Módulo Temporal Central (CTM), que explora los patrones temporales en los datos de reflectancia superficial y meteorológicos para finalmente predecir el rendimiento. Es importante destacar que los cuatro módulos se aprenden de manera conjunta en un proceso de extremo a extremo mediante el descenso de gradiente estocástico y el aprendizaje contrastivo aplicado a datos de satélite (reflectancia superficial y datos del suelo).

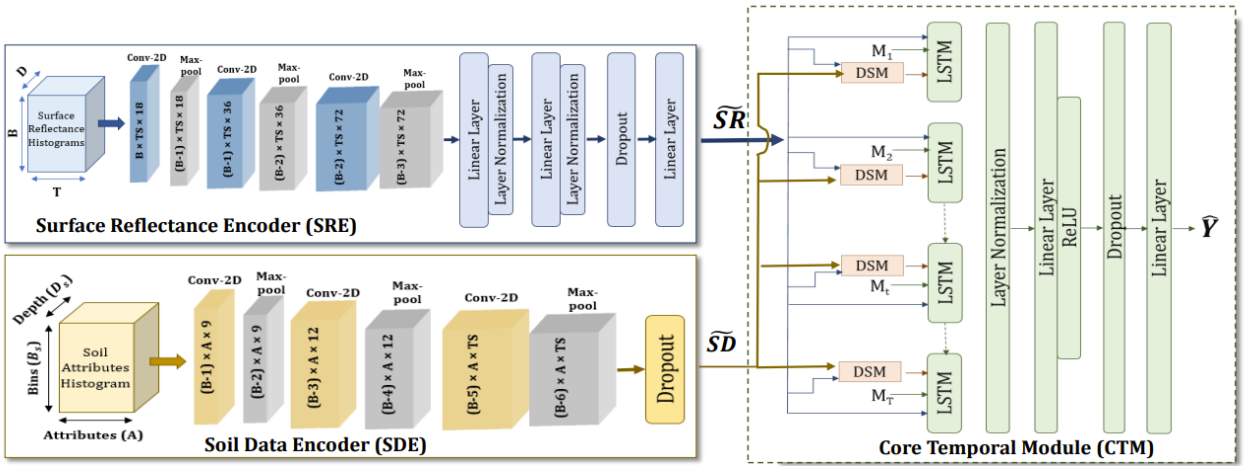

Figura 2. Arquitectura del modelo CropYieldNet.

2.3. Experiencia nacional

En Chile, ha habido un número muy limitado de estudios académicos que investiguen PBD, CTC o CYE a través de RS, y ninguno de ellos ha aplicado métodos de aprendizaje profundo a este problema. Por ejemplo, ha habido estudios utilizando métodos clásicos de aprendizaje automático como "support vector machines", "linear discriminant analysis" y "random forest" para clasificar el tipo de árboles frutales en la región del valle del Maipo [39], [40]. Más recientemente, un estudio se ha centrado en predecir la reducción de la productividad agrícola chilena inducida por sequía [41]. Por lo tanto, el desarrollo de modelos de clasificación y estimación, así como tecnologías para recopilar los datos necesarios in situ (datos reales), están aún ausentes del paisaje tecnológico chileno.

En el ámbito corporativo, hay algunas instituciones dedicadas a usar datos de RS en diferentes áreas relacionadas a detectar el estado de la tierra de cultivo, su calidad y estimar la producción de cultivos. La compañía AGROSAT ofrece un abanico de productos relacionados con el uso de RS para determinar desde la fertilidad, nivel de nutrientes y componentes del suelo hasta estimación del rendimiento. Entre estos productos se destaca Cropsense, definido como "[un] servicio de sensoramiento que hace una estimación de los rendimientos productivos previos a cosecha". Incluye facilidades para identificar y organizar zonas productivas y para evaluar la producción y sus potenciales beneficios económicos. En la oferta no se revelan estrategias de automatización de PBD. Por su parte, el Centro de Investigación e Innovación de la Viña Concha y Toro en 2018 publicó una noticia que involucra drones e inteligencia artificial [42] para amortiguar los errores en la predicción del volumen de cosecha de las uvas. El proyecto recoge imágenes multiespectrales con drones, que son procesadas por modelos de inteligencia artificial que identifica, cuantifica y predice la producción. Sin embargo, el uso de drones presenta importantes desafíos en el escalamiento de la solución, considerando la necesidad de disponer de esta información con una cobertura a nivel nacional.

3. Metodología de delineación, clasificación, estimación de superficie de parcelas, y predicción de la producción de trigo en Chile.

Nuestra solución se basa en un proceso de seis pasos asociado con la implementación o adaptación de tecnologías listas para su uso, las cuales han alcanzado el nivel de madurez tecnológica requerido para esta aplicación. La Figura 3 muestra una visión general de los procesos involucrados en nuestro plan de trabajo. Basándose en imágenes satelitales, se pre-procesan los datos (paso 1) y se recopila información sobre polígonos (es decir, datos de delimitación de parcelas) y etiquetas de cultivos (paso 2). Esto permite segmentar secuencialmente las parcelas de cultivo (paso 3), clasificar el tipo y estimar el área de cultivo (paso 4) y estimar la producción (paso 5). Cada paso de la solución se describe en las siguientes secciones. Nuestra solución tecnológica se basa en código abierto, así como en artículos publicados que se re-implementan, utilizando, cuando están disponibles, códigos sin licencias restrictivas. Posteriormente, todos nuestros modelos y conjuntos de datos serán compartidos como recursos de código abierto para la comunidad de investigación.

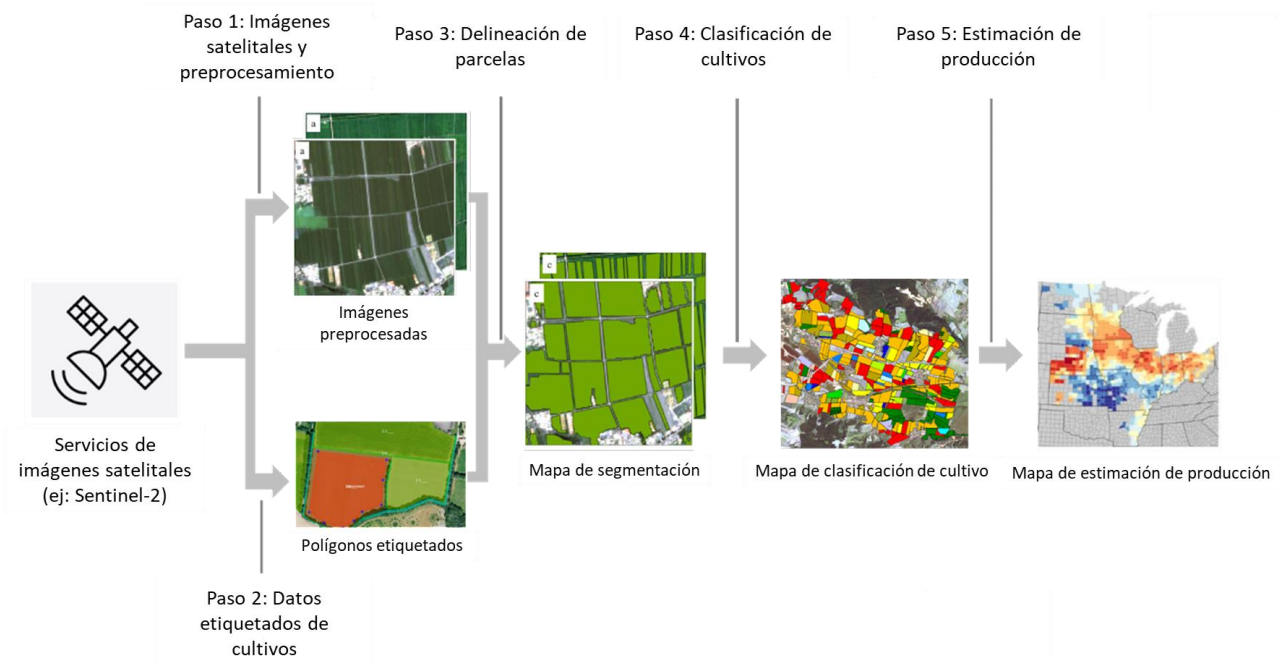

Figura 3. Metodología general.

3.1. Preparación del dataset

Para poder entrenar los modelos de delineación de parcela y de clasificación de tipo de cultivo se prepara un dataset específico. En esta sección describimos los pasos necesarios para poder generar este dataset y sus componentes.

3.1.1. Recopilación y preprocesamiento de datos satelitales (Paso 1)

Los datos satelitales se obtienen desde el sistema Copernicus. Se creó código usando librerías de Python para descargar productos desde Sentinel-2 L2A (es necesario tener credenciales de una cuenta de Copernicus para la descarga).

El proceso de descarga de estos archivos presenta múltiples desafíos:

- El peso de los archivos *.SAFE: varía de 0.5 a 1.5GB, lo que hace lenta la descarga. Esto es debido a que se maneja a nivel de celda (110 x 110 km²) y contiene todas las bandas espectrales.
- Superposición de celdas: las celdas tienen un alto nivel de superposición, por lo que es importante considerarlo al momento de trabajar con los datos.
- Formato de los resultados: el formato *.SAFE consiste en una estructura de carpetas. Dentro, los datos de los sensores satelitales están en formato XML, lo que implica un procesamiento desde XML a un formato de cubo (ej: NetCDF4) para su uso en librerías y modelos que trabajan sobre datos satelitales.
- Long-time Archive (LTA): después de cierta cantidad de tiempo, los productos de Sentinel-2 pasan al archivado de largo plazo. Actualmente pasan a LTA luego de 18 meses en el caso de productos L2A. Las cuentas de Copernicus tienen límites sobre la cantidad de solicitudes al LTA que se pueden hacer en un mismo día.
- Bases de procesamiento: las diferentes bases de procesamiento implican modificaciones a los datos. Este proceso ejecuta cambios en los valores de las bandas espectrales y por lo tanto modifica los rangos de valores que se trabajan en las bandas.
- Se recomienda revisar el documento de convención de nombres para identificar las diferentes variables en la recolección de datos.

Durante este proceso, se descargan archivos de diferentes años para Ñuble. Se tienen datos del 2021 hasta 2023. Se tiene al menos una o dos imágenes de cada mes mencionado. Las diferentes solicitudes hechas durante el proyecto corresponden a las siguientes:

- 2021-febrero hasta 2021-Noviembre, 2022-Enero hasta 2022 Abril, 2022-Junio a 2022-Agosto
    - Bases de procesamiento: 'N0214', 'N0300', 'N0301', 'N0400'
    - Celdas: 'T18HXD', 'T18HXE', 'T18HXF', 'T18HYD', 'T18HYE', 'T18HYF', 'T19HBA', 'T19HBU', 'T19HBV', 'T19HCA', 'T19HCU', 'T19HCV'
    - Órbitas relativas: 'R053', 'R096', 'R139'
- 2022-mayo (al no obtener datos de mayo en la solicitud anterior, se trabajó independientemente)
    - Fechas: '20220528', '20220531
    - Bases de procesamiento: 'N0400'
    - Celdas: 'T18HXD', 'T18HXE', 'T18HXF', 'T18HYD', 'T18HYE', 'T19HBU', 'T19HBV', 'T19HCU', 'T19HCV'
    - Órbitas relativas: 'R096', 'R139'

- 2022-febrero a 2022-Abril (aumentar cantidad de imágenes dado que en la consulta anterior no se obtuvieron suficientes)
- 2022-agosto hasta 2022-Diciembre
    - Bases de procesamiento: 'N0400'
    - Celdas: 'T18HXD', 'T18HXE', 'T18HXF', 'T18HYD', 'T18HYE', 'T18HYF', 'T19HBU', 'T19HBV', 'T19HCU', 'T19HCV'
    - Órbitas relativas: 'R053', 'R096', 'R139'
- 2022-octubre hasta 2023-Abril (se aplicó un filtro para obtener una celda específica, órbita relativa e imágenes con bajo porcentaje de cobertura por nubes)
    - Bases de procesamiento: 'N0400', 'N0509'
    - Celda: 'T19HBV'
    - Órbitas relativas: 'R096'
- 2023-julio
    - Bases de procesamiento: 'N0509'
    - Celdas: 'T18HXD', 'T18HXE', 'T18HXF', 'T18HYD', 'T18HYE', 'T18HYF', 'T19HBU', 'T19HBV', 'T19HCU', 'T19HCV'
    - Órbitas relativas: 'R053', 'R096', 'R139'

Para recopilar datos a nivel nacional, se contactó con el centro de modelamiento matemático (CMM de la Universidad de Chile), un centro enfocado en la investigación y aplicación de modelos matemáticos en diferentes áreas. Una de ellas es la disponibilización de datos de satélites de Copernicus para Chile. El CMM es el copernicus hub para Chile. Inicialmente, con la idea de entrenar el modelo con datos anotados sobre la región de Ñuble, los contactamos para gestionar la obtención de datos satelitales de esta misma región. Solicitamos las celdas 18HYE y 19HBV, con procesamiento L2A para el periodo del mes de enero 2021 hasta septiembre 2022. Las celdas mencionadas fueron elegidas por el porcentaje de cobertura que tienen sobre la región de Ñuble, siendo las dos con mayor superposición con esta zona. En la figura 4 se puede apreciar que ambas celdas cubren un gran porcentaje de la región.

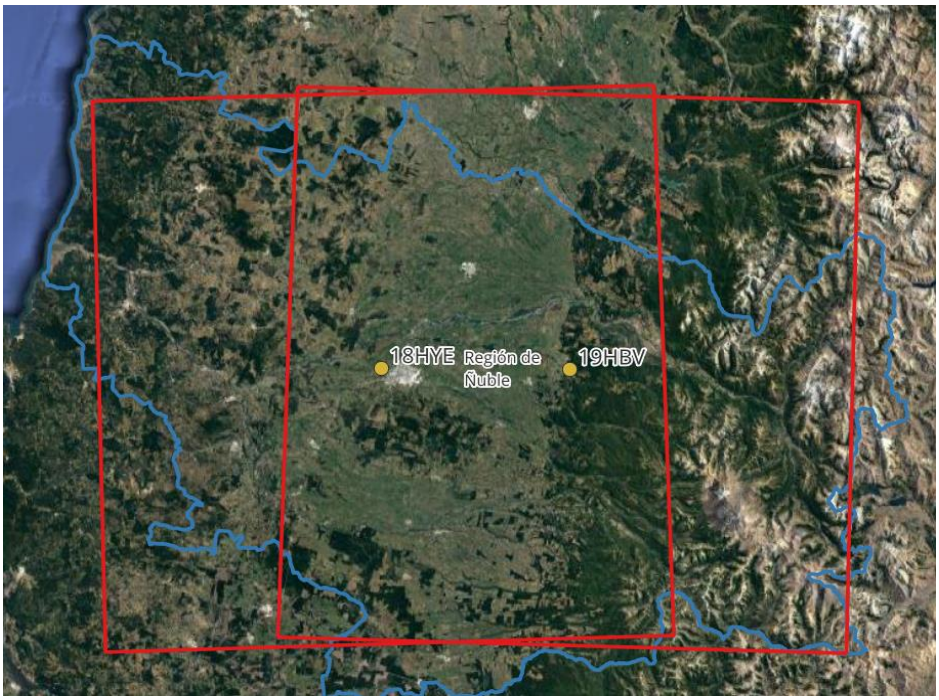

Figura 4. Celdas 18HYE y 19HBV para descarga de imágenes satelitales.

Los datos disponibles actualmente y facilitados por CMM son los siguientes:

- Año 2021 y 2022 completos
    - Bases de procesamiento: 'N9999'
    - Celdas: '18HYE' y '19HBV
    - Órbitas relativas: 'R053', 'R096'

Los datos incluyen varias fechas mensuales para cada celda, con un rango de tres a seis fechas por celda/mes.

Adicionalmente, se solicitan los datos de años anteriores y aumentar la cantidad de celdas a circundantes las celdas 18HYE y 19HBV. Específicamente, los años 2020, 2019, 2018 y 2017, y las celdas circundantes 19HBA, 18HYF, 19HBU, y 18HXD (ver figura 5).

