UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE AGRONOMÍA Y ZOOTECNIA

ESCUELA PROFESIONAL DE AGRONOMÍA

TESIS

"APLICACION DE LA METODOLOGIA DE DESAGREGACION ESPACIO-TEMPORAL DE ESTIMACIONES REMOTAS DE LA HUMEDAD DEL SUELO MEDIANTE TECNICAS DE APRENDIZAJE AUTOMATICO EN LA SUB CUENCA HUATANAY, MICROCUENCA HUANACAURE, KAYRA-CUSCO EN EL PERIODO 2015-2022".

Presentada por:

Br. MARCELO BUENO DUEÑAS,

para optar al título profesional de
INGENIERO AGRÓNOMO.

ASESORES:

DR. CARLOS JESÚS BACA GARCÍA.
DR. PEDRO CHRISTOPHER RAU LAVADO.

CO-FINANCIADO POR: CONCYTEC Y NERC

Cusco-Perú

2023

DEDICATORIA

A Dios y a mi familia.

AGRADECIMIENTOS

A la Universidad Nacional de San Antonio Abad del Cusco, a la Facultad de Ciencias Agrarias, a la hermosa Escuela Profesional de Agronomía y a sus excelentes Docentes.

A las maravillosas personas del proyecto “RAHU: Seguridad hídrica y adaptación al cambio climático en las cuencas Hidrográficas de los glaciares peruanos”, financiado por el Fondo Newton - Paulet, NERC, la embajada británica en Perú y CONCYTEC, a través de su unidad ejecutora ProCiencia.

A mis asesores *Carlos Jesús Baca García* y *Pedro Christopher Rau Lavado* por la confianza otorgada en mí y sus valiosos consejos.

Al Servicio Nacional de Hidrología y Meteorología (SENAMHI), especialmente a *Elsa* por su valiosa ayuda.

CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El presente trabajo de investigación titulado: “Desagregación espacio-temporal de estimaciones remotas de la humedad del suelo mediante técnicas de aprendizaje automático en la cuenca del alto Urubamba en el periodo 2015-2022”. Se realizó entre mayo del 2021 y agosto del 2022. El área de estudio seleccionada fue una superficie de aproximadamente 8 328 km^2, que va desde 72.30°O a 70.83° O y desde 13.13°S a 14.68°S correspondiendo a aproximadamente el 10 % de la superficie del departamento del Cusco, abarcando total o parcialmente las provincias de Calca, Canchis, Canas, Acomayo, Cusco, Anta, Urubamba, y un área pequeña de la provincia de Paucartambo.

El contenido de agua del suelo es capaz de predecir el impacto de sequias en el rendimiento agrícola mejor que la precipitación. En zonas sin instrumentación la teledetección es una alternativa viable para obtener información de la humedad del suelo. Sin embargo, las estimaciones de la humedad del suelo desde el espacio son poco adecuadas para aplicaciones agrícolas, hidrológicas y ambientales que requieren información diaria y a alto detalle espacial.

El objetivo de la investigación fue analizar, en el contexto de la cuenca Urubamba Vilcanota, la posible utilidad de la técnica de aprendizaje automático *random forest* para mejorar la resolución espacio-temporal del producto SMAP-L3-E de humedad del suelo del satélite SMAP mediante comparación con datos *in situ*.

Para cumplir tal objetivo se construyeron modelos de desagregación espacio-temporal del producto SMAP-L3-E mediante *random forest*, se evaluaron visual y estadísticamente para comprobar su capacidad de desagregación; seguidamente se determinó la influencia de la topografía, las propiedades del suelo y la precipitación en la dinámica espacial del producto SMAP-L3-E desagregado; y finalmente se analizó la relación entre el producto SMAP-L3-E desagregado mediante *random forest* con la humedad del suelo medida *in situ* en el área bajo estudio.

Después de entrenar los modelos de desagregación espacio-temporales usando *random forest* como función básica de desagregación, quedó demostrado que la desagregación temporal (reconstrucción de las series de tiempo) asemeja adecuadamente la dinámica temporal del producto SMAP-L3-E. Respecto a la desagregación espacial, la validación visual mostró que la desagregación es coherente con la distribución original de la humedad del suelo, pero además la mejora significativamente. El análisis de la validación estadística en ambos casos mostró que el error de generalización de los modelos de desagregación es adecuado para aplicaciones científicas y prácticas. Se demostró que *random forest* es capaz de desagregar espacio-temporalmente el producto SMAP-L3-E en el área de estudio.

Mediante análisis espacial y análisis de componentes principales se encontró que el producto SMAP-L3-E desagregado depende fundamentalmente de la elevación, del contenido de carbono orgánico del suelo, del contenido de arcilla y la conductividad hidráulica saturada del suelo, pero solo en condiciones de cercanas a saturación.

Respecto a la validación con datos *in situ* se encontró que el producto SMAP-L3-E desagregado mediante *random forest* explica adecuadamente las mediciones *in situ* de la

humedad del suelo en el área bajo monitoreo en condiciones de bajo contenido de agua en el suelo, sin embargo, la relación diverge de ese comportamiento en condiciones de contenidos de humedad entre 0.4 a 0.5 cm^3 cm^{-3}, por lo que el esquema de desagregación propuesto en este estudio no dio resultados adecuados en esas condiciones específicas y en frecuencias temporales diarias. Sin embargo, cuando se consideraron agregaciones temporales más gruesas, entre 7 a 10 días, la correlación entre las dos series de tiempo fue en promedio 0.98, independientemente de las condiciones de saturación.

Palabras Clave: Humedad del suelo, Aprendizaje Automático, Teledetección, Desagregación espacial.

INTRODUCCIÓN

El agua es el componente más dinámico del sistema suelo y se relaciona con una amplia variedad de procesos hidrológicos y edáficos que permiten entender la ocurrencia de sequías, la lixiviación de contaminantes, la escorrentía superficial, el ciclo del carbono, la capacidad del suelo para almacenar y proporcionar nutrientes (Weil & Brady, 2017, p. 232), etc.

Por lo tanto, el acceso oportuno a información de la distribución espacial y temporal de la humedad del suelo es de gran importancia para la conservación de los recursos naturales, la producción agropecuaria, la predicción de riesgos y en general la utilización sostenible de los recursos hídricos.

Se han desarrollado una serie de técnicas que permiten medir la humedad del suelo con instrumentos de campo o laboratorio, que incluyen métodos gravimétricos (Topp & Ferré, 2018), reflectometría de dominio de tiempo, TDR (Topp et al., 1980), sensores de capacitancia (Gaskin & Miller, 1996), sondas de neutrones, mediciones de resistividad eléctrica, sensores de pulso de calor y sensores de fibra óptica.

Con estas técnicas, se pueden obtener mediciones altamente precisas de la humedad del suelo a escala puntual. Estas técnicas tienen las ventajas de una fácil instalación, relativa madurez técnica y la capacidad de medir la humedad del suelo a diferentes profundidades del perfil (Peng et al., 2017).