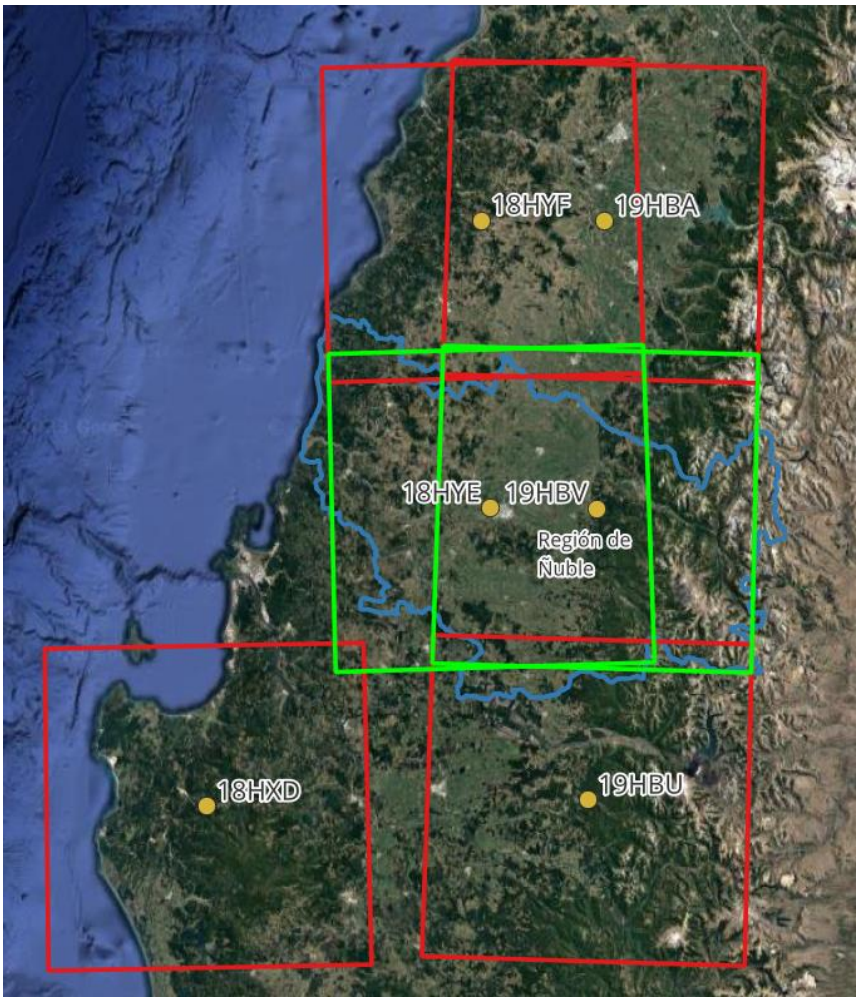

Figura 5. Distribución de las celdas adicionales que se solicitan al CMM (rojo) junto a las ya obtenidas (verde).

Actualmente, y con el ambito de tener una cobertura a nivel nacional, se ha solicitado al CMM la descarga de las celdas: “19JBG”, “19JCG”, “19JBF”, “19JCF”, “19HBE”, “19HCE”, “19HBD”, “19HCD”, “19HBC”, “19HCC”, “19HBB”, “19HCB”, “18HYF”, “19HBA”, “18HXE”, “18HYE”, “18HXD”, “18HYD”, “18HXC”, “18HYC”, “18HXB”, “18HYB”, “18GXA”, “18GYA”, “18GXV”, “18GYV”, “18GWU”, “18GXU”, “18GWT”, “18GXT”, “19KCV”, “19KDV”, “19KCU”, “19KDU”, “19KCT”, “19KDT”, “19KCS”, “19KDS”, “19KCR”, “19KDR”, “19KCQ”, “19KDQ”, “19KCP”, “19KDP”, “19JCN”, “19JDN”, “19JCM”, “19JDM”, “19JCL”, “19JDL”, “19JCK”, “19JDK”, “19JBJ”, “19JCJ”, “19JBH”, “19JCH”.

3.1.2. Recopilar información de terreno a través de múltiples métodos de recopilación de datos (Paso 2).

En esta sección se detalla la recopilación de datos asociados a los cultivos agrícolas en la región de Ñuble. Estos datos serán utilizados tanto para el entrenamiento como para la validación del modelo de clasificación de cultivo descrito en la Sección 3.3.4. Dada la diversidad de fuentes de información que ha sido levantada y compartida por ODEPA, éstas se analizan

por separado. Para mayor facilidad en la descripción y análisis de la información, ésta se divide en dos subgrupos: polígonos multi-campos, y polígonos uni-campo. Como su nombre lo indica, el primer grupo se refiere a los polígonos que contienen más de un campo o parcela de cultivo, mientras que el segundo grupo incluye aquellos polígonos que describen campos o parcelas de cultivo de forma individual.

3.1.2.1. Polígonos multi-campos

Catastro de propiedades rurales

Esta capa de polígonos, compartida por ODEPA (Oficina de Estudios y Políticas Agrarias), es generada desde el CIREN (Centro de Información de Recursos Naturales) y el SII (Servicio de Impuestos Internos de Chile) para el año 2017. Cada polígono representa una propiedad. La mayoría de ellos tendrán más de un terreno de cultivo. Este conjunto de datos incluye los sitios rurales en detalle, y sitios urbanos de manera simplificada (fusionados). El tamaño de las propiedades varía desde las más grandes, que son aproximadamente 45,000 hectáreas en la zona montañosa, hasta las más pequeñas, que son 0.0069 hectáreas. El conjunto de datos para la región de Ñuble consta de 56,634 polígonos y para cada comuna dentro de la región, cada uno de los polígonos tiene un identificador único o ROL. Un ejemplo de estos polígonos se puede ver en la figura 6.

Figura 6. Ejemplo de polígonos del catastro de propiedades rurales.

Seguros INDAP

Datos relacionados con los seguros otorgados por INDAP a los productores. Incluye datos desde 2017 hasta 2022. Cubre 300 tipos de cultivos, incluyendo 6 variedades de trigo. La base de datos original de los seguros INDAP corresponde a polígonos multi-campo. Para georreferenciar estos polígonos, se utiliza el identificador ROL asociado al seguro, por lo que su definición geométrica es equivalente a la definida en el catastro de propiedades rurales (descrito anteriormente).

Atributo	Descripción
poliza	identificador de la póliza
rubro	clase del cultivo
cultivo	subclase del cultivo
riego	tipo de riego (riego o secano)
región	Nombre de la región
comuna	Nombre de la comuna
rol	Identificador ROL asociado a la póliza
rendimiento	Rendimiento del cultivo asegurado
superficie	Superficie del cultivo asegurado
fecha siembra	Fecha de la siembra
fecha cosecha	Fecha de la cosecha
fecha inicio vigencia poliza	Fecha de inicio de vigencia de la póliza
fecha término vigencia poliza	Fecha de término de vigencia de la póliza
año	Año de la póliza

Tabla 1. Descripción de atributos de tablas de seguros INDAP.

En la tabla 2 se muestra el número total de seguros INDAP y aquellos asociados a cultivos de trigo por año.

Año	Seguros	Seguros de trigo
2017	1090	526
2018	978	464
2019	1236	625
2020	1035	441
2021	839	347
2022	940	452

Tabla 2. Número total de seguros INDAP por año

3.1.2.2. Polígonos uni-campo

Seguros INDAP (sólo trigo)

A partir de los datos de seguros INDAP de carácter multi-campo, ODEPA apoyó en la delineación manual de polígonos dentro de cada uno de los sitios utilizando como apoyo, imágenes satelitales (Sentinel-2) proporcionadas por el equipo de CENIA. La figura 7 muestra un ejemplo de polígonos definidos por este método.

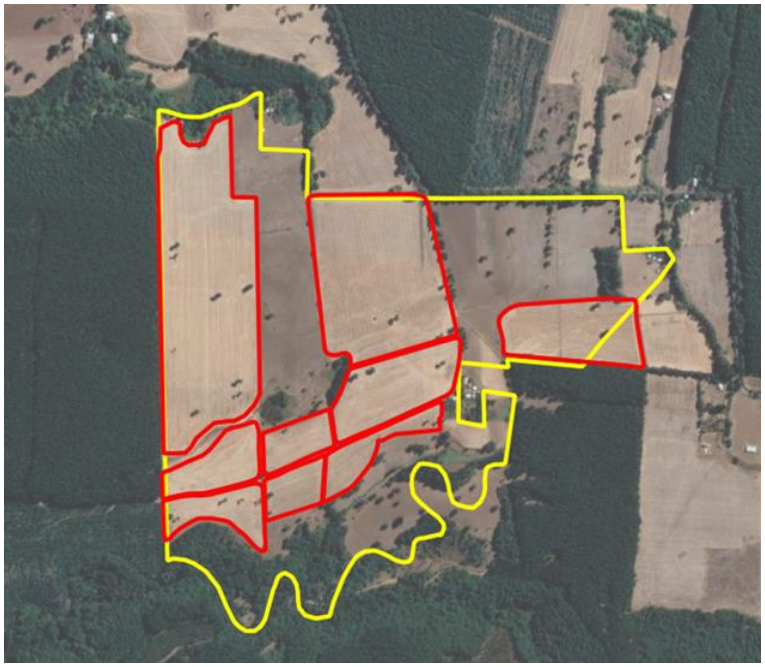

Figura 7. Ejemplo de polígonos de trigo definidos manualmente por ODEPA según la información de seguros INDAP e imágenes Sentinel-2.

Aún se requiere discutir si este enfoque es escalable a otras regiones del país. Por un lado, se introduce una nueva fuente de error humano, pero por otro lado, se aumentan significativamente los datos etiquetados, los cuales son indispensables para el entrenamiento y validación de los modelos de clasificación de cultivos.

Programa de gestión INDAP

Esta capa de polígonos uni-campo tiene información sobre el rendimiento del trigo (en quintales por hectárea) para una pequeña proporción de parcelas (no propiedades) en un municipio específico (El Carmen, Yungay). Los datos están disponibles para los años 2021 y 2022 y provienen del programa de gestión de INDAP (este programa abarca solo la región de Ñuble). Los polígonos son creados por el representante de INDAP. Para 2021, hay 105 polígonos (38 para Yungay y 67 para El Carmen). Para 2022, hay 87 polígonos (30 para Yungay y 57 para El Carmen). Los polígonos no tienen un ROL asociado. La figura 8 muestra algunos ejemplos (en amarillo) de polígonos asociados al programa de gestión INDAP.

Figura 8. Ejemplo de polígonos (en amarillo) del programa de gestión de INDAP.

Catastro frutícola

Datos relacionados con el catastro frutícola en la región para el año 2022. Éste proviene del Centro de Información de Recursos Naturales (CIREN). El conjunto de datos consta de 6,452 polígonos con 6 atributos (Id, Provincia, Comuna, ROL, Especie y Superficie). La Figura 9 muestra algunos ejemplos de polígonos del catastro frutícola.

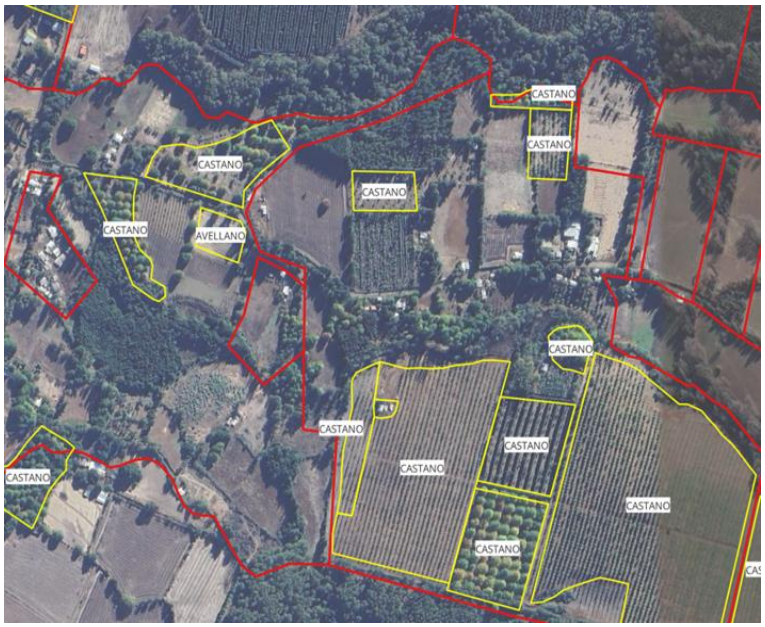

Figura 9. Ejemplo de polígonos (en amarillo) de catastro frutícola.

En la tabla 3 se muestra la superficie total por especie frutícola, considerando “otras categorías” para agregar especies con baja superficie.

Especie	Superficie total (ha)
Avellano	6.558,7
Arándano Americano	4.142,5
Cerezo	2.973,4
Nogal	1.972,5
Frambuesa	1.096,7
Castaño	868,8
Manzano rojo	726,6
Kiwi	233,1
Moras cultivadas e híbridas	215,8
Otras categorías	343,9

Tabla 3. Superficie total por especie frutícola.

En resumen, la cantidad de polígonos que representan una parcela (uni-campo) con información anotada (etiquetada) relativa al tipo de cultivo se compone de los polígonos del programa de gestión INDAP y datos de seguros INDAP para el caso de cultivos de trigo, y del catastro frutícola para otros cultivos. El resumen con el número de polígonos se muestra en la tabla 4, para los años 2021 y 2022. Cabe destacar que, si bien el catastro frutícola incluye información del 2022, se asume que las parcelas tienen el mismo tipo de cultivo en el año 2021.

Año	Trigo	Especies frutícolas
2021	266	6452
2022	281	6452

Tabla 4. Número de parcelas etiquetadas con tipo de cultivo.

En la Figura 10 se muestran los polígonos anotados para la región de Ñuble para el año 2021.

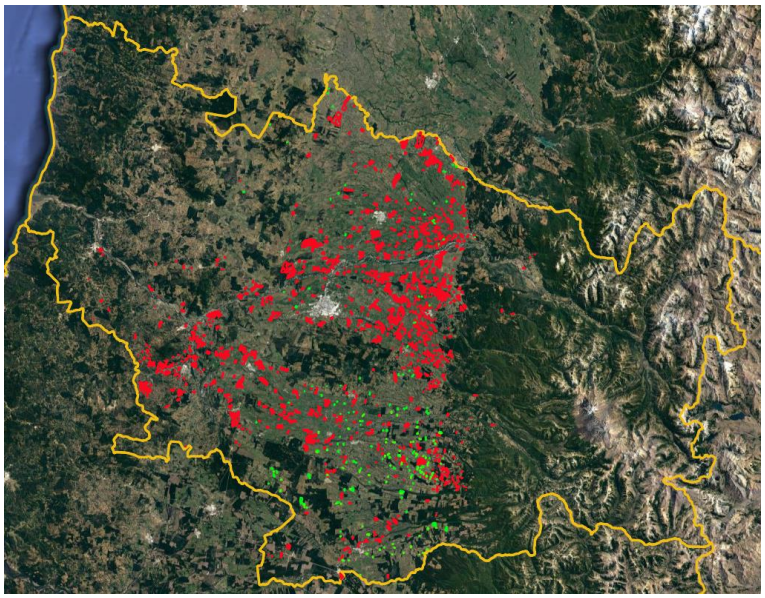

Figura 10. Polígonos anotados para Ñuble 2021. En verde polígonos de trigo, y en rojo polígonos con especies frutícolas.

3.2. Método de delineación de parcelas (Paso 3)

3.2.1. Descargar y armar dataset AI4Boundaries

El primer paso para entrenar el modelo de delineación de parcela es descargar el dataset AI4Boundaries[11], un dataset para delineación de parcelas. Este dataset contiene 7831 muestras en formato NetCDF4., de tamaño 256 x 256, obtenidas desde el Satélite Sentinel-2. Cada muestra tiene 5 bandas multiespectrales con granularidad 10m. Las bandas usadas son la B2 (azul), B3 (verde), B4 (rojo), B8 (infrarrojo cercano) y NVDI (índice de vegetación de diferencia normalizada). El peso aproximado del dataset es de 60GB, y ejemplos de imágenes de este dataset pueden ser visualizadas en la figura 11.

También contiene las etiquetas para las muestras; estas consisten en 4 máscaras de segmentación (c), bordes (d), distancia a los bordes (e) y enumeración de parcelas (f). Las imágenes provienen de 7 países diferentes de la Unión Europea. Están en formato TIFF.