Sin embargo, estas mediciones puntuales sufren de numerosas desventajas, a saber, normalmente solo es posible realizar pocos muestreos, son laboriosas, consumen mucho tiempo y principalmente no son representativas debido a la heterogeneidad de la humedad del suelo (Crow et al., 2012). Esta heterogeneidad depende de diversas variables ambientales, entre ellas las propiedades físicas y químicas del suelo, las características topográficas, al tipo de cobertura vegetal y a factores meteorológicos (Brocca et al., 2007; Mohanty & Skaggs, 2001).

Actualmente este problema se viene resolviendo progresivamente con el desarrollo de técnicas de teledetección (*remote sensing*, en inglés) que, en comparación con las técnicas mencionadas, pueden proporcionar datos de humedad del suelo de forma dinámica, barata, exacta y para áreas más extensas (Peng et al., 2017).

La teledetección de la humedad del suelo se basa principalmente en observaciones satelitales mediante sensores de reflectancia y/o emisión de radiación electromagnética (Babaeian et al., 2019; Mohanty et al., 2017) dentro del óptico, infrarrojo térmico (Zhang & Zhou, 2016), y microondas activo o pasivo (Chan et al., 2018; Jackson et al., 2010; Kerr et al., 2016).

Las observaciones satelitales con sensores de microondas son las más adecuadas para la estimación de la humedad del suelo (Mohanty et al., 2017; Schmugge et al., 2002) ya que minimizan la influencia negativa de la atmósfera, de la cobertura vegetal y en cierto grado de la topografía local en el proceso de estimación.

En los últimos veinte años se han lanzado varios satélites con sensores de microondas ya sean pasivos como activos capaces de estimar la humedad del suelo de forma remota.

Entre los más relevantes se pueden nombrar el SMOS de la ESA, y más recientemente el SMAP de la NASA, con diferentes grados de precisión y resolución espacio-temporal (Mohanty et al., 2017).

I. PROBLEMA OBJETO DE INVESTIGACIÓN

1.1. Identificación del problema objeto de investigación

Ha sido demostrado que el contenido de agua del suelo es un indicador más directo de la disponibilidad de agua para los cultivos y es capaz de predecir el impacto de sequias en el rendimiento agrícola mejor que la precipitación (Xia et al., 2014).

En países en vías de desarrollo como el Perú, se cuenta con escasa información de las propiedades y procesos del suelo a alta resolución espacio-temporal y por lo tanto es muy difícil disponer de esta para distintas aplicaciones (Rojas et al., 2017, p. 13); particularmente casi no existe a la fecha infraestructura de monitoreo de la humedad del suelo a largo plazo en países en vías de desarrollo (Brocca et al., 2017, p. 2).

Más aún, en el contexto del cambio climático existe más que nunca la necesidad de disponer de información continua y a largo plazo de la humedad del suelo (Dorigo & de Jeu, 2016, p. 3)**.**

En zonas sin instrumentación la teledetección es una alternativa viable para obtener información de la humedad del suelo de alta resolución y casi en tiempo real.

Actualmente la misión SMAP (*Soil Moisture Active Passive*) lanzada por la NASA el 31 de enero de 2015 es la principal fuente remota dedicada exclusivamente a proveer información continua de humedad del suelo a nivel global.

El producto SMAP-L3-E de la misión SMAP (Chan et al., 2018) es capaz de entregar información de la humedad del suelo a nivel global a través de observaciones pasivas del

radiómetro a bordo del SMAP, y permite una precisión media de 0.05 (5 %) cm^3cm^{-3} (Das et al., 2019).

Sin embargo, las estimaciones de la humedad del suelo del producto SMAP-L3-E se producen aproximadamente a nueve kilómetros de resolución espacial y con una periodicidad aproximada de cuatro días, lo cual constituye su principal limitación (Das et al., 2019), haciéndolas poco adecuadas para aplicaciones agrícolas, hidrológicas y ambientales que requieran información diaria y a alto detalle espacial (Vergopolan et al., 2021).

Se han propuesto varios métodos para mejorar la resolución espacial y temporal de estimaciones remotas de la humedad del suelo (proceso denominado como desagregación o "*downscaling*") (Mao et al., 2019; Peng et al., 2017).

Recientemente a través técnicas de aprendizaje automático o *machine learning*, como *random forest* (Liu et al., 2020) se han logrado avances en la desagregación de estimaciones remotas de la humedad del suelo, ya sea espacialmente (Bai et al., 2019; Chen et al., 2019; Zappa et al., 2019; Zhao et al., 2018) o temporalmente (Lu et al., 2015ª; Mao et al., 2019; Xing et al., 2017).

Aunque la teledetección ha probado ser una herramienta valiosa en la medición de la humedad del suelo, observaciones *in situ* aún son fundamentales para evaluar la precisión de los productos de humedad del suelo derivados de técnicas remotas de estimación (Dorigo & de Jeu, 2016).

La presente tesis propuso evaluar una técnica de aprendizaje automático denominada *random forest* para mejorar la resolución espacio-temporal de las estimaciones remotas de

la humedad del suelo del producto SMAP-L3-E del satélite SMAP desde el 2015 hasta el 2022 y evaluar tales predicciones durante un año hidrológico en un área de estudio perteneciente a la microcuenca K´ayra**.**

1.2.Planteamiento del problema

1.2.1. Problema general

¿Es posible usar la técnica de aprendizaje automático *random forest* para mejorar la resolución espacio-temporal del producto SMAP-L3-E de humedad del suelo del satélite SMAP con adecuada precisión respecto a mediciones *in situ* en la microcuenca K´ayra?

1.2.2. Problemas específicos

- ¿Cuál es la capacidad de modelos basados en *random forest* para desagregar espacio-temporalmente el producto SMAP-L3-E?
- ¿En qué medida la topografía, las propiedades del suelo y la precipitación describen la dinámica espacial del producto SMAP-L3-E desagregado mediante *random forest*?
- ¿Cuál es la relación entre el producto SMAP-L3-E desagregado mediante *random forest* con la humedad del suelo medida *in situ* en la microcuenca K´ayra?

II. OBJETIVOS Y JUSTIFICACIÓN

2.1. Objetivo general

Indagar la utilidad de la técnica de aprendizaje automático *random forest* para mejorar la resolución espacio-temporal del producto SMAP-L3-E de humedad del suelo del satélite SMAP mediante comparación con datos *in situ* en la microcuenca K´ayra,

2.2. Objetivos específicos

Para lograr este objetivo principal se proponen los siguientes objetivos específicos:

- Evaluar la capacidad de desagregación espacio-temporal mediante *random forest* del producto SMAP-L3-E.
- Determinar la influencia de la topografía, las propiedades del suelo y la precipitación en la dinámica espacial del producto SMAP-L3-E desagregado mediante *random forest*.
- Analizar la relación entre el producto SMAP-L3-E desagregado mediante *random forest* con la humedad del suelo medida *in situ* en la microcuenca K´ayra.

2.3. Justificación

2.3.1. Conveniencia

El estudio permitirá acceder a información de humedad del suelo de forma continua, exacta y de alta resolución espacio-temporal en la zona de estudio.

2.3.2. Relevancia social

El estudio propuesto permitirá contar con una herramienta validada que beneficiará futuros proyectos a nivel agronómico, ambiental e hidrológico y podría implementarse en programas de monitoreo de riesgos agroclimáticos y de adaptación al cambio climático en beneficio de comunidades agrícolas.