[11] https://github.com/waldnerf/ai4boundaries

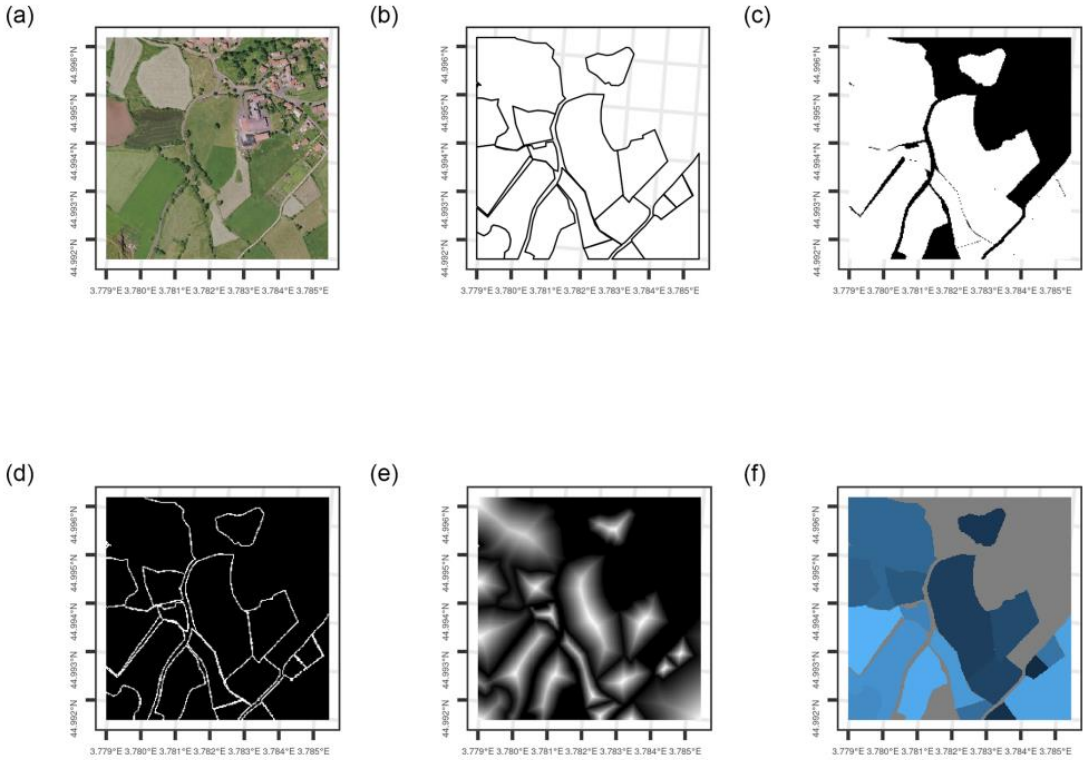

Figura 11. Ejemplo de AI4Boundaries.

Respecto a la dimensión temporal, en la mayoría de las zonas estudiadas se incorporaron muestras de diferentes fechas. En la figura 12 siguiente se muestran para cada país la cantidad de muestras temporales disponibles en el dataset por zona.

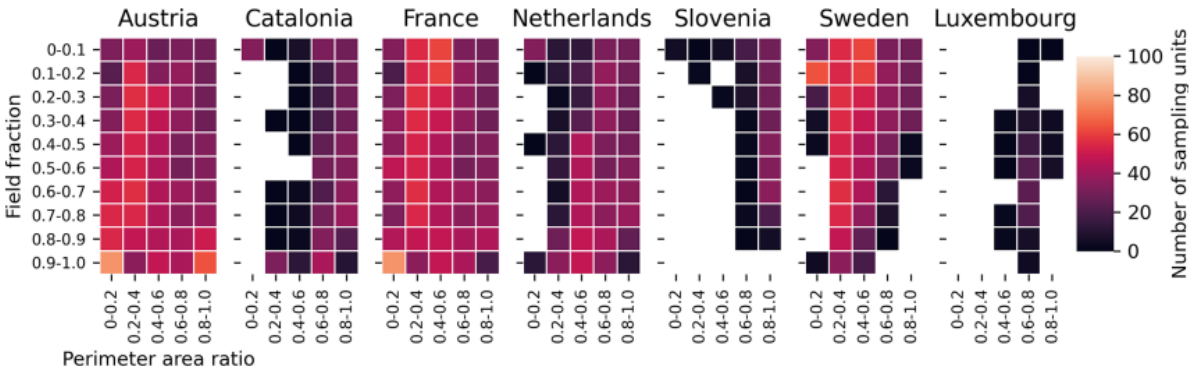

Figura 12. Muestras temporales por zona.

El dataset se compone de archivos .nc que se deben transformar en .png de dimensión 256x256. Para cada timestamp del archivo .nc (netcdf) se genera una foto (.png) con sus respectivas máscaras (ver figura 13). Para poder entrenar el modelo descrito en la sección siguiente se prepara un dataset de entrenamiento, uno de validación, y uno de testeo. Este

"split" de datos es un proceso común para poder evaluar la calidad del modelo en generalizar a datos que no fueron usados durante el entrenamiento[12].

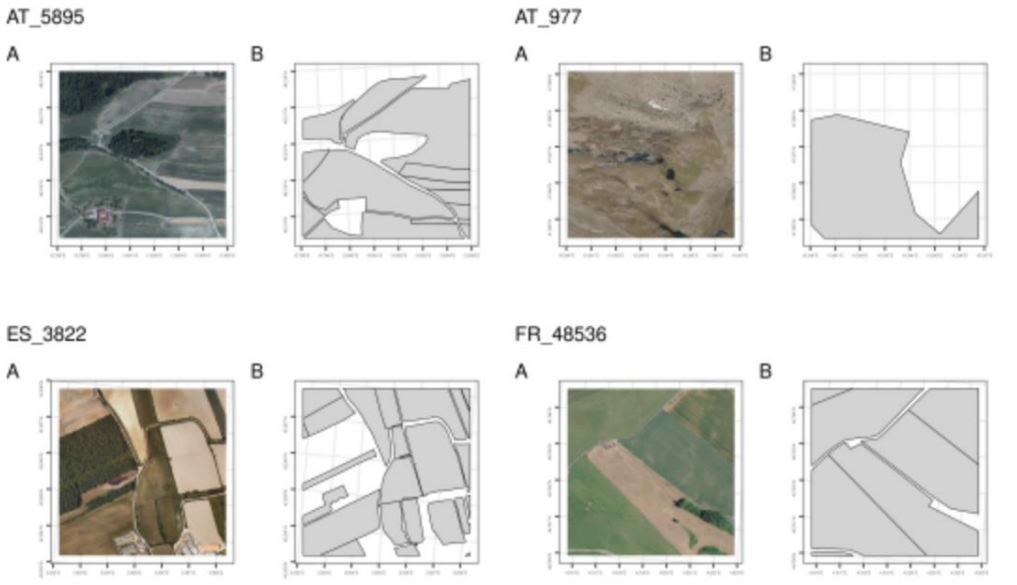

Figura 13. Ejemplos de png. (256x256 pixeles) del dataset AI4Boundaries con sus máscaras respectivas.

3.2.2. Pre-procesamiento datos satelitales

Los datos Sentinel-2 MSI con datos de reflectancia superficial de Nivel-2A se descargan del Copernicus Open Access Hub. Los datos L2A son preprocesados para enmascarar información no relacionada con la superficie, i.e., píxeles señalados como nube, sombra de nube, cirros y nieve. Los mapas de variables biofísicas de Índice de Área Foliar (LAI) y Fracción de Radiación Fotosintéticamente Activa Absorbida (FAPAR) se derivan de los productos S2 L2A (resolución espacial de 20m) utilizando el algoritmo BV-NET[13]. Este algoritmo permite recuperar las dos variables biofísicas a partir de la reflectancia multiespectral utilizando la inversión del modelo de transferencia radiativa PROSAIL y una red neuronal artificial de retropropagación (ANN).

Se hace uso de las bandas roja (B04) e infrarroja cercana (NIR, B08). Por lo tanto, se usan mapas basados en píxeles de cuatro variables (LAI, FAPAR, banda roja y banda NIR) para derivar series temporales por campo. A pesar de filtrar los datos con estas etapas de pre-procesamiento, las series temporales basadas en parcelas resultantes siguen contaminadas por la presencia de nubes, sombras de nubes, niebla o atmósfera densa. Para eliminar estos valores atípicos restantes, se aplica un filtro Hampel usando las bandas roja y NIR para descartar nubes y sombras de nubes en las series temporales, respectivamente. Finalmente, las series temporales filtradas de las cuatro variables (banda roja, banda NIR, LAI y FAPAR)

[12] https://encord.com/blog/train-val-test-split/
[13] M. Weiss and F. Baret, "Evaluation of canopy biophysical variable retrieval perfor-mances from the accumulation of large swath satellite data," Remote Sens Envi-ron, vol. 70, no. 3, pp. 293–306, 1999

agregadas a nivel de parcela se suavizaron individualmente utilizando el método Whittaker[14]. Las series temporales se rellenan para cerrar los vacíos temporales utilizando una interpolación lineal y un intervalo de tiempo de 2 días. Se aplica la configuración de Whittaker con optimización de la curva V del parámetro de suavizado y suavizado de “expectile” utilizando un peso asimétrico. Utilizamos un valor de "Envoltura" de 0.9 y un rango de lambda probado entre -1 y 1. Esto resultará en una serie temporal suavizada con un intervalo de tiempo de 4 días y sin interrupciones entre las estaciones.

Además, se ejecutaron los siguientes pasos. Se generó un archivo de bandas B2, B3, B4, B8 (RGB y NIR), con una resolución de 10m por pixel, y uno archivo de banda SCL (por “scene classification”, para manejar la presencia de nubes), con una resolución de 20m por pixel. Unimos las bandas de colores interpolando el SCL. Durante estos pasos, se segmentó la imagen original en subceldas menores de 2100 x 2100 antes de iniciar los pasos de preprocesamiento debido al tamaño de cada archivo. Luego, se aplica el algoritmo de preprocesamiento para eliminar la presencia de nubes.

El paso final corresponde a combinar las bandas existentes para generar imágenes RGB en formato .png. Esto conlleva una segunda subdivisión para trabajar en la resolución del modelo, de 256 x 256 pixeles. Es importante notar que las subceldas tienen dimensiones no divisibles por esta resolución. Se decidió que en los casos bordes de esta división se repitan los pixeles anteriores necesarios para completar el tamaño completo. Como se tienen múltiples timestamps, se generó una foto para cada momento en el tiempo. La figura 14 presenta una imagen de muestra de 256 x 256 pixeles de Ñuble. Estos resultados requieren pasos de procesamiento de imágenes descritos más adelante.

Figura 14. Ejemplo de la celda 18HYE de Ñuble. La celda completa tiene tamaño 2100 x 2100, por lo que antes de seguir, se divide en imágenes pequeñas de 256 x 256 pixeles.

[14] https://github.com/WFP-VAM/vam.whittaker

En la figura 14, se nota que la imagen de la derecha es el resultado de combinar las bandas 2, 3 y 4 del archivo NetCDF4 después de normalizar los valores y ajustarlos al rango 0 a 255 del formato PNG. Se puede notar que la imagen se ve opaca, por lo que es necesario aplicar una corrección de color para subir la iluminación. Las imágenes siguientes ilustran el proceso para la corrección de brillo (figura 15).

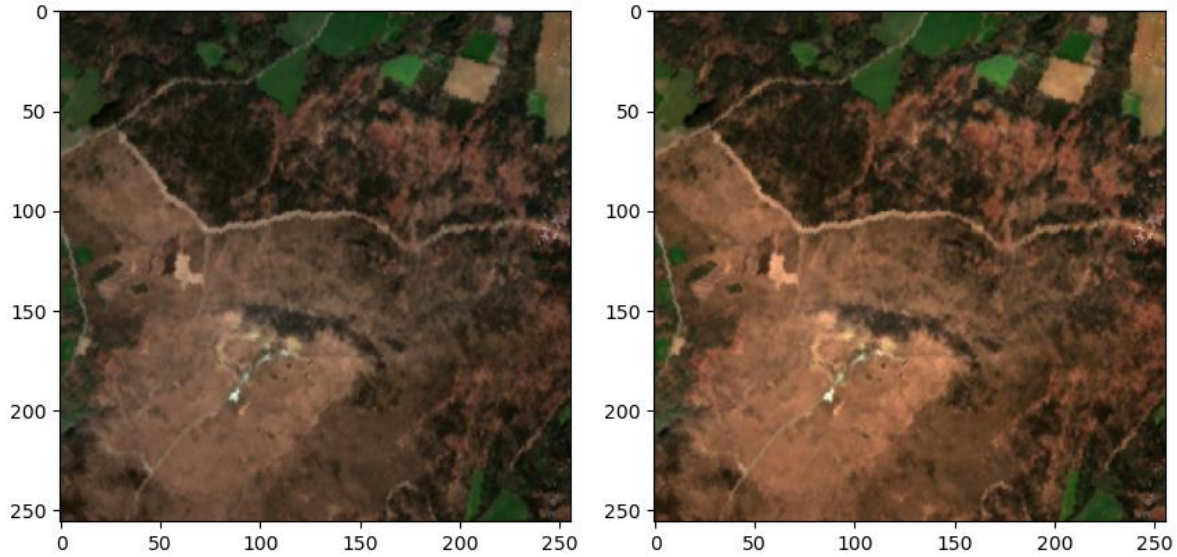

Figura 15. Corrección de brillo.

También se aplica una corrección de color para que los datos de Ñuble tengan una distribución de colores semejante a los datos originales. La imagen siguiente muestra una foto de Ñuble como la fuente (Source) y una foto del dataset para la referencia (Reference). La foto de la derecha es el resultado de aplicar la corrección (figura 16).

Figura 16. Corrección de color.

3.2.3. Entrenar modelo ResUNet-a

ResUNet-a es un modelo enfocado en hacer segmentación semántica en imágenes de alta resolución[15]. En la figura 17 se muestra el funcionamiento del modelo:

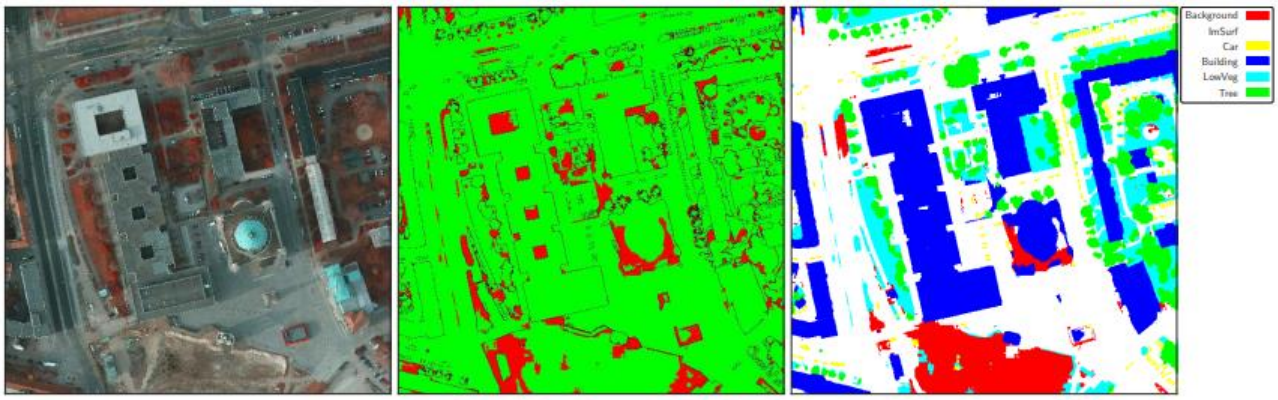

Figura 17. Ejemplo de segmentación con modelo ResUNet-a.

Este modelo se condujo sobre imágenes de ciudades, pero la versión del método utilizada en este trabajo se aplica para hacer la tarea de delineación de parcelas. Es necesario realizar un entrenamiento desde cero enfocado en imágenes de zonas de cultivo para que logre identificar los campos. Para ello utilizamos el dataset mencionado en la sección 3.2.1.

Las entradas y salidas de este modelo son en formato PNG, o sea, son imágenes. Por eso es necesario que cualquier fuente de datos que se use sea procesada para lograr este formato. Además, se debe aplicar una etapa de post-procesamiento sobre los resultados finales para trabajar los bordes como polígonos georeferenciados.

El procesamiento de los datos de entrada consiste en transformar las muestras del dataset en imágenes PNG. Estos archivos contienen información de bandas multiespectrales B2, B3 y B4 que corresponden al modelo de colores RGB, lo que permite hacer una combinación para obtener imágenes en color. Por otro lado, las etiquetas representan máscaras binarias, por lo que se transforman a imágenes en blanco y negro. La siguiente imagen muestra el formato de las entradas, etiquetas y salidas del modelo (Figura 18).

[15] Más información en https://arxiv.org/pdf/1904.00592.pdf

Figura 18. Entradas, etiquetas y salidas del modelo.

La etapa de post-procesamiento considera el paso de imágenes PNG a un formato que se pueda transformar a polígonos utilizables en herramientas GIS. En este caso, se prefiere el formato TIFF y la librería PyJeo para la poligonización. Durante este proceso se deben reincorporar las referencias temporales y espaciales de las imágenes, que fueron previamente guardadas en la transformación de muestras del dataset a archivos PNG.

La figura 19 muestra los resultados de implementar el modelo ResUNet-a sobre el dataset de Ñuble en el set de entrenamiento y el set de testeo. El modelo fue entrenado por 49 épocas.

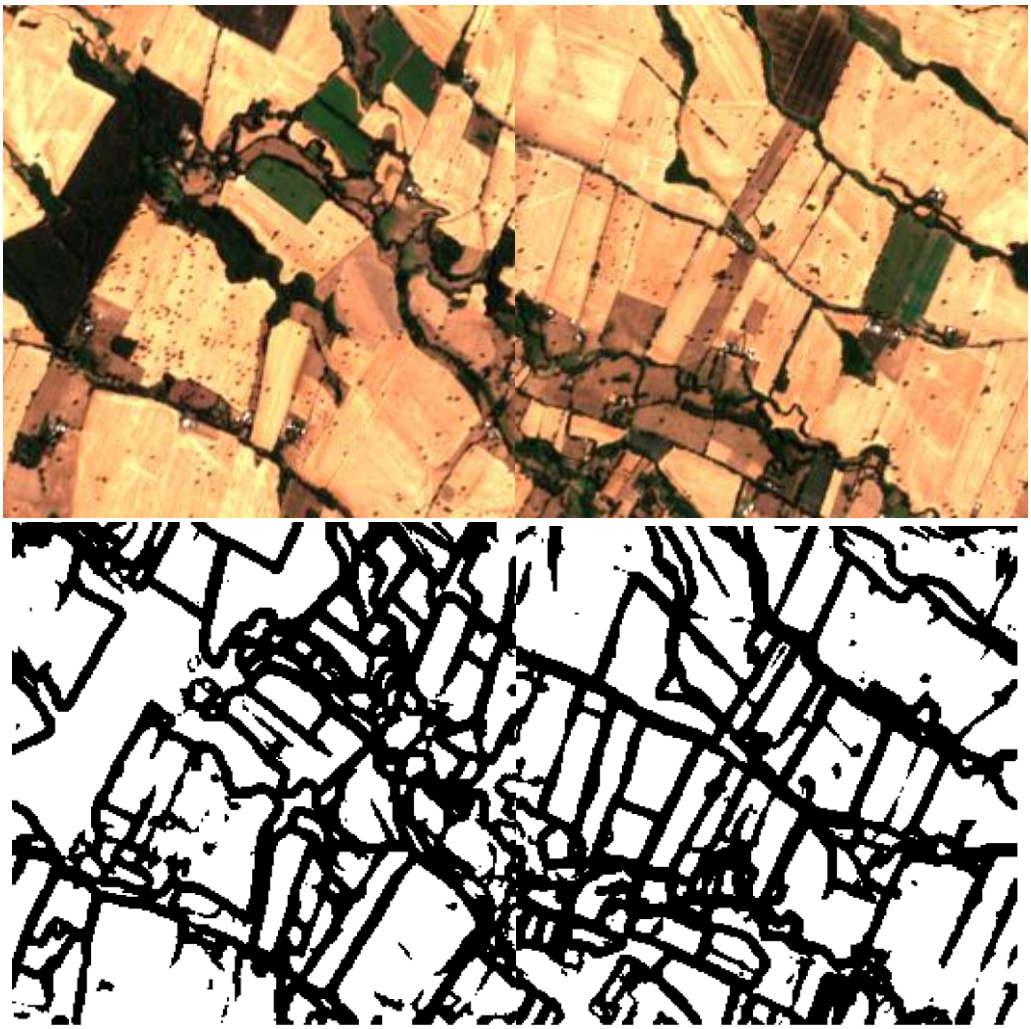

Figura 19. Resultados preliminares de segmentación y delineación. Se muestran algunos ejemplos de la delineación en diferentes muestras de Ñuble.

Se puede apreciar que se logra realizar la segmentación de las áreas de cultivo. La segmentación no es perfecta (en comparación con el desempeño del modelo original) por dos razones principales. Primero, el modelo fue entrenado con datos europeos, que no necesariamente permiten generalizar a datos chilenos. El segundo punto lo discutimos en el siguiente párrafo.

El entrenamiento completo de este modelo requiere varios días y una considerable cantidad de cómputo. En la documentación se señala que entrenaron el modelo sobre 4 GPUs NVIDIA Tesla P100, tamaño de batch de 256 y entrenamiento distribuido. Actualmente, la infraestructura disponible permite que cada época de entrenamiento con tamaños de batches de 8 ejemplos tenga una duración de una hora sobre el dataset. Esto significa que requiere 5 días para un entrenamiento de 120 épocas. La versión actual del modelo corresponde a 70 épocas, con una exactitud de 88,13% en datos de entrenamiento y 81.35% en datos de validación. El modelo original fue entrenado durante 300 épocas. Por lo tanto, se estima que más datos locales y más tiempo de cómputo, pueden sustantivamente mejorar la predicción del modelo.

3.2.4. Inferencia con el modelo ResUNet-a

Los pesos del modelo entregado al final de la etapa de entrenamiento se utilizan para la etapa de inferencia sobre datos de Chile. La inferencia comienza por descargar los archivos SAFE desde el copernicus hub. Como ejemplo nos centraremos en los datos disponibles de zonas de la región del Ñuble. Estos datos tienen que transformarse en formato NetCDF4, pesan 12GB cada uno y es importante notar que hay que aplicar sobre ellos el procesamiento mencionado en el punto 3.2.2, que asegura que el porcentaje de nubosidad no impida el uso de las muestras. Debido al tamaño de cada archivo, esto toma alrededor de 24 horas para una celda completa de Sentinel-2. Para manejar la memoria disponible para generar la inferencia, se deben dividir las imágenes en "subtiles". El tamaño de estas "subtiles" va a depender de la memoria disponible para la inferencia (i.e., propiedad del hardware). La figura 20a muestra los resultados de la inferencia del modelo ResUNet-a[16] que, tomando como input una imagen .png de 256x256 pixeles, entrega como salida una máscara de valores binarios (0/1) de mismo tamaño.

Figura 20a. Inferencia con el modelo ResUNet-a. La primera imagen es la entrada del modelo. La imagen derecha, es la máscara binaria resultante de binarizar las salidas del modelo.

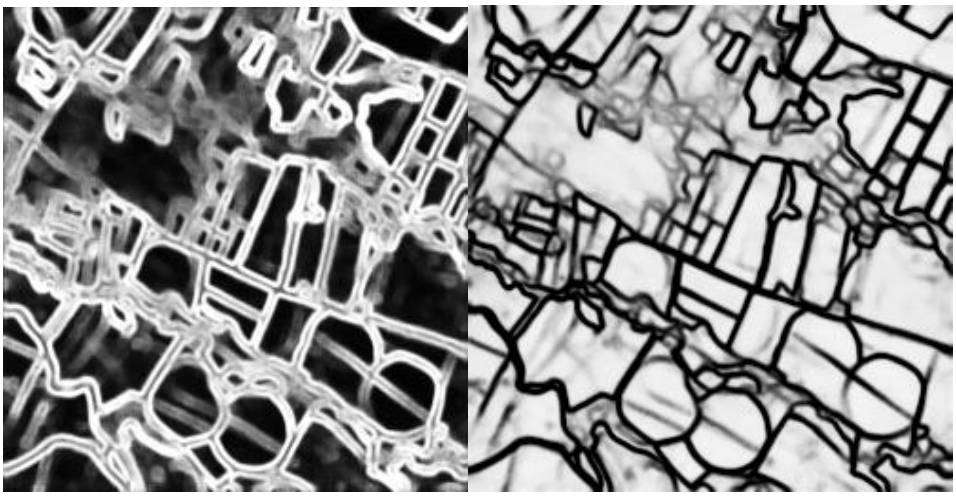

Figura 20b. Máscaras en el rango 0 a 1 de ResUNet-a. La imagen izquierda es la máscara de "boundary" o bordes. La imagen de la derecha es la máscara de "extent" o extensión.

[16] Se utilizar el código de una implementación no oficial, disponible en repositorio Github: https://github.com/Akhilesh64/ResUnet-a

A este nivel se generan tres pasos de post-procesamiento sobre la máscara compuesta de números binarios, lo cual permite generar la poligonización necesaria para el proceso de clasificación de tipo de cultivo (que presentamos en el próximo punto). Primero, se aplica un método de cutoff que aplica un umbral de binarización a las máscaras se bordes y segmentación del modelo; la figura 20b muestra un ejemplo de estas máscaras. Luego se combinan con una función que resta los bordes a la segmentación, que corresponde a la imagen binaria de la figura 20a. El objetivo es discriminar entre la parcela y el "fondo", generando una máscara en formato .tiff y lograr tener separaciones limpias entre parcelas para la poligonización. Ajustar el umbral es necesario para capturar con precisión los límites de la delineación. Este paso es importante, ya que afecta directamente la precisión y la exactitud del proceso de delineación. Segundo, se restaura el georeferenciamiento de los .tiff usando los archivos .nc originales. Tercero, y paso final en la inferencia de la delineación de parcelas, se genera la poligonización de las parcelas. Estos polígonos recopilan información de la delineación de parcelas en formato vectorizado, y almacenados en un archivo shapefile. Un archivo shapefile es un formato digital vectorial de almacenamiento de datos geográficos. Los archivos shapefile son utilizados para describir entidades geográficas como puntos, líneas y polígonos, lo que incluye, en nuestro caso, la localización (i.e., coordenadas geográficas) de parcelas (polígonos).

3.3. Clasificación de tipo de cultivos (Paso 4)

3.3.1. Preprocesamiento de datos satelitales

Para generar la clasificación de tipo de cultivos (CTC), se empieza por descargar los archivos SAFE por el periodo de interés. Para entrenar el modelo de clasificación, descargamos los archivos SAFE de Sentinel-2 desde el 01/02/2021 al 29/01/2021 de los tiles de Ñuble mencionados en el punto. Como mencionado previamente, los archivos SAFE contienen información sobre las bandas B2, B3, B4, B8, y SCL, que serán necesarios para la clasificación.

En un siguiente paso, y similar al pre-procesamiento efectuado para el modelo de delineación de parcelas, se dividen las imágenes del archivo SAFE en sub-imágenes de tamaño adecuado para no sobrepasar la capacidad de memoria del sistema de procesamiento. Además, estas imágenes pasan por un procesamiento de "limpieza" de datos similar al descrito en el punto 3.2.2.

3.3.2. Transformación de SAFE a Xarray

Una vez el preprocesamiento finalizado, se necesita generar una matriz de datos en formato Xarray[17]. Esta matriz se compone de 4 dimensiones (x, y, tiempo, banda ID). En conjunto, las dos primeras dimensiones representan valores en un plano (x, y) que define la imagen. La tercera dimensión representa información temporal (i.e., la marca de tiempo del dato asociado). La cuarta dimensión representa la banda espectral. La figura 21 representa la estructura del archivo Xarray necesario para el entrenamiento del modelo CTC.

[17] https://docs.xarray.dev/en/stable/

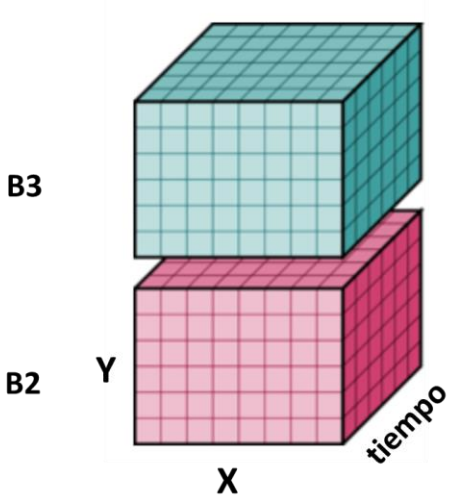

Figura 21. Representación de las 4 dimensiones incluidas en el archivo Xarray: ejes x, y de la imagen, el tiempo, y las bandas espectrales.

3.3.3. Extracción de características

Para extraer las características necesarias al entrenamiento e inferencia del modelo CTC, hicimos uso de los archivos Shapes que contienen la información con respecto a la localización de las parcelas (i.e., polígonos vectorizados). Primero, los valores asociados a cada banda fueron promediados al nivel de cada parcela, y por cada marca de tiempo. En otras palabras, se reduce el archivo Xarray a 2 dimensiones (tiempo, banda); un tiempo de serie por banda. Segundo, se aplicó un proceso de ventana deslizante (rango = 30 días, paso = 15 días). En esta ventana, se extrajo 7 estadísticas descriptivas de los tiempos de series: media, desviación estándar, primer cuartil, mediana, tercer cuartil, mínimo y máximo. Por lo tanto, en el caso que los datos agreguen información sobre 1 año, y que se usan 4 bandas espectrales (o compositivos de bandas espectrales), se generan 7 * 4 = 28 características por ventana, generando 700 características por año. En el margen de este proyecto usamos las bandas B2/B3/B4/B8 (i.e., las bandas S2 con resolución de 10 m/pixel).

3.3.4. Descripción, entrenamiento e inferencia con del modelo de clasificación de tipo de cultivo

Para clasificar el tipo de cultivo hicimos uso del modelo ilustrado en la figura 22. En su esencia, este modelo es una red neuronal recurrente (RNN) compuesta de células “long-short term memory” (LSTM). Las particularidades de este modelo son su bi-directionalidad y mecanismos de atención. Esto significa que para clasificar el tipo de cultivo en una parcela, el modelo incorpora información pasada, presente y futura, aprendiendo cuáles características temporales son más valiosas para una clasificación correcta (la parte de atención)[18].

[18] Una descripción formal (matemática) del modelo se encuentra aquí: https://arxiv.org/pdf/2208.10838.pdf.

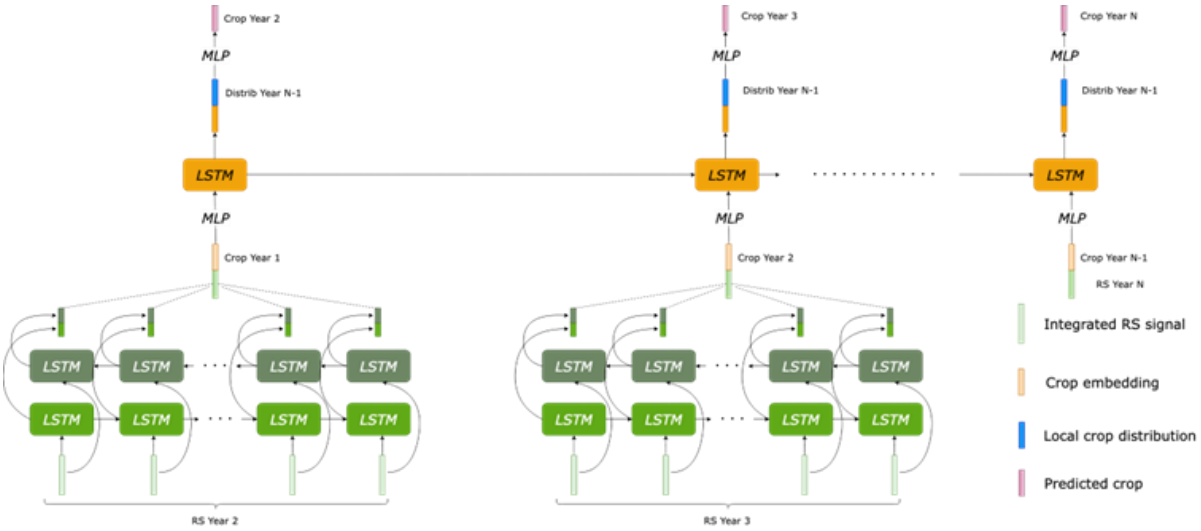

Figura 22. Arquitectura del modelo CTC.

Para entrenar el modelo, se recopiló información con ODEPA, y se creó un dataset de polígonos anotados en la región de Ñuble (Figura 9), descrito en el punto 3.1.2.2 (ver tabla 7). El dataset se compone de polígonos geolocalizados, las extracciones de características (punto 3.3.3), como inputs, y las etiquetas de tipo de cultivo, como output. El modelo está entrenado de la misma manera que un modelo de lenguaje, i.e., su tarea es predecir cuál es el tipo de cultivo debido a información previa[19].