2.3.3. Implicaciones prácticas

La información continua de la humedad del suelo de alta resolución espacio-temporal permitirá plantear soluciones científicas a la ocurrencia de sequías agrícolas , incendios forestales, contaminación de aguas subterráneas (Zhang et al., 2020), erosión hídrica, inundaciones (Brocca et al., 2017), deslizamientos y movimientos en masa, estimación de rendimientos de cosecha, programación del riego, etc.

2.3.4. Valor teórico

Actualmente en Perú no hay ningún estudio sobre estimaciones remotas de la humedad del suelo ni aplicaciones de aprendizaje automático para mejorar la resolución espacio-temporal de tales estimaciones. Mediante el estudio propuesto se podrá comprender por primera vez la dinámica espacio-temporal del agua del suelo a nivel de microcuencas dentro de la cuenta Vilcanota-Urubamba.

2.3.5. Utilidad metodológica

La investigación haría uso por primera vez en el contexto nacional de instrumentos modernos de medición de la humedad del suelo ya sea a nivel de campo como mediante teledetección satelital, y a través de la metodología propuesta mejoraría la disponibilidad de información de humedad del suelo. La investigación sugiere un método novedoso basado en aprendizaje automático para resolver el problema de la baja resolución espacio-temporal del producto de humedad del suelo L2-SM-SP de la misión SMAP.

III. HIPÓTESIS

3.1. Hipótesis general

Es posible utilizar el método *random forest* para mejorar la resolución espacio-temporal del producto SMAP-L3-E de humedad del suelo del satélite SMAP con adecuada precisión respecto a mediciones *in situ* en la microcuenca K´ayra.

3.2. Hipótesis específicas

- El modelo *random forest* es capaz de desagregar espacio-temporalmente el producto SMAP-L3-E.
- La precipitación explica la mayor parte de la dinámica espacial del producto SMAP-L3-E desagregado con *random forest* seguida por la topografía y las propiedades del suelo.
- El producto SMAP-L3-E desagregado mediante *random forest* explica adecuadamente las mediciones *in situ* de la humedad del suelo en la microcuenca K´ayra.

IV. MARCO TEÓRICO

4.1. Contenido de humedad del suelo

El contenido de agua del suelo, o la humedad del suelo, es el volumen o la masa de agua que ocupa espacio dentro de los poros del suelo (Tindal & Kunkel, 1999, p. 30).

La humedad del suelo puede ser expresada en forma gravimétrica, volumétrica o como grado de saturación (Hillel, 2004, p. 14).

La humedad del suelo puede ser definida en términos de masa la cual es referida como contenido gravimétrico de humedad o humedad másica la que se define como la masa de agua en relación a la masa del material solido seco de la muestra de suelo bajo análisis. Para su determinación se asumen dos condiciones, se debe remover el suelo de su ubicación original y debe existir un criterio para definir cuándo el suelo se considera seco. Por lo tanto, según la Sociedad Americana de Ciencia de suelos (2008) el contenido gravimétrico de humedad es la relación entre la masa de agua perdida de un suelo a 105 °C y la masa de dicho suelo. El contenido gravimétrico de humedad varía entre diferentes tipos de suelos entre 25% a 60% dependiendo de la densidad aparente (Hillel, 2004).

El contenido volumétrico de humedad generalmente es el índice de humedad más útil para estudios a nivel de campo o laboratorio principalmente por dos motivos, primeramente es la forma en que la humedad del suelo es registrada por equipos de atenuación gamma, sonda de neutrones o reflectometría en el dominio del tiempo (TDR) (Tindal & Kunkel, 1999) y además es más directamente aplicable en el cálculo de los flujos de agua en el suelo (infiltración, percolación y drenaje) (Hillel, 2004, p. 15).

El contenido volumétrico de humedad puede expresarse de la siguiente manera:

$$\theta = \frac{V_{agua}}{V_{suelo}}$$

Ecuación 1

Donde V_{agua} es el volumen de agua dentro del volumen de suelo analizado, V_{suelo}. Es una proporción entre volúmenes y por lo tanto adimensional, pero usualmente se expresa en centímetros cúbicos por centímetros cúbicos [L^3L^{-3}]. En suelos arenosos en saturación Θ es aproximadamente 40%, en suelos francos 50% y en suelos arcillosos se aproxima a 60% (Hillel, 2004).

El concepto de saturación efectiva o contenido relativo de humedad es usado a menudo en estudios de flujo del agua en el suelo (Vereecken et al., 2019) y se define de la siguiente manera:

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r}$$

Ecuación 2

Donde Θ es la saturación efectiva, θ_s es contenido volumétrico de humedad en saturación (igual a la porosidad total si no existe aire atrapado y todos los poros están llenos de agua) y θ_r es el contenido volumétrico residual de humedad. El contenido volumétrico residual de humedad es definido como el contenido volumétrico de humedad donde la conductividad hidráulica se aproxima a cero $K(\theta) = 0$ para $\theta = \theta_r$. Comúnmente se asume que $\theta_r = 0$ (van Genuchten, 1980). Este índice expresa el contenido de humedad presente en el suelo en relación al contenido de humedad de saturación (Hillel, 2004). Este índice va desde cero cuando el suelo está completamente seco hasta 100% en saturación.

Como la humedad del suelo es un cociente entre volúmenes o masas esta magnitud es adimensional, cuando este valor es multiplicado por 100 se convierte en valores porcentuales sin embargo se recomienda el uso de las unidades respectivas para evitar ambigüedades de interpretación (Strawn et al., 2020), por ejemplo un contenido volumétrico de humedad de 32% debe expresarse como 0.32 cm^3/cm^3.

4.2. Medición remota de la humedad del suelo

La medición remota de la humedad del suelo se basa principalmente en observaciones satelitales mediante sensores de reflectancia y/o emisión de radiación electromagnética (Babaeian et al., 2019; Mohanty et al., 2017) dentro del óptico, infrarrojo térmico (Zhang & Zhou, 2016), microondas activo y microondas pasivo (Chan et al., 2018; Jackson et al., 2010; Kerr et al., 2016), cada una con ventajas y desventajas inherentes al método de medición. Los principales métodos de medición remota de la humedad del suelo se presentan en la tabla 1.

Tabla 1*.Métodos de medición remota de la humedad del suelo*

Método de teledetección de la humedad del suelo	Ventajas	Desventajas	Referencias
Métodos ópticos (Vis-NIR- SWIR⋆)	Cobertura espacial amplia, alta resolución espacial, potencialidad para aplicaciones en tiempo real, aplicabilidad de sensores multi e hiper-espectrales.	Limitada profundidad de medición (milímetros), alta perturbación por nubes y la vegetación, baja resolución.	(Zhang et al., 2013)

Continúa Tabla 1

Métodos térmicos	Cobertura espacial amplia, potencialmente alta resolución espacial, potencialidad para aplicaciones en tiempo real (drones), alta correlación entre la HS⋆⋆ y la temperatura del suelo.	Limitada profundidad de medición (milímetros), alta perturbación por nubes y la vegetación, atenuación atmosférica, baja resolución temporal.
Métodos de microondas activos	Sensibilidad a la permitividad del suelo, cobertura espacial amplia (global), profundidad de medición de hasta 5 cm, alta resolución espacial, alta correlación entre la dispersión del radar y la HS, insensible a nubes y condiciones atmosféricas, no depende de la radiación solar.	Perturbación por la topografía y la vegetación.