La figura 23 muestra los resultados del modelo CTC en el test set del dataset creado para entrenarlo sobre la clasificación bi-clase de trigo contra el resto. En la figura se puede apreciar que el valor f-1, i.e., la media armónica de la precisión y recolección (recall), es de 0.891. Esto significa que el modelo de CTC clasifica las parcelas de trigo con un rendimiento de ~90%. En particular, una detección de verdadero positivo (la precisión), llega a un nivel de ~96%. Esto significa que, dentro de la distribución de datos de entrenamiento, este modelo es capaz de detectar el 96.1% de las parcelas de trigos, y por ende de la superficie real de trigo. Por lo tanto, con un dataset de entrenamiento suficientemente amplio y variado, reflejando la realidad del terreno chileno, este modelo podría dar excelentes resultados de clasificación.

	precision	recall	f1-score	support
other crops	0.992	0.998	0.995	1283
unspecified_season_common_soft_wheat	0.961	0.831	0.891	59
accuracy			0.991	1342
macro avg	0.977	0.914	0.943	1342
weighted avg	0.991	0.991	0.991	1342

Figura 23. Resultados del modelo de CTC.

Para generar una inferencia sobre el tipo de cultivo en cualquier imagen, primero hay que implementar el proceso de delineación de parcela como descrito en el punto 3.2.4 que finaliza

[19] Ver ecuaciones 1, 2 y 3 en paper: https://arxiv.org/pdf/2208.10838.pdf

con un archivo shape conteniendo la información vectorizada de los polígonos detectados en la imagen. Después, uno repite el proceso de extracción de características (3.3.3) con datos históricos (el rango temporal de los datos es una variable importante, con más rango mejor predicción). Como explicado en el párrafo precedente, los tiempos de series de características se usan como input al modelo, que genera una clasificación del tipo de cultivo por parcela.

3.3.5. Estimación de superficie de trigo plantado

La estimación de la superficie de trigo plantado es relativamente simple y directa, y depende de los resultados de delineación y clasificación. Por cada polígono detectado y clasificado como trigo, se suma la cantidad de píxeles en el polígono. Ya que trabajamos con imágenes que tienen una resolución de 10m² por pixel, generar una estimación de superficie simplemente requiere sumar la cantidad de pixeles en los polígonos detectados como trigo, y multiplicar ese valor por 10, para tener una estimación en m². Por otra parte, la estimación de la superficie de polígonos se hace automáticamente a través de las herramientas GIS.

3.4. Estimación de producción de trigo (paso 5)

Para estimar la producción de trigo, primero analizamos los datos del programa de gestión INDAP (ver sección 3.1.2.2) en los cuales teníamos la información del área de la parcela y el rendimiento asociado (en quintal/hectárea). Luego, generamos un análisis por “bins” con respecto al área de las parcelas. En este análisis, las parcelas cayendo entre dos valores de un bin se promediaron para generar una estimación de producción en función del área de las parcelas. Esto es equivalente a generar “maximum likelihood estimates” (MLEs) de producción dentro de un rango de área de parcelas.

Nuestro análisis inicial de la relación entre el tamaño de los polígonos y rendimiento de trigo demuestra que no existe una relación clara entre tamaño y rendimiento (figura 24); por lo menos en los datos disponibles para generar este análisis.

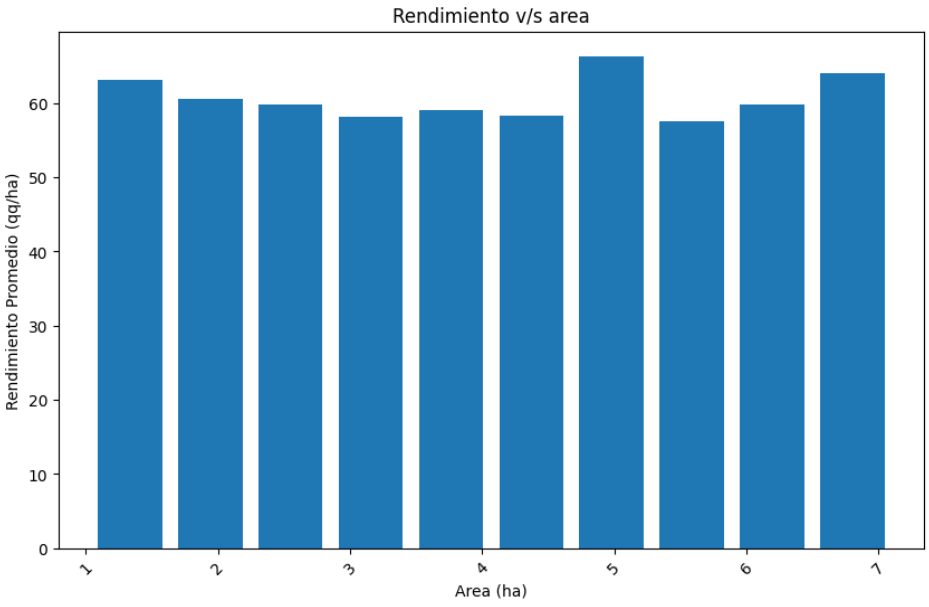

Figura 24. Rendimiento promedio por bins de hectáreas (bins = 10).

Por lo tanto, por lo que es de estos datos, la estimación de la producción se hará usando el promedio del rendimiento. Nuestro análisis indica que el rendimiento promedio es de 60.52 qq/ha. La figura 25 muestra un histograma de la frecuencia de polígonos (eje *y*) por rendimiento (en qq/ha, eje *x*). Para generar una estimación de producción en quintales, multiplicamos la estimación de superficie total de trigo en hectáreas por el rendimiento promedio.

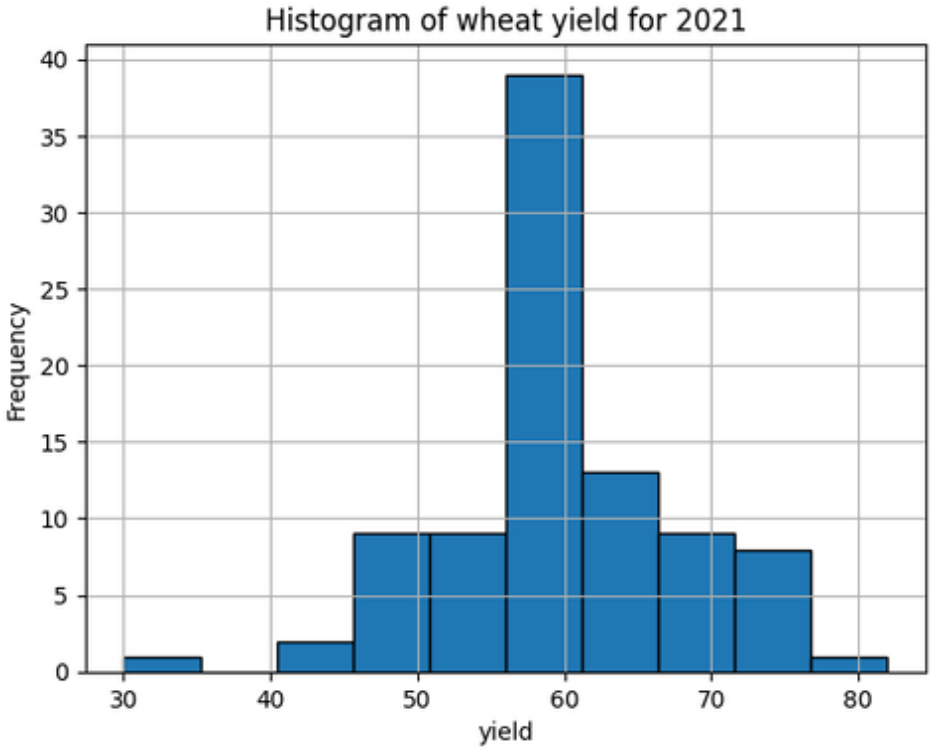

Figura 25. Histograma de rendimiento en qq/ha de las parcelas/polígonos que contienen información de rendimiento.

4. Resultados

Para generar estimaciones de estadísticas al nivel nacional, regional, y subregional, el análisis de datos se focalizó en las celdas: '19HBA', '19HBB', '19HBC', '19HCC', '19HBU', '19HBT', '19HBS', '19HBV', '18HYE', '18HXD', '18HXC', '18HXB', '18GXA', y '18GXV'. La figura 26 muestra la localización de las celdas utilizadas en la etapa de inferencia de los modelos de delineación de parcelas y clasificación de cultivos.

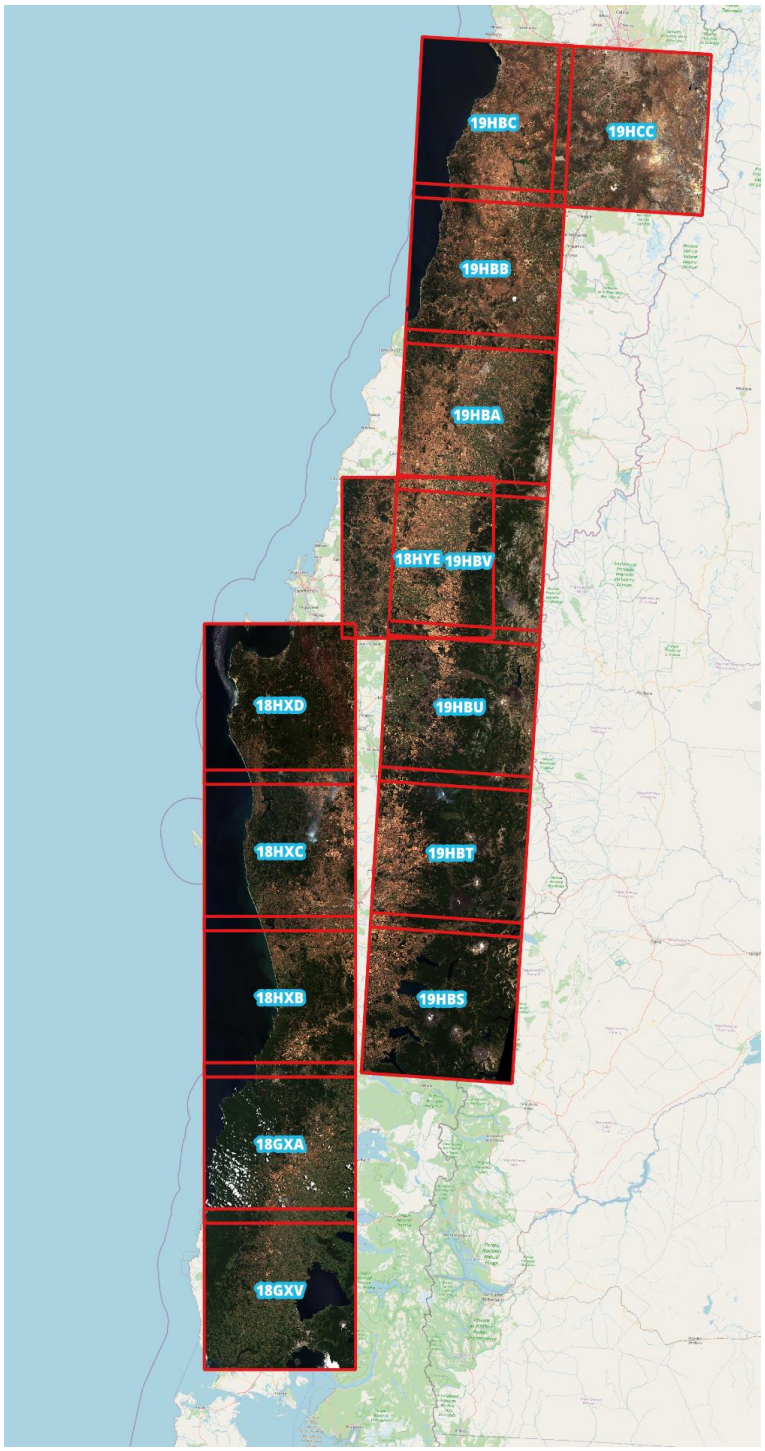

Figura 26. Celdas utilizadas en la inferencia de los modelos.

Los datos de estas celdas fueron recolectados para la temporada 2022-2023, dado que cubre la gran mayoría de las regiones de Ñuble, del Maule, del Biobío, de Araucanía. Los análisis posteriores se focalizaron en estas regiones debido a que cubren la gran mayoría de la producción de trigo en Chile. La tecnología que se entregará durante la transferencia tecnológica hará posible la generación de estimaciones estadísticas sobre cualquier celda seleccionada e introducida al sistema. Cabe mencionar que las celdas definidas por el servicio de imágenes satelitales (Sentinel-2) presentan un nivel de superposición. Este efecto podría generar una sobreestimación del número de parcelas y la superficie de ciertos cultivos. A la

luz de los resultados que se presentan posteriormente, y considerando que los modelos se encuentran en una etapa preliminar de entrenamiento (en particular el modelo de delineación de parcelas), no se considera relevante resolver este detalle cartográfico en esta etapa.[20]

En lo que sigue, separamos la información a nivel regional y subregional, por las tres métricas de interés: número de paralelas de trigo, superficie de trigo, estimación de producción. Para una visualización más intuitiva de los resultados, proporcionamos figuras ilustrando "mapas de calor" de las tres métricas de interés. Además, proporcionamos tablas con los valores exactos.

En términos de delineación de parcelas, el modelo ResUNet-a en conjunto con el modelo de CTC detectaron 12.127 parcelas de trigo a nivel nacional (definido por las regiones en la figura 26). La tabla 5 (primera columna) contiene la distribución del número de parcelas por región.

Región	Cantidad de parcelas	Superficie (ha)	Producción (qq)
Maule	1.466	1.482	89.692
Ñuble	10.126	13.413	811.776
Biobío	426	356	21.553
Araucanía	109	81	4.871

Tabla 5. Estimación de superficie y producción de trigo por región.

La figura 27 representa esta información a través de un mapa de calor. Del análisis de esta figura, se puede deducir que los modelos estiman que la región de Ñuble presenta la mayor cantidad de parcelas de trigo.

[20] Para estimaciones futuras, este detalle se podría resolver definiendo una cuadrícula que calce con los límites de las celdas, considerando solo una de ellas en el caso que exista superposición con otras celdas.

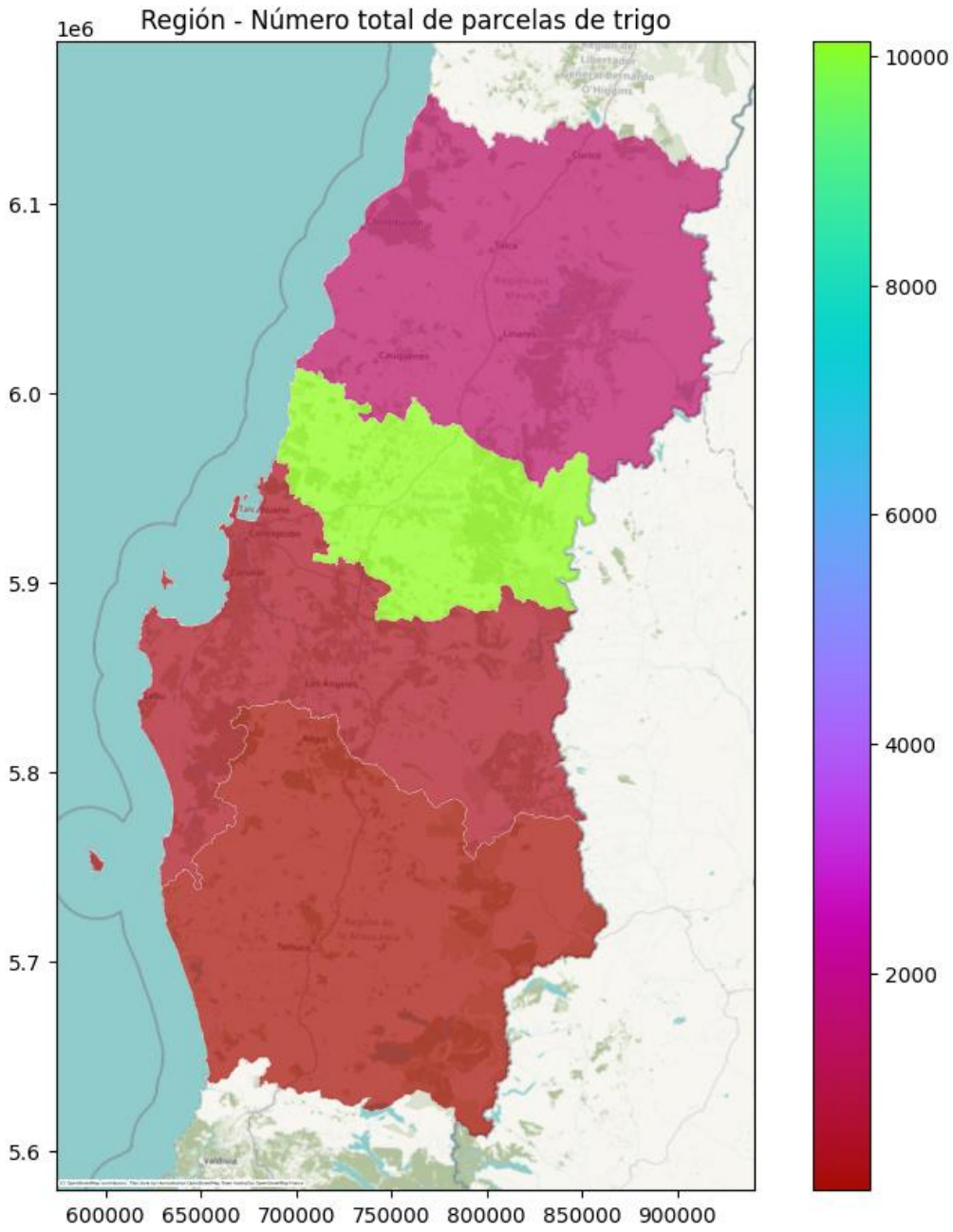

Figura 27. Mapa de color reflejando la cantidad de parcelas detectadas por región.