Continúa Tabla 1

Métodos de microondas pasivos	Sensibilidad a la permitividad del suelo, cobertura espacial amplia, profundidad de medición de hasta 5 cm, alta resolución espacial, la temperatura de brillo es insensible a nubes y condiciones atmosféricas.	Baja resolución espacial comprada con los otros métodos y perturbación por la topografía y la vegetación.	(Entekhabi et al., 2010) (Kerr, 2007)

⋆ Vis = longitudes de onda visible, NIR = infrarrojo próximo y SWIR = infrarrojos de onda corta.

⋆⋆HS = Humedad del suelo

Fuente: Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H., & Tuller, M. (2019). Ground, Proximal, and Satellite Remote Sensing of Soil Moisture (Teledetección satelital, proximal y a nivel de campo de la humedad del suelo). Reviews of Geophysics, 57(2), 530-616.

4.3. Teledetección de la humedad del suelo basada en radiación de microondas

La región de microondas del espectro electromagnético ha demostrado un inmenso potencial en la medición precisa y eficiente de la humedad del suelo debido principalmente a la gran disparidad entre la permitividad dieléctrica del agua respecto a la permitividad de los materiales minerales u orgánicos que conforman el suelo (Das, 2019).

Las propiedades dieléctricas del agua, minerales y de la materia orgánica afectan la emisividad y la retrodispersión de microondas del suelo (Topp et al., 1980).

La humedad del suelo puede ser estimada mediante mediciones de emisividad y retrodispersión de microondas en diferentes bandas de frecuencia incluyendo la banda P, banda L, banda C y banda X. Se han propuesto modelos empíricos denominados *modelos de mezcla dieléctrica* para inferir la permitividad del suelo y su relación con el contenido de humedad (Mironov et al., 2009; Wang & Schmugge, 1980).

La teledetección de la humedad del suelo mediante microondas se puede clasificar en dos categorías según la fuente de las señales que utiliza, 1) el radar activo, que mide la señal de retrodispersión después de transmitir un pulso electromagnético y 2) el radiómetro pasivo que mide las emisiones naturales de la superficie terrestre (Mao et al., 2019).

4.4. Teledetección pasiva de la humedad del suelo.

En radiometría pasiva de microondas, los sensores, también llamados radiómetros, miden la emisión de microondas a través de la temperatura del brillo, T_β, que es una magnitud que describe la cantidad natural (de ahí el término pasiva) de radicación de microondas, también llamada emisión térmica, emitida por un medio específico (Babaeian et al., 2019, p. 20; Montzka et al., 2020, p. 26; Vereecken et al., 2014, p. 4). La intensidad de esta radiación (para medios naturales como el suelo) depende de sus propiedades dieléctricas y de la temperatura de dicho medio.

Según la teoría electromagnética, la temperatura del brillo T_β puede ser usada para calcular el contenido de humedad del suelo. En procesos naturales de humedecimiento y secamiento los suelos emiten un amplio rango de T_β (pudiendo darse variaciones de hasta 70 K). Dada la incertidumbre radiométrica típica de ~1 K o menos para los radiómetros modernos, este rango dinámico de 90 K de T_β entre suelos húmedos y secos proporciona un

método muy favorable para la estimación precisa de la humedad del suelo (Montzka et al., 2020).

A pesar de las incertidumbres mencionadas y la degradación de la señal causadas por estos factores, algunos de sus impactos pueden mitigarse fácilmente con observaciones de T_β adquiridas a frecuencias más bajas como la banda L (1,4 GHz) que, por ejemplo, la banda C (6,9 GHz) o la X (10,7 GHz) como se puede observar en la tabla 2. En las frecuencias de banda L, los impactos de la rugosidad superficial, la cobertura de la vegetación, la atenuación atmosférica y los efectos ionosféricos son más fácilmente corregibles o mucho menos dominantes que la señal de emisión debida a la humedad del suelo.

Tabla 2.*Profundidad de medición aproximada para el suelo y diferentes tipos de uso de suelo mediante diferentes bandas de microondas.*

Tipo de superficie	Banda X (8-12 GHz)	Banda C (4-8 GHz)	Banda L (1-2 GHz) ⋆	Banda P (0.3-1 GHz)
Suelo	~1.25–1.87	~1.87–3.75	~7.5–15	~15–50
Cultivo o pastura	~0.5–0.75	~0.75–1.5	~3–6	~6–20
Bosque	~0.25–0.37	~0.37–0.75	~1.5–3	~3–10

Fuente: Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H., & Tuller, M. (2019). Ground, Proximal, and Satellite Remote Sensing of Soil Moisture (Teledetección satelital, proximal y a nivel de campo de la humedad del suelo). Reviews of Geophysics, 57(2), 530-616.
⋆ La misión SMAP utiliza un Radar y un radiómetro en la banda L.

Por lo tanto, la banda L es considerada el rango de frecuencia más adecuado para la teledetección de la humedad del suelo, y ha sido utilizada por las misiones de teledetección

de la humedad del suelo en la última década (por ejemplo, Aquarius, SMOS y SMAP). Se han validado algoritmos de estimación de la humedad del suelo basados en la relación de la T_β en la banda L, la contante dieléctrica del suelo y la humedad del suelo llegando a una precisión de estimación de 0.04 cm^3/cm^3 y una correlación de más de 0.80 en relación a observaciones de la humedad del suelo realizadas en campo (Chan et al., 2018; Montzka et al., 2020, p. 21).

Como resultado de múltiples estudios de campo en las dos últimas décadas, los algoritmos de estimación de la humedad del suelo mediante microondas pasivo han evolucionado, convirtiéndose en un método robusto y fiable (NASA, 2014).

4.5. La misión SMAP

Lanzado al espacio en enero de 2015, el satélite humedad del suelo activo pasivo (SMAP, por sus siglas en ingles) de la Administración Nacional de Aeronáutica y del Espacio (NASA) fue diseñado para proporcionar un mapeo global de la humedad del suelo a alta resolución espacial y temporal (Chan et al., 2018).

Después de SMOS y Aquarius, SMAP es la tercera misión en menos de una década utilizando un radiómetro de banda L para estimar la humedad del suelo desde el espacio (Chan et al., 2018)

El satélite SMAP posee un sensor de microondas diseñado para medir y mapear la humedad del suelo a nivel global con una incertidumbre de ~±0.04 cm^3cm^{-3} en condiciones de media a baja vegetación (NASA, 2014); el sistema de medición de SMAP consiste en un radiómetro y un radar de apertura sintética que funcionan con polarización múltiple en el rango de la banda L.

Mientras que el satélite SMAP fue diseñado para pasar al menos 3 años en órbita, el radar a bordo del SMAP falló en julio del 2015 (tres meses después del lanzamiento), dejando solo el radiómetro del SMAP operativo hasta el día de hoy (Babaeian et al., 2019, p. 37).

Los principales objetivos científicos de SMAP son entender los procesos que vinculan los ciclos del agua y del carbono del suelo, y mejorar las predicciones meteorológicas y climáticas

(Entekhabi et al., 2010). Los parámetros técnicos del satélite SMAP se resumen en la siguiente tabla 3:

Tabla 3.Características técnicas del satélite SMAP

Nombre del modelo	Descripción
Instrumentos	Radiómetro de banda L (1.41 GHz). Radar de banda L⋆ (1.22 a 1.3 GHz).
Ancho de franja de medición	1000 Km.
Altitud de órbita	685 Km.
Inclinación	98 grados, sol-sincrónico.
Tiempo local de órbita de ascenso	6:00 p.m.

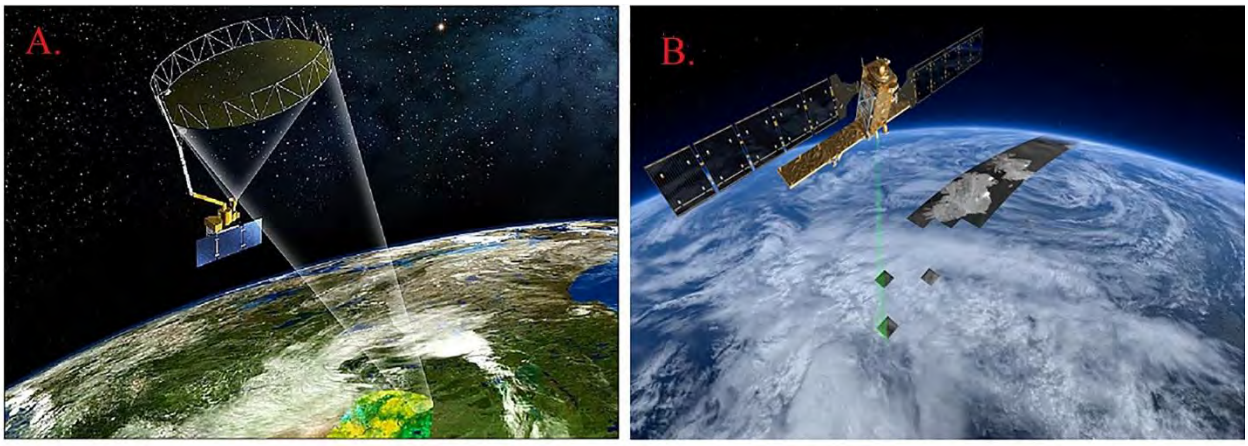
Continúa Tabla 3

Tiempo local de órbita de descenso .	6:00 a.m.
Tiempo de revisita	2 a 3 días

⋆El radar dejó de funcionar por una falla en el sistema de circuitos el 2015. Y fue remplazado por el radar de la misión Sentinel 1 (banda C).
Fuente: Elaboración propia.

Recientemente, las mediciones de retrodispersión del radar a bordo del Sentinel-1 (banda C) se han combinado con observaciones de temperatura de brillo del radiómetro a bordo del SMAP (banda L) para generar un producto de humedad del suelo (L2-SM-SP) de alta resolución espacial que proporciona información desde abril del 2015 hasta la actualidad y disponible para el público a través de https://nsidc.org/data/SPL2SMAP_S/versions/3 (Das, 2019). En la figura 1 se muestran imágenes renderizadas del SMAP y del Sentinel-1.

Figura 1. *Representaciones digitales de los satélites SMAP y Sentinel-1*



Nota: A. SMAP, B. Sentinel-1. Adaptado de Ground, Proximal, and Satellite Remote Sensing of Soil Moisture [Teledetección satelital, proximal y a nivel de campo de la humedad de suelo], por Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H., & Tuller, M. (2019) en Reviews of Geophysics, 57(2), 530-616.

4.6. Estimación pasiva de la humedad del suelo de SMAP

La estimación de la humedad del suelo a través de las observaciones de temperatura de brillo del radiómetro del SMAP se basa en una aproximación a la ecuación de transferencia radiativa, comúnmente conocida como modelo tau-omega (O'Neill et al., 2020, p. 17).

En forma general hay siete pasos involucrados en la estimación de la humedad del suelo utilizando la teledetección activa-pasiva de microondas (NASA, 2014, p. 53; O'Neill et al., 2020ª):

1. La data de retrodispersión del *Sentinel*-1 es topográficamente corregida, filtrada para reducir la influencia de estructuras artificiales y agregada a 1 km de resolución espacial.

2. La data de retrodispersión se empareja con la observación de temperatura de brillo de la órbita de descenso (6: 00 a.m.) del SMAP.
3. La temperatura del brillo se calcula espacialmente mediante el algoritmo pasivo (ecuación de transferencia radiativa) (Das et al., 2019).
4. La temperatura de brillo desagregada es convertida a emisividad usando la aproximación Rayleigh-Jeans (O'Neill et al., 2020)
5. Se estima la emisividad del suelo considerando el efecto de la vegetación a través del modelo tau-omega, implementado en el algoritmo de O'Neill et al. (2020).
6. Se calcula la permitividad del suelo.
7. Se aplica un modelo dieléctrico para relacionar la constante dieléctrica estimada con la humedad del suelo, la misión SMAP implementa los modelos de Mironov et al. (2009), y Wang y Schmugge (1980).

4.7. Estructura de datos del SMAP

La información que libera la misión SMAP a la comunidad científica se divide en 4 niveles en función de las aplicaciones y el nivel de procesamiento realizado para obtenerla (NASA, 2014) esquematizados en al tabla 4.

Tabla 4.Tipos de niveles de información de los productos generados por la misión SMAP.

Nivel de productos de datos del SMAP	Descripción
Nivel 0	Los productos de data de nivel 0 contienen información instrumental no procesada.
Nivel 1	El nivel 1 contiene información de la temperatura de brillo geolocalizada y calibrada a las condiciones de vegetación y topografía local; adicionalmente contiene las covariables usadas para resolver el modelo tau- omega (O'Neill et al., 2020).
Nivel 2	Los productos de data de nivel 2 contienen estimaciones de la humedad del suelo en base a la fusión de datos del radiómetro para la órbita ascendente (6:00 p.m.) y la órbita descendente (6:00 a.m.) (Liu et al., 2020).
Nivel 3	Los productos de data de nivel 3 contienen estimaciones medias de la humedad del suelo en base al radiómetro para ambas órbitas, es una media diaria (Liu et al., 2020).
Nivel 4	Los productos de data de nivel 4 contienen productos científicos derivados como humedad del suelo en la zona radicular o flujo de dióxido de carbono del suelo, que están bajo desarrollo y validación científica.