La superficie de parcelas de trigo, a nivel nacional, es de 15.332,02 hectáreas. La distribución por región se encuentra en la columna 2 de la tabla 5. La figura 28 representa la información de superficie a través de un mapa de calor. En términos de superficie por región, esta figura muestra que la región Ñuble contiene la mayor superficie de parcelas de trigo.

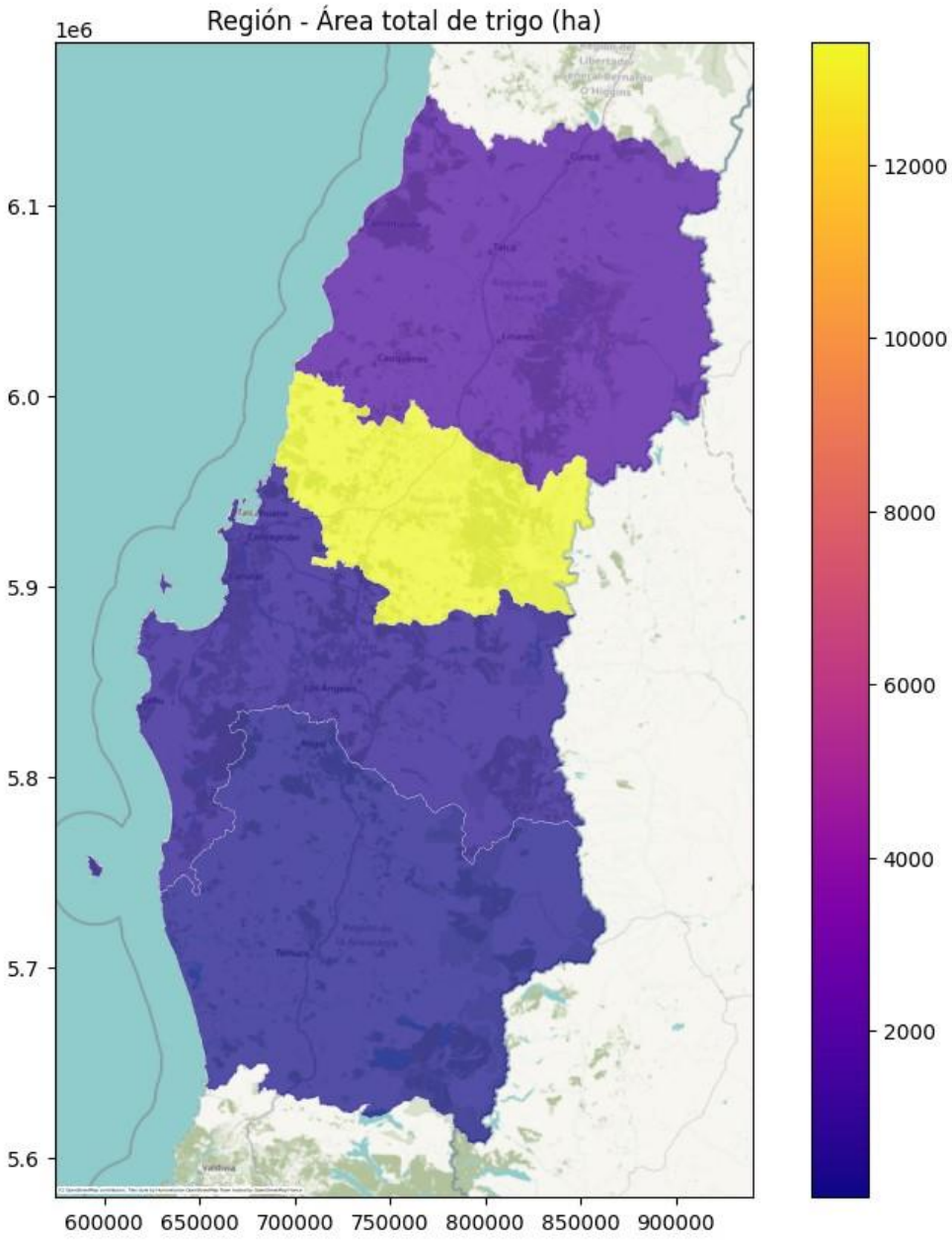

Figura 28. Mapa de color reflejando la estimación de superficie (en hectáreas) de trigo por región.

La estimación de producción a nivel nacional es de 927.892,5 quintales. La distribución por región se encuentra en la columna 3 de la tabla 5. La figura 29 representa la información de producción a través de un mapa. En términos de producción por región, esta figura muestra que la región Ñuble contiene la mayor superficie de parcelas de trigo.

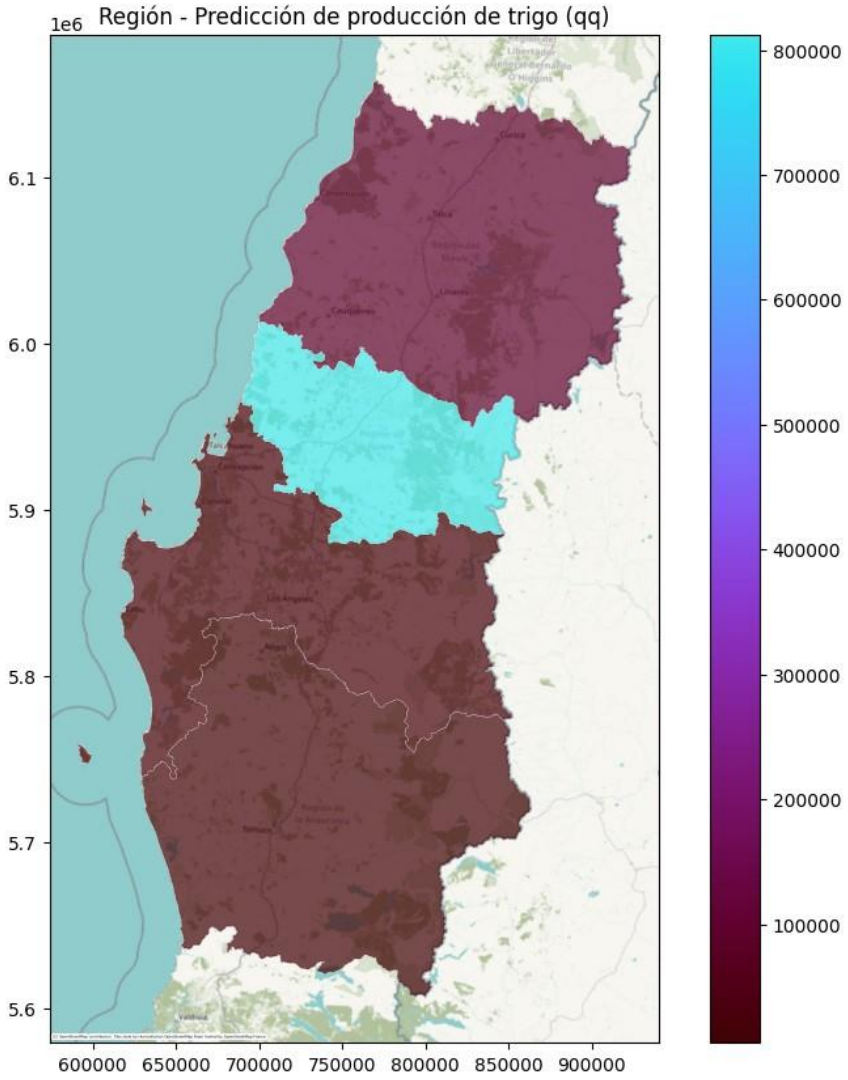

Figura 29. Mapa de color reflejando la estimación de producción (en qq) de trigo por región.

Para la estimación de la producción, ésta se genera directamente a partir de nuestra estimación total de hectáreas de trigo, multiplicando esta estimación por el promedio de rendimiento (en quintales/hectárea).

No es sorprendente que nuestros modelos estimen a Ñuble como la región que tiene mayores valores en las métricas de interés. El modelo fue entrenado con datos de esa zona, y la estimación de estas métricas fue generada con datos de la misma temporada (i.e., 2021-2022). Volvemos a discutir este punto en las limitaciones del estudio (punto 5).

La información a nivel subregional (i.e., comunal) se encuentra en la tabla A.1 (ver Anexos). En esta tabla uno puede observar los valores de número de parcelas de trigo, estimación de superficie, y producción, por cada comuna de las regiones descritas.

La información desplegada en la tabla A.1 se entrega igualmente en forma de mapa de calor. Las figuras 30, 31, 32 muestran las tres métricas de interés a nivel comunal.

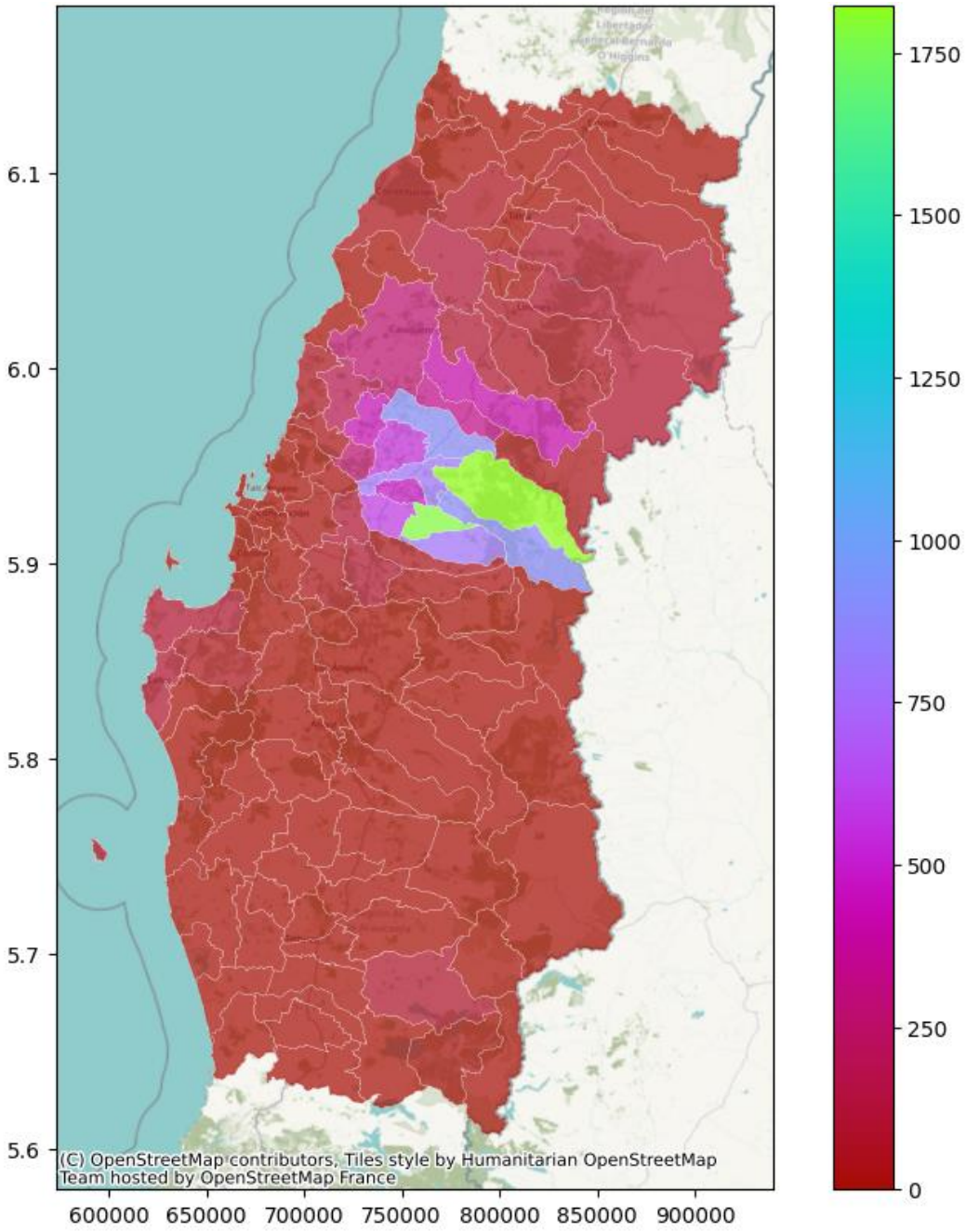

Figura 30. Mapa de calor de número de parcelas de trigo por comuna de las regiones de Ñuble, del Maule, del Biobío, de Araucanía. La barra de color representa los valores de numero de parcelas.

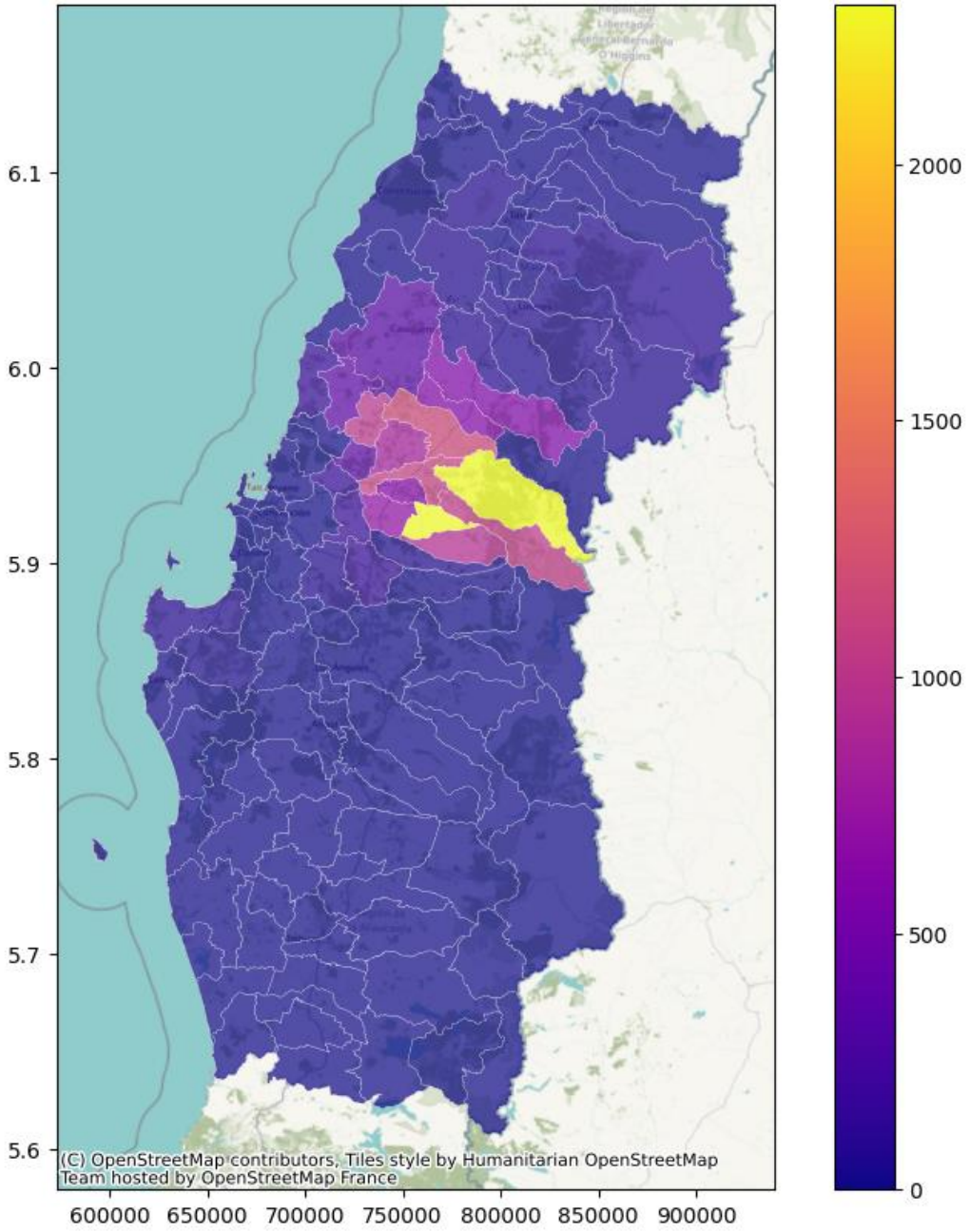

Figura 31. Mapa de calor de superficie plantada de trigo por comuna de las regiones de Ñuble, del Maule, del Biobío, de Araucanía. La barra de color representa los valores de superficie (en ha).