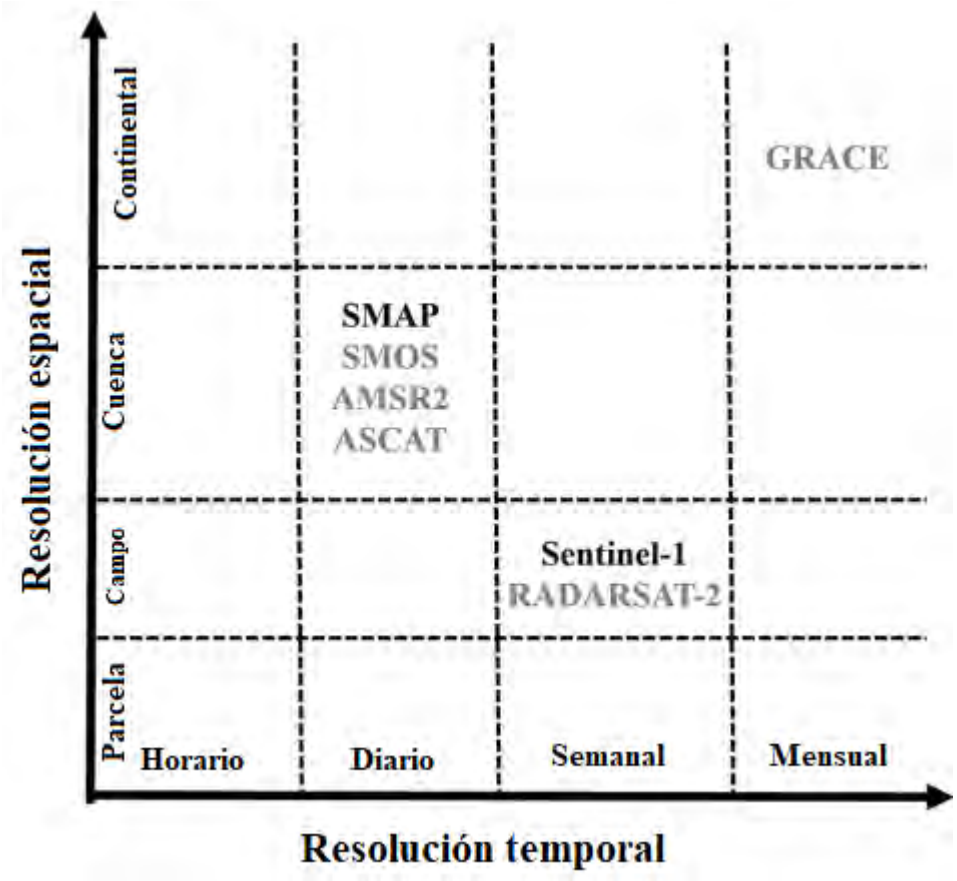
Fuente: NASA. (2014). SMAP Handbook Soil Moisture Active Passive (Manual de la misión de Humedad del Suelo Activa Pasiva). Jet Propulsion Laboratory California Institute of Technology

.

4.8. El producto SMAP-L3-E

Es un producto de nivel 3 de la misión SMAP que contiene estimaciones diarias de la humedad del suelo a nueve kilómetros de resolución espacial basadas en la temperatura de brillo medida por el radiómetro del SMAP mediante el algoritmo descrito por Entekhabi et al. (2010). Actualmente es el producto satelital pasivo de humedad del suelo con mayor resolución espacial a nivel global (Chan et al., 2018; Das et al., 2019) figura 2.

Figura 2.*Escalas espacio temporales y misiones satelitales de monitoreo de la humedad del suelo.*



Elaborado por Marcelo Bueno Dueñas.

En un reciente estudio Das et al. (2019) realizaron una evaluaron del producto SMAP-L3-E basado en observaciones in situ de la humedad del suelo de las redes de sensores: Red de Referencia Climática (CRN) del NOAA, la red de Análisis del Clima y Suelo (SCAN) del USDA, la red *Mesonet* en el estado de Oklahoma, la red MAHASRI (Mongolia), red SMOSMania (Europa) y Red las Pampas (Argentina), desde abril del 2015 a octubre del 2018. Se encontró que el producto SMAP-L3-E posee una incertidumbre de estimación (error cuadrático medio) aproximado a 0.062 cm3/cm3 y un coeficiente de correlación por encima de 0.600 en función de la cobertura vegetal sobre el suelo como se observa en la tabla 5. Cabe mencionar que una característica de las redes mencionadas anteriormente es su baja densidad de puntos de muestreo, resultando en un punto de medición por pixel del producto SMAP-L3-E.

Tabla 5*.Evaluación del producto de humedad del suelo activo pasivo SMAP-L3-E en diferentes estaciones de monitoreo en función del tipo de cobertura vegetal.*

Tipo de cobertura vegetal	ubRMSE (cm^3cm^{-3})	Sesgo (cm^3cm^{-3})	RMSE (cm^3cm^{-3})	R (-)	N⋆
Arbustal	0.046	0.008	0.046	0.544	43
Bosque	0.056	-0.001	0.065	0.489	7
Pastura	0.060	-0.036	0.069	0.647	236
Cultivo	0.076	-0.041	0.094	0.468	80

Continuación de la tabla 5

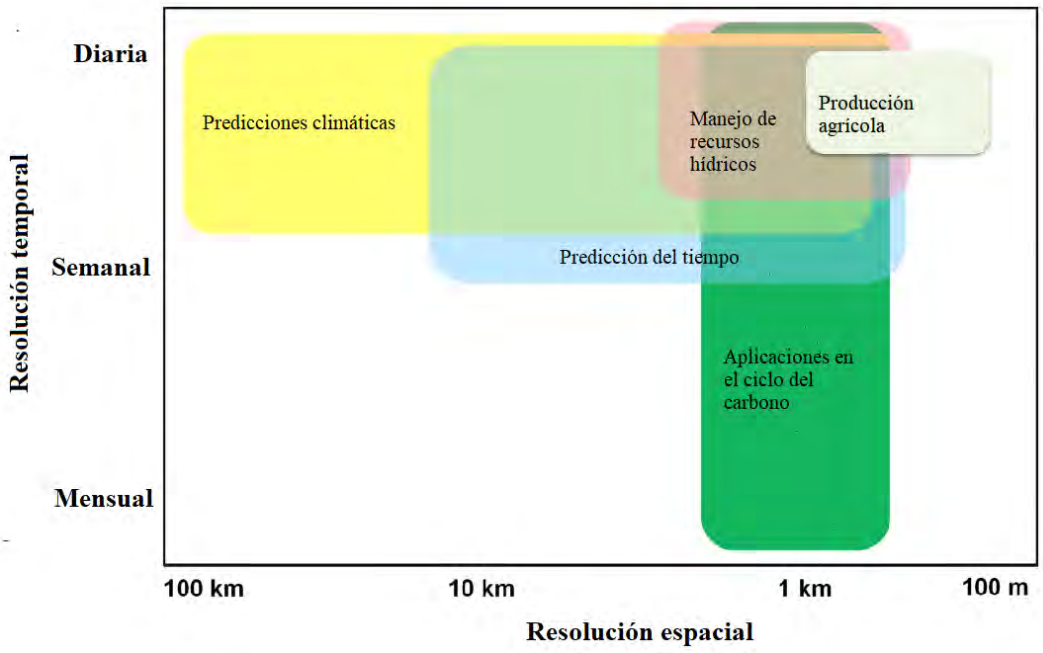
Cultivo/ vegetación natural (mosaico)	0.068	-0.008	0.077	0.349	8
Suelo sin vegetación	0.023	0.018	0.036	0.592	6
Promedio	0.052	-0.028	0.064	0.548	

Fuente: Das, N. N., Entekhabi, D., Dunbar, R. S., Chaubell, M. J., Colliander, A., Yueh, S., Jagdhuber, T., Chen, F., Crow, W., O'Neill, P. E., Walker, J. P., Berg, A., Bosch, D. D., Caldwell, T., Cosh, M. H., Collins, C. H., Lopez-Baeza, E., & Thibeault, M. (2019). The SMAP and Copernicus Sentinel 1ª/B microwave active-passive high resolution 24specia soil moisture 24special [El producto activo-pasivo de humedad del suelo de alta resolución del SMAP y Sentinel 1ª/B]. Remote Sensing of Environment, 233, 111380. https://doi.org/10.1016/j.rse.2019.111380

⋆ Número total de puntos de validación.

El producto SMAP-L3-E posee una resolución espacial nunca antes alcanzada en teledetección satelital pasiva de la humedad del suelo con una precisión adecuada para aplicaciones científica a nivel global o regional (Das et al., 2019); sin embargo para que el SMAP-L3-E sea útil en aplicaciones agrícolas, hidrológicas (figura 3) y ambientales que requieran información de la humedad el suelo de forma diaria y a mayor detalle espacial (Babaeian et al., 2019, p. 27) es necesario implementar algún método de desagregación (*downscaling*) en la data del SMAP-L3-E .

Figura 3*.Escalas de aplicación de información de humedad del suelo.*



Elaborado por Marcelo Bueno Dueñas.

4.9. Propiedades del suelo.

La heterogeneidad de las propiedades del suelo como la textura, el contenido de materia orgánica, la porosidad, la estructura y la microporosidad afecta la distribución de la humedad del suelo principalmente a pequeñas escalas (Crow et al., 2012; Vereecken et al., 2014).

SoilGrids (Hengl et al., 2017) es un 25specia de predicción 25special de propiedades y clases del suelo desarrollado por ISRIC (*Centro de Información Internacional del Suelo*) el cual integra técnicas de aprendizaje automático, bases de datos de perfiles de suelos a nivel mundial y data remota (de Sousa et al., 2020) para generar predicciones de propiedades del suelo a alta resolución espacial a las cuales es posible acceder de forma gratuita a través de *https://soilgrids.org/*.

4.10. Modelo digital de elevación (DEM) e índice de humedad topográfica (TWI)

Varios estudios han demostrado la utilidad de incluir parámetros topográficos y geomorfométricos derivados de modelos digitales de elevación (DEM) en modelos predictivos de propiedades y procesos del suelo (Li et al., 2020; Raduła et al., 2018).

Un modelo digital de elevación (DEM, por sus siglas en inglés) se refiere a un conjunto de puntos grillados con valores de elevación que aproximan la superficie de la tierra (Hengl et al., 2017).

El DEM MERIT es un modelo digital de elevación de alta precisión de 90 m de resolcuion espacial derivado del DEM SRTM a través de un proceso de corrección de errores (Yamazaki et al., 2017), el cual está disponible gratuitamente con fines de investigación y educación a través *http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/*.

El índice de humedad topográfica IHT, también llamado índice topográfico o índice topográfico compuesto (Qin et al., 2007) es un parámetro que describe la tendencia de un píxel a acumular agua. El índice de humedad topográfica se define como:

$$TWI = ln\left[\frac{A}{\tan(\beta)}\right]$$

Ecuación 3

Donde A es el área de drenaje o captación (AdC), β es la pendiente topográfica local. El concepto de TWI se basa en el principio de conservación de masa donde el área de captación total es un parámetro de la tendencia a recibir agua y la pendiente local es un parámetro de la tendencia a evacuar el agua. El TWI asume condiciones de estado

estacionario y condiciones espacialmente invariables para la infiltración y transmisividad del agua en el suelo (Gruber & Peckham, 2009).

Es posible calcular el *AdC* de varias formas. Los diferentes algoritmos de cálculo del *AdC*, comúnmente denominados algoritmos de flujo, generalmente se pueden agrupar en dos categorías: algoritmos de dirección de flujo único (SFD) y dirección de flujo múltiple (MFD), dependiendo de cómo se distribuya el flujo de agua potencial entre los píxeles de un DEM. Los algoritmos SFD no muestran divergencia en la dirección del flujo y están restringidos al movimiento en una dirección descendente de un *pixel* a otro a la vez, mientras que los algoritmos MFD son divergentes en la dirección del flujo, el flujo se extiende a varios (dos o más) *pixels* adyacentes según el gradiente topográfico del DEM (Hengl et al., 2009)

4.11. Desagregación de data remota

Los datos derivados de sensores remotos se utilizan para estudiar una amplia gama de procesos terrestres, estos datos tienen la ventaja de una amplia cobertura espacial y temporal, sin embargo, una desventaja es que a menudo contienen valores faltantes o resoluciones no adecuadas para aplicaciones específicas (Gerber et al., 2018).

Según Peng et al. (2017) la desagregación, reducción de escala o *downscaling* se refiere al proceso de generación de información a una escala temporal y/o espacial específica dada la información original a cierta escala temporal y/o espacial mayor (de resolución más baja).

Según Gerber et al. (2018) con el fin de desagregar data remota, los valores faltantes a o pixeles a resoluciones muy bajas a menudo se reemplazan con predicciones de una variedad de métodos de predicción (también llamados métodos de *llenado de brechas* o

métodos de imputación o de *cambio de escala*). Muchos de estos métodos de predicción explotan la correlación con otras covariables predictoras.

La predicción de valores faltantes en data remota, ya sea en el dominio espacial, temporal o espacio-temporal puede realizarse mediante diferentes técnicas (Van der Meer, 2012), principalmente: técnicas geoestadísticas, análisis de series de tiempo y aprendizaje automático.

Los métodos de análisis de series de tiempo son comúnmente utilizados en teledetección óptica e infrarroja (Landsat, MODIS, etc.) para predecir datos faltantes (Gerber et al., 2018).