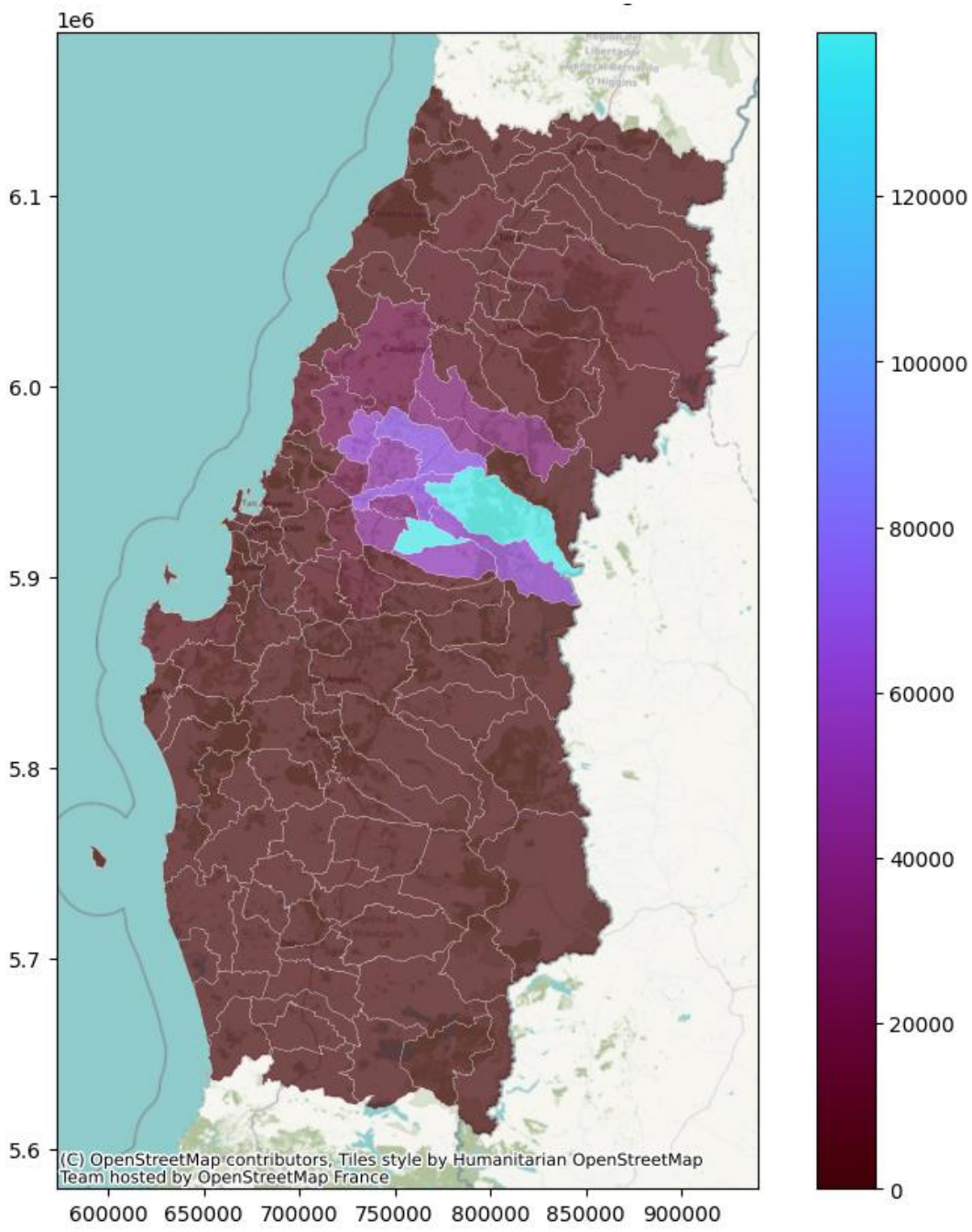

Figura 32. Mapa de calor de estimación de producción de trigo por comuna de las regiones de Ñuble, del Maule, del Biobío, de Araucanía. La barra de color representa los valores de producción (en qq).

Para completar la información analizada, ilustramos la visualización de las parcelas detectadas en las dos comunas de la región de Ñuble asociadas con la mayor cantidad de parcelas de trigo detectadas (ver figuras 33 y 34).

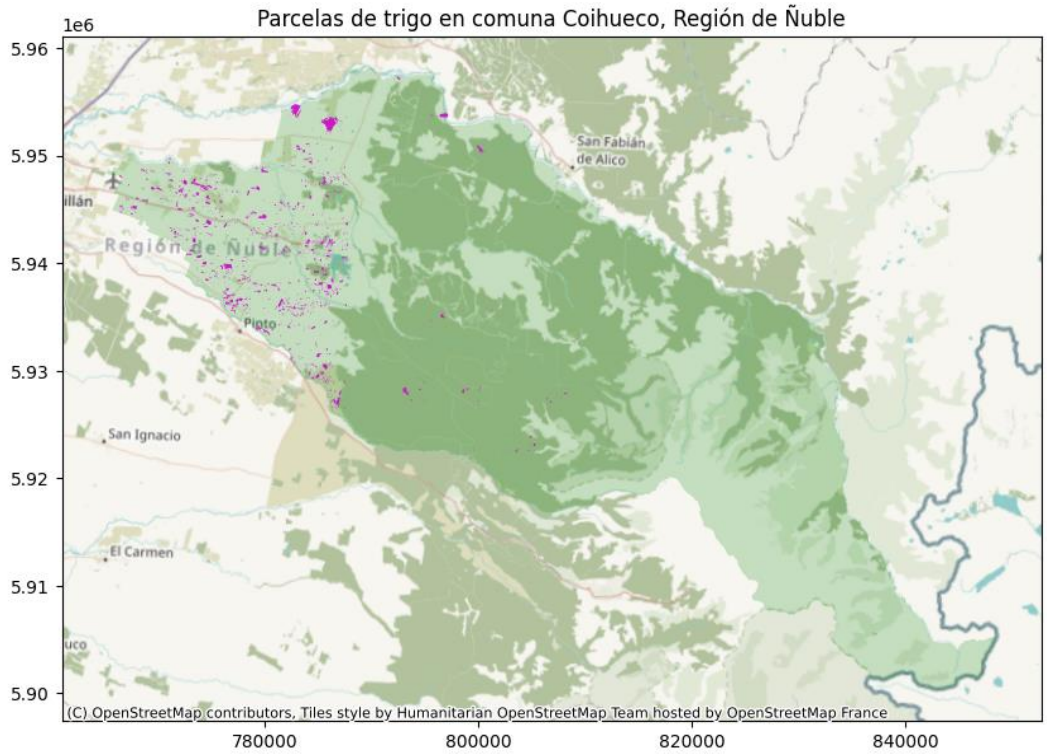

Figura 33. Parcelas detectadas en la comuna de Coihueco.

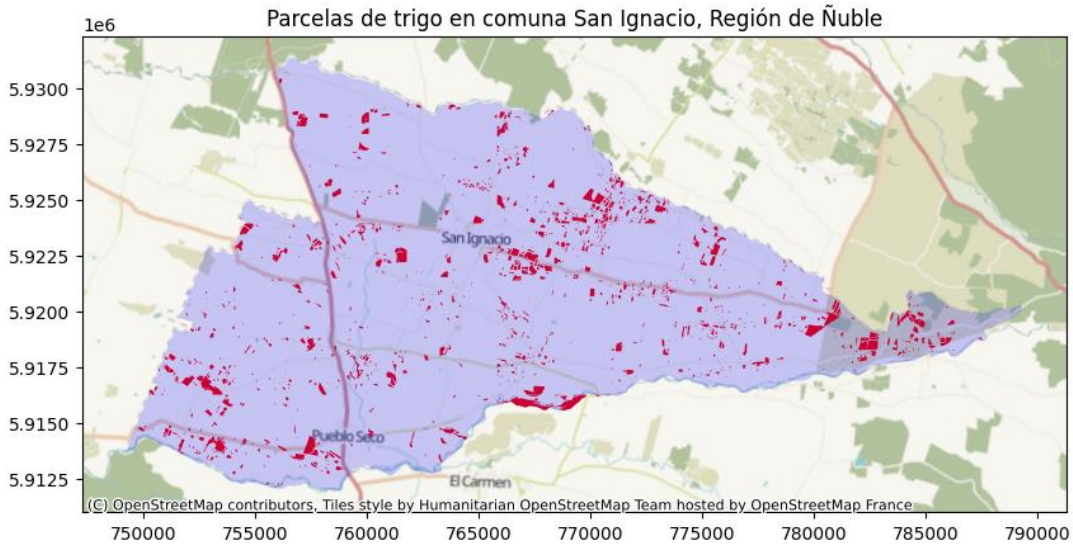

Figura 34. Parcelas detectadas en la comuna de San Ignacio.

Los resultados detallados presentados en este capítulo se disponen en formatos shapefile (incluyendo archivos data.info con la información de atributos) y csv, como material complementario a este informe.

Para efectos de contraste de los resultados obtenidos con estadísticas nacionales, se consideran las estimaciones del INE (2023) de la superficie regional de Trigo, presentadas en la tabla 7.

Región	Superficie (ha)
Coquimbo	0
Valparaíso	573
Metropolitana	5.944
O'Higgins	14.072
Maule	25.015
Ñuble	36.564
Biobío	31.440
La Araucanía	81.365
Los Ríos	7.843
Los Lagos	13.871
Resto del País	46
Total	**216.733**

Tabla 6. Estadísticas de INE (2023) de superficie de trigo.

Comparando las estadísticas de la tabla 6 con los resultados de la tabla 5 se puede apreciar que el modelo subestima la superficie de trigo para todas las regiones del país, con Ñuble la región en la que la estimación de los modelos (13.413 ha) se acerca más al valor indicado por INE (2023) (36.564 ha). Esto sugiere que la aplicación de los modelos en zonas geográficas diferentes a las empleadas durante su entrenamiento se debe considerar con cautela. Es de esperar que cada región tenga sus particularidades en términos de condiciones climáticas, prácticas agrícolas, temporadas de siembra y cosecha, variedades de cultivos, entre otros. Por lo tanto, existen riesgos y limitaciones asociadas a la generalización del modelo a regiones distintas de aquellas para las cuales fue originalmente desarrollado el piloto de este estudio. En el siguiente capítulo se abordan estas limitaciones de forma más detallada. Sin embargo, se sugiere seguir explorando las posibles causas de las diferencias entre los valores estimados y los entregados por INE (2023) de una manera más exhaustiva. Esto permitirá una comprensión más precisa de la aplicabilidad y validez del modelo en diferentes contextos, evitando extrapolaciones no fundamentadas que podrían comprometer la fiabilidad de los resultados.

5. Límites del estudio, sugerencias para estudios futuros, datos técnicos, transferencia, y protección de datos

La solución desarrollada durante este proyecto, incluyendo la metodología, desarrollo e implementación de modelos de procesamiento de imágenes satelitales, junto con los resultados obtenidos, cumplen con los objetivos del estudio. Sin embargo, estos resultados deben considerarse preliminares en el sentido que, si bien son consistentes, no cumplen satisfactoriamente con un nivel de confiabilidad y robustez adecuado. Lo anterior no se debe a deficiencias en la estructura de los modelos utilizados, si no que al acceso limitado a datos etiquetados de cultivos a nivel nacional, los cuales son la información de entrada de los modelos, utilizados durante la etapa de entrenamiento de los modelos de clasificación de cultivos. En esta sección se describen las limitaciones de este estudio asociadas a los distintos pasos de nuestra metodología, y se proponen sugerencias para poder superar estas limitaciones en futuras estimaciones.

5.1. Limitaciones asociadas a la delineación de parcelas

En el estado actual de inferencia, ResUNet-a fue entrenado con 49 épocas. Idealmente, para aumentar/maximizar el rendimiento del modelo, habría que seguir entrenando el modelo hasta llegar aproximadamente a 300 épocas (similar al paper original de Waldner y colegas[21]). Adicionalmente, lo ideal sería poder generar un dataset idéntico al del AI4boundaries pero sobre el territorio chileno. Un tal dataset permitiría aumentar el rendimiento del modelo ResUNet-a en lo que es su capacidad para delinear parcelas.

Una manera directa sería generar una plataforma similar al Registro Gráfico de Parcelas implementado por el gobierno francés (ver punto 2.2.1). Otra posibilidad es implementar una campaña de crowd-sourcing, usando aplicaciones open source como crop observe[22], desarrollada por el "International Institute for Applied Systems Analysis" (IIASA). Para llevar a cabo estas estrategias, se recomienda buscar mecanismos de colaboración con productores agrícolas para así poder comunicar de manera más efectiva los beneficios de disponer esta información, y así fomentar una mejor participación y apoyo por parte de los productores en el levantamiento de la información.

5.2. Limitaciones asociadas a la clasificación de tipo de cultivo

El modelo CTC depende altamente de la cantidad y variabilidad de clases en los datos. Para generar clasificaciones, usamos datos que no incorporan estas dos propiedades. En efecto, para generar una buena capacidad de generalización, se necesitan del orden de miles o diez miles de parcelas etiquetadas por clase (i.e., tipo de cultivo). En el caso presente, cuya meta es detectar parcelas de trigos, se pudieron recopilar 547 polígonos etiquetados como "trigo", y 12.904 como "especies frutícolas". Por lo tanto, no es posible evaluar, de manera exhaustiva, la capacidad de generalización del modelo CTC con datos que presentan una distribución distinta a los datos de entrenamiento. Por otro lado, el set de entrenamiento solo incorpora algunos tipos de cultivo, dejando fuera a cultivos que puedan presentar características agrícolas similares al trigo, como por ejemplo la avena, la cebada, o el centeno. Tener un set de entrenamiento en el cual la cantidad de clases a clasificar es baja genera aproximaciones que pueden "inflar" las predicciones. Por ejemplo, cultivos que tienen características visuales similares a las del trigo, van a tener una tendencia a ser categorizados como trigo. De la misma forma, si el set de entrenamiento está desbalanceado, i.e., si una clase está sobre representada, esto entrena un "overfitting" de esta clase, que será inflada en los datos de clasificación. Por lo cual, en estudios futuros, es fundamental poder trabajar con un dataset de entrenamiento que incorpora a una gran variedad de tipos de cultivos, con la mayor cantidad posible de polígonos etiquetados. Para esto, se puede seguir una estrategia similar a la sugerida en el punto anterior, aprovechando de levantar de forma simultánea la información correspondiente a la delineación y clasificación de cultivos con apoyo de los productores.

Como punto de comparación, el mismo modelo de clasificación descrito en este estudio fue utilizado en un dataset obtenido a través del sistema de identificación de parcelas de los

[21] https://www.sciencedirect.com/science/article/abs/pii/S0034425720301115
[22] https://play.google.com/store/apps/details?id=com.iiasa.cropobserve&hl=en&gl=US

países bajos ('basis registratie percelen' o BRP). En este sistema, los agricultores holandeses tienen que subir información sobre la delineación de sus parcelas y tipo de cultivo. El set de datos fue recopilado sobre 12 años de datos (2009-2020), llegando a una cantidad de 974,000 polígonos (i.e., parcelas) anotados, cubriendo una área de 1,600 Mha[23]. Las anotaciones fueron reagrupadas en 10 clases. En base de este set de datos de 974,000 polígonos, el modelo usado en este estudio generó un rendimiento de 90% de exactitud en la clasificación, y una predicción equivalente a 80% de exactitud 5 meses previo a la cosecha. Por tanto, es importante considerar que la utilidad de este modelo ya fue demostrada con otro set de datos, y que los resultados sesgados en este estudio provienen únicamente de la limitación en cantidad de polígonos anotados y cantidades de clases. Para generar rendimientos similares en Chile, usando un modelo entrenado desde cero, se necesita una cantidad similar de datos, i.e., del orden del millón de polígonos anotados. Eso equivale a 100 veces más polígonos de los que fueron usados para entrenar este modelo, y con una mayor distribución de clase.

Una alternativa potencial a la recopilación de 1.000.000 de polígonos anotados, sería de inicialmente entrenar el modelo sobre datos accesible en zonas con terrenos y condiciones meteorológicas similares a las de chile, quizás a través de convenios gubernamentales, y subsecuentemente afinar ("fine-tune") el modelo con la mayor cantidad posible de datos recopilados sobre el terreno chileno.

5.3. Limitaciones asociadas a la estimación de producción

En el marco de este estudio, no se pudo trabajar con métodos de machine learning para generar una estimación de producción. No se pudo recopilar una cantidad de datos suficiente para implementar métodos de visión por computadora. En el contexto de este proyecto, 91 puntos de datos (i.e., polígonos) pudieron ser recopilados que disponían de información de superficie y rendimiento. En estudios futuros, se podría mejorar la predicción de producción usando datos multimodales, incluyendo, además de imágenes satelitales, variables agrícolas que afectan el rendimiento de los cultivos (ej: datos meteorológicos, terreno, calidad de suelo, etc.).

5.4. Limitaciones metodológicas

Es importante entender que el modelo de clasificación de cultivos fue entrenado únicamente sobre datos de la región de Ñuble para el periodo 2021-2022. En cambio, para el proceso de inferencia, se utilizaron imágenes del periodo 2021-2022 sólo para la región de Ñuble. Para las otras tres regiones (Maule, Biobío y Araucanía), se utilizaron imágenes de la temporada 2022-2023. De esta decisión metodológica podemos generar dos recomendaciones cruciales. Primero, se necesita significativamente agrandar el set de datos etiquetados tanto al nivel de instancias, de clases de cultivo, y de localización espacial. Segundo, es importante poder contar con datos etiquetados que cubran varios años de recolección.

[23] https://arxiv.org/pdf/2208.10838.pdf

5.5. Datos técnicos

Para maximizar la eficiencia de entrenamiento e inferencia de los modelos, se sugiere trabajar con un sistema que tenga 100G de RAM, GPU RTX-8000, y 32 Cores.

5.6. Transferencia

La transferencia tecnológica se facilita a través de manuales de uso, incluidos como material complementario a este informe. Estos manuales incluyen:

- Requerimientos del modelo PBD
- Pipeline (flujo de trabajo) para inferencia modelo PBD
- Pipeline (flujo de trabajo) para inferencia modelo CTC

Estos puntos se explicarán a través del análisis del Github open-acces (https://github.com/CENIA-DEV/agro-satelite) que hemos creado para replicar los procesos necesarios a la generación de los resultados presentados en este informe.

5.7. Protección de datos

Cualquier publicación de datos seguirá el “General Data Protection Regulation (GDPR) guidelines”, implementado en la Unión Europea, la cual asegura una anonimidad total de los datos publicados.

6. Referencias

[1] E. Saralioglu and O. Gungor, “Crowdsourcing in remote sensing: A review of applications and future directions,” IEEE Geosci Remote Sens Mag, vol. 8, no. 4, pp. 89–110, 2020.

[2] J. Minet et al., “Crowdsourcing for agricultural applications: A review of uses and opportunities for a farmsourcing approach,” Comput Electron Agric, vol. 142, pp. 126–138, 2017.

[3] S. Talukdar, P. Singha, S. Mahato, S. Pal, Y.-A. Liou, and A. Rahman, “Land-use land-cover classification by machine learning classifiers for satellite observations—A review,” Remote Sens (Basel), vol. 12, no. 7, p. 1135, 2020.

[4] J. Minet, B. Robert, and B. Tychon, “The potential of OpenStreetMap for land use/land cover mapping,” in FOSS4G Belgium 2015, 2015.

[5] S. Fritz et al., “Geo-Wiki: An online platform for improving global land cover,” Environmental Modelling & Software, vol. 31, pp. 110–123, 2012.

[6] L. Estes et al., “Diylandcover: Crowdsourcing the creation of systematic, accurate landcover maps,” PeerJPrePrints, vol. 3, p. e1266, 2015.

[7] A. Bey et al., “Collect earth: Land use and land cover assessment through augmented visual interpretation,” Remote Sens (Basel), vol. 8, no. 10, p. 807, 2016.

[8] S. Crommelinck, R. Bennett, M. Gerke, F. Nex, M. Y. Yang, and G. Vosselman, “Review of automatic feature extraction from high-resolution optical sensor data for UAV-based cadastral mapping,” Remote Sensing, vol. 8, no. 8. MDPI AG, 2016. doi: 10.3390/rs8080689.

[9] S. A. Ramadhani, R. M. Bennett, and F. C. Nex, “Exploring UAV in Indonesian cadastral boundary data acquisition,” Earth Sci Inform, vol. 11, pp. 129–146, 2018.

[10] G. Mountrakis, J. Im, and C. Ogole, “Support vector machines in remote sensing: A review,” ISPRS journal of photogrammetry and remote sensing, vol. 66, no. 3, pp. 247–259, 2011.

[11] N. Yager and A. Sowmya, “Support vector machines for road extraction from remotely sensed images,” in Computer Analysis of Images and Patterns: 10th International Conference, CAIP 2003, Groningen, The Netherlands, August 25-27, 2003. Proceedings 10, Springer, 2003, pp. 285–292.

[12] O. Vlachopoulos, B. Leblon, J. Wang, A. Haddadi, A. LaRocque, and G. Patterson, “Delineation of bare soil field areas from unmanned aircraft system imagery with the mean shift unsupervised clustering and the random forest supervised classification,” Canadian Journal of Remote Sensing, vol. 46, no. 4, pp. 489–500, 2020.

[13] O. Vlachopoulos, B. Leblon, J. Wang, A. Haddadi, A. LaRocque, and G. Patterson, “Delineation of crop field areas and boundaries from UAS imagery using PBIA and GEOBIA with random forest classification,” Remote Sens (Basel), vol. 12, no. 16, p. 2640, 2020.

[14] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Ger-many, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[15] A. Garcia-Pedrero, M. Lillo-Saavedra, D. Rodriguez-Esparragon, and C. Gonzalo-Martin, “Deep learning for automatic outlining agricultural parcels: Exploiting the land parcel identification system,” IEEE access, vol. 7, pp. 158223–158236, 2019.

[16] V. S. F. Garnot and L. Landrieu, “Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks,” Proceedings of the IEEE International Conference on Computer Vision, pp. 4852–4861, 2021, doi: 10.1109/ICCV48922.2021.00483.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 630–645.

[18] R. d’Andrimont et al., “AI4Boundaries: an open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography,” Earth Syst Sci Data, vol. 15, no. 1, pp. 317–329, 2023, doi: 10.5194/essd-15-317-2023.

[19] F. Waldner et al., “Detect, consolidate, delineate: Scalable mapping of field boundaries using satellite images,” Remote Sens (Basel), vol. 13, no. 11, p. 2197, 2021.

[20] A. Orynbaikyzy, U. Gessner, and C. Conrad, “Crop type classification using a combination of optical and radar remote sensing data: A review,” Int J Remote Sens, vol. 40, no. 17, pp. 6553–6595, 2019.

[21] S. Feng, J. Zhao, T. Liu, H. Zhang, Z. Zhang, and X. Guo, “Crop type identification and mapping using machine learning algorithms and sentinel-2 time series data,” IEEE J Sel Top Appl Earth Obs Remote Sens, vol. 12, no. 9, pp. 3295–3306, 2019.

[22] S. Wang, G. Azzari, and D. B. Lobell, “Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques,” Remote Sens Environ, vol. 222, pp. 303–317, 2019.

[23] Z. Ma et al., “An unsupervised crop classification method based on principal components isometric binning,” ISPRS Int J Geoinf, vol. 9, no. 11, p. 648, 2020.

[24] Q. Song et al., “In-season crop mapping with GF-1/WFV data by combining object-based image analysis and random forest,” Remote Sens (Basel), vol. 9, no. 11, p. 1184, 2017.

[25] N. Farmonov et al., “Crop type classification by DESIS hyperspectral imagery and machine learning algorithms,” IEEE J Sel Top Appl Earth Obs Remote Sens, vol. 16, pp. 1576–1588, 2023.

[26] V. Barriere and M. Claverie, “Multimodal Crop Type Classification Fusing Multi-Spectral Satellite Time Series with Farmers Crop Rotations and Local Crop Distribution,” in Proceedings of 2nd Workshop on Complex Data Challenges in Earth Observation, IJCAI, 2022, pp. 50–57.

[27] P. Muruganantham, S. Wibowo, S. Grandhi, N. H. Samrat, and N. Islam, “A systematic literature review on crop yield prediction with deep learning and remote sensing,” Remote Sens (Basel), vol. 14, no. 9, p. 1990, 2022.

[28] J. Ansarifar, L. Wang, and S. V Archontoulis, “An interaction regression model for crop yield prediction,” Sci Rep, vol. 11, no. 1, p. 17754, 2021.

[29] R. Bhatnagar and G. B. Gohain, “Crop yield estimation using decision trees and random forest machine learning algorithms on data from terra (EOS AM-1) & Aqua (EOS PM-1) satellite data,” Machine Learning and Data Mining in Aerospace Technology, pp. 107–124, 2020.

[30] R. Li et al., “Winter wheat yield estimation based on support vector machine regression and multi-temporal remote sensing data.,” Transactions of the Chinese Society of Agricultural Engineering, vol. 25, no. 7, pp. 114–117, 2009.

[31] K. Kuwata and R. Shibasaki, “Estimating crop yields with deep learning and remotely sensed data,” in 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2015, pp. 858–861.

[32] K. Alibabaei, P. D. Gaspar, and T. M. Lima, “Crop yield estimation using deep learning based on climate big data and irrigation scheduling,” Energies (Basel), vol. 14, no. 11, p. 3004, 2021.

[33] H. Yalcin, “An approximation for a relative crop yield estimate from field images using deep learning,” in 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), 2019, pp. 1–6.

[34] T. Van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” Comput. Electron. Agric., vol. 177, p. 105709, 2020.

[35] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and F. B. Fritschi, “Soybean yield prediction from UAV using multimodal data fusion and deep learning,” Remote Sens Environ, vol. 237, p. 111599, 2020.

[36] A. Kaur, P. Goyal, K. Sharma, L. Sharma, and N. Goyal, “A Generalized Multi-modal Deep Learning Model for Early Crop Yield Prediction,” in 2022 IEEE International Conference on Big Data (Big Data), 2022, pp. 1272–1279.

[37] T. Sakamoto, “Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm,” ISPRS Journal of Photogrammetry and Re-mote Sensing, vol. 160, pp. 208–228, 2020.

[38] M. Marszalek, M. Körner, and U. Schmidhalter, “Prediction of multi-year winter wheat yields at the field level with satellite and climatological data,” Comput. Electron. Agric., vol. 194, p. 106777, 2022.

[39] M. A. Peña and A. Brenning, “Assessing fruit-tree crop classification from Landsat-8 time series for the Maipo Valley, Chile,” Remote Sens Environ, vol. 171, pp. 234–244, 2015.

[40] M. A. Peña, R. Liao, and A. Brenning, “Using spectrotemporal indices to improve the fruit-tree crop classification accuracy,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 128, pp. 158–169, 2017.

[41] F. Zambrano, A. Vrieling, A. Nelson, M. Meroni, and T. Tadesse, “Prediction of drought-induced reduction of agricultural productivity in Chile from MODIS, rainfall estimates, and climate oscillation indices,” Remote Sens Environ, vol. 219, pp. 15–30, 2018.

[42] Centro de Investigación e Innovación de la Viña Concha y Toro. (2018, March 1). Inteligencia artificial y drones: Nuevo Sistema de Predicción del Volumen de Cosecha. Viña Concha y Toro. https://vinacyt.com/noticia/innovacion/inteligencia-artificial-y-drones-nuevo-sistema-de-prediccion-del-volumen-de-cosecha-2/

7. Anexos

En la tabla A.1 se presentan los resultados a nivel sub-regional (número de parcelas, superficie y producción estimada de trigo por comuna) para las regiones de Maule, Ñuble, Bio-Bio y Araucanía.

Nombre Región	Nombre Comuna	Cantidad de parcelas	Superficie (ha)	Producción (qq)
Región del Maule	Talca	13	4.2	253.2
Región del Maule	Constitución	43	20.8	1259.6
Región del Maule	Curepto	11	3.4	208.5
Región del Maule	Empedrado	3	0.1	7.1
Región del Maule	Maule	11	3.0	178.7
Región del Maule	Pelarco	42	26.9	1629.1
Región del Maule	Pencahue	57	13.3	804.1
Región del Maule	Río Claro	11	4.6	281.0
Región del Maule	San Clemente	78	38.6	2333.8
Región del Maule	San Rafael	7	1.3	77.8
Región del Maule	Cauquenes	267	463.9	28072.4
Región del Maule	Hualañé	4	0.9	55.8
Región del Maule	Licantén	6	6.3	381.4
Región del Maule	Molina	32	44.5	2690.6
Región del Maule	Rauco	5	4.3	259.7
Región del Maule	Vichuquén	1	2.4	147.0
Región del Maule	Linares	63	39.9	2412.4
Región del Maule	Colbún	87	40.5	2452.0
Región del Maule	Longaví	62	55.7	3369.7
Región del Maule	Parral	408	488.1	29542.6
Región del Maule	Retiro	106	101.6	6150.9
Región del Maule	San Javier	114	92.9	5621.9
Región del Maule	Villa Alegre	11	14.9	901.0
Región del Maule	Yerbas Buenas	24	9.9	601.6
Región de Ñuble	Chillán	789	1123.2	67974.8
Región de Ñuble	Bulnes	584	722.1	43699.0
Región de Ñuble	Chillán Viejo	481	734.0	44421.1
Región de Ñuble	El Carmen	779	1253.5	75863.5
Región de Ñuble	Pemuco	52	72.8	4408.0
Región de Ñuble	Pinto	887	1261.6	76352.7
Región de Ñuble	Quillón	113	85.3	5159.5
Región de Ñuble	San Ignacio	1781	1752.9	106082.7
Región de Ñuble	Yungay	24	24.7	1494.3
Región de Ñuble	Quirihue	183	437.8	26496.3
Región de Ñuble	Cobquecura	52	38.6	2336.7
Región de Ñuble	Coelemu	9	15.2	919.7

Región de Ñuble	Ninhue	437	838.0	50717.1
Región de Ñuble	Portezuelo	271	224.0	13556.2
Región de Ñuble	Ranquil	16	48.6	2943.2
Región de Ñuble	Treguaco	40	44.4	2688.9
Región de Ñuble	San Carlos	893	1843.8	111584.3
Región de Ñuble	Coihueco	1822	1697.2	102713.0
Región de Ñuble	Ñiquén	331	387.3	23440.9
Región de Ñuble	San Fabián	38	46.3	2804.7
Región de Ñuble	San Nicolás	544	762.1	46119.3
Región del Biobío	Florida	17	10.0	605.9
Región del Biobío	Tomé	2	67.1	4063.1
Región del Biobío	Lebu	96	12.1	730.0
Región del Biobío	Arauco	98	22.5	1358.8
Región del Biobío	Cañete	4	0.4	21.6
Región del Biobío	Curanilahue	44	7.1	429.7
Región del Biobío	Los Alamos	10	0.7	43.0
Región del Biobío	Antuco	3	0.2	11.6
Región del Biobío	Cabrero	98	210.7	12753.6
Región del Biobío	Quilaco	2	0.0	2.0
Región del Biobío	Quilleco	4	0.2	10.3
Región del Biobío	Santa Bárbara	3	0.1	7.4
Región del Biobío	Tucapel	2	0.2	10.7
Región del Biobío	Yumbel	42	24.8	1500.9
Región del Biobío	Alto Biobío	1	0.1	4.8
Región de La Araucanía	Cunco	79	54.2	3278.8
Región de La Araucanía	Curarrehue	4	0.1	4.1
Región de La Araucanía	Melipeuco	4	3.3	198.3
Región de La Araucanía	Pucón	16	19.7	1192.9
Región de La Araucanía	Villarrica	6	3.3	197.3

Tabla A.1. Resultados a nivel sub-regional (número de parcelas, superficie y producción estimada de trigo por comuna).