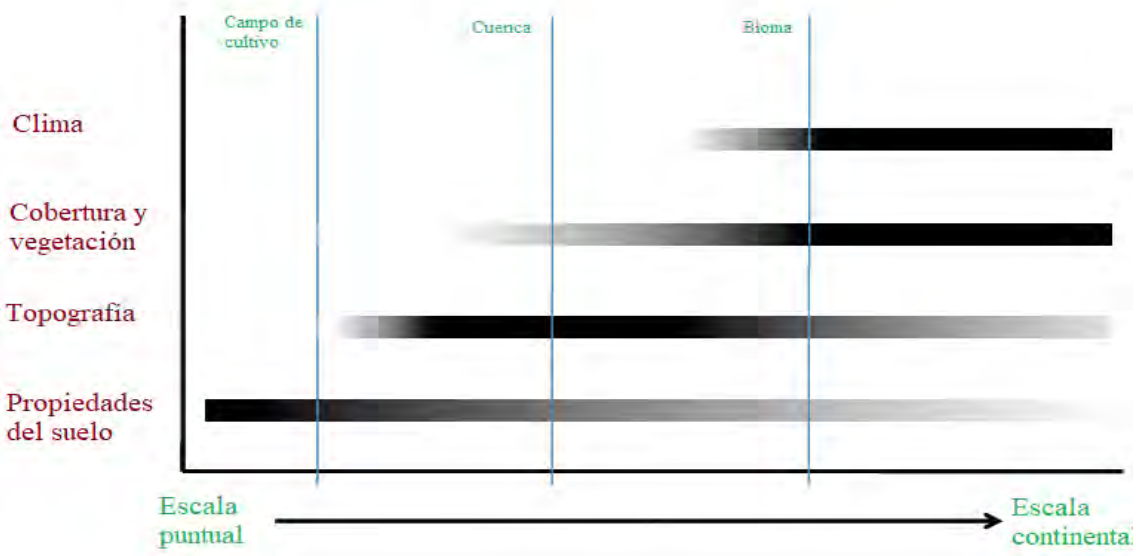
Los métodos de aprendizaje automático han sido ampliamente utilizados en la comunidad de la ciencia del suelo con resultados satisfactorios, principalmente con fines de predicción espacio-temporal de propiedades y procesos del suelo (Khaledian & Miller, 2020) y particularmente en la desagregación espacio-temporal de productos pasivos de la humedad del suelo (Mao et al., 2019; Qu et al., 2021), son generalmente son más versátiles ya que no requieren satisfacer condiciones específicas de la data y permiten usar una gran cantidad de covariables predictoras (Fang et al., 2019).

4.12. Desagregación de estimaciones remotas de la humedad del suelo mediante técnicas de aprendizaje automático.

Dado que la humedad del suelo está correlacionada con las propiedades del suelo, con el tipo de cobertura y densidad de la vegetación , con parámetros topográficos y la

variabilidad de la precipitación en un área determinada (Crow et al., 2012) esta relación podría ser utilizada dentro de un proceso de desagregación (figura 4).

Figura 4*.Variables que controlan la distribución de la humedad del suelo a diferentes escalas*



Modificado de Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D., & Walker, J. P. (2012). Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products [Agregación de observaciones in situ de humedad del suelo con fines de validación de productos remotos de baja resolución]. Reviews of Geophysics, 50(2), Article 2. Nota: las barras reflejan la importancia relativa de cada variable a diferente escala.

A través de esta relación es posible crear un modelo que relacione estas variables geoespaciales con las estimaciones remotas de la humedad del suelo; y utilizar tal modelo en la predicción de la humedad del suelo ya sea en el dominio espacial como temporal.

4.12.1. Desagregación espacial mediante técnicas de aprendizaje automático.

Una forma sencilla de desagregación es mediante regresión lineal (Kutner et al., 2004). Según Qu et al. (2021) este método consiste en ajustar un modelo lineal mediante mínimos cuadrados a estimación remotas de la humedad del suelo con variables predictoras con el fin de determinar los coeficientes de regresión apropiados, a continuación, la humedad del suelo se estima mediante la aplicación del modelo a las covariables predictoras en un dominio espacial o temporal diferente.

Aunque los modelos de regresión lineal son bastante comunes en la investigación actual (Liu et al., 2020; Qu et al., 2021; Tu, 2019), el problema de la no linealidad de los predictores generalmente degrada la capacidad predictiva del modelo.

Recientemente se han propuesto nuevos métodos de desagregación de estimaciones remotas de humedad del suelo (Bai et al., 2019; Chen et al., 2019; Zappa et al., 2019; Zhao et al., 2018) a través técnicas de *machine learning* o aprendizaje automático, entre ellas *random forest,* redes neuronales y modelos bayesianos (Liu et al., 2020).

Srivastava et al. (2013) usaron covariables del MODIS para aumentar la resolución de estimaciones de humedad del suelo del satélite SMOS en el dominio espacial a través de técnicas de aprendizaje automático, incluyendo máquinas vectoriales de apoyo (SVM) y redes neuronales (ANN). Esta investigación fue es el primer intento de aplicar las técnicas de aprendizaje automático para mejorar la resolución de estimaciones satelitales de la humedad del suelo en el dominio del espacio.

Im et al. (2016) aplicaron tres técnicas de aprendizaje automático basadas en árboles de regresión con el objetivo de aumentar la resolución espacial de estimaciones de humedad

del suelo del satélite AMSR-E usando covariables del MODIS, concluyeron que *random forest* es la técnica más apropiada para mejorar la resolución espacial de estimaciones satelitales de la humedad del suelo.

4.13. Procedimiento de desagregación de estimaciones remotas de la humedad del suelo mediante técnicas de aprendizaje automático.

En general para mejorar la resolución espacio-temporal de estimaciones remotas de la humedad mediante modelamiento estadístico o *machine learning* se siguen los siguientes pasos (Bai et al., 2019; Chen et al., 2019; Im et al., 2016; Khaledian & Miller, 2020; Qu et al., 2021; Srivastava et al., 2013ª; Zhao et al., 2018):

1. Las covariables predictoras se promedian a la resolución espacial de las estimaciones remotas de la humedad del suelo; y se construye el modelo estadístico, mediante regresores de aprendizaje automático en el dominio espacial y temporal del estudio.
2. El modelo se calibra y evalúa a través de técnicas como validación cruzada, *bagging* o simulación de *Monte Carlo* (Kuhn & Johnson, 2013).
3. El modelo se aplica a las covariables predictoras para obtener predicciones de humedad del suelo en el dominio del tiempo (fechas sin datos, si es necesario) o del espacio (mejoramiento de resolución espacial).
4. Comparación de los valores predichos con data de campo en diferentes puntos del dominio espacial del estudio mediante técnicas estadísticas, a este proceso se le denominado validación (Das et al., 2019; Singh et al., 2019).

4.14. Aprendizaje automático

El aprendizaje automático es un campo de estudio que analiza el uso de algoritmos computacionales, modelos estadísticos y métodos numéricos con el fin de generar modelos predictivos en base a observaciones de un conjunto de variables predictoras (Hastie et al., 2009).

El proceso de aprendizaje consiste en generar la mejor predicción $\hat{Y}$ dados los vectores X e Y que representan a la variable predictora y a la variable de respuesta respectivamente.

Los modelos predictivos pueden clasificarse según el tipo y la disponibilidad de la variable de respuesta:

Si la variable de respuesta es cuantitativa, el modelo predictivo se denomina modelo de regresión, y si la variable de respuesta es cualitativa el modelo de predicción se denomina modelo de clasificación (Hastie et al., 2009)

El proceso de aprendizaje es supervisado si genera una predicción $\hat{Y}$ dadas los vectores *X* e *Y*.

El proceso de aprendizaje es no supervisado si para cada elemento de *X* no existe un elemento de *Y*.

Los modelos lineales clásicos imponen restricciones sobre la estructura de la data (Kutner, 2005) y suelen generar predicciones estables, pero poco exactas si el proceso bajo estudio es complejo, mientras que los modelos de aprendizaje automático permiten resolver los problemas de no linealidad, multicolinealidad y no normalidad de la data generando predicciones más exactas (Hastie et al., 2009), siempre y cuando sean usados

correctamente, estos métodos pueden extraer la mayor cantidad posible de información de las variables.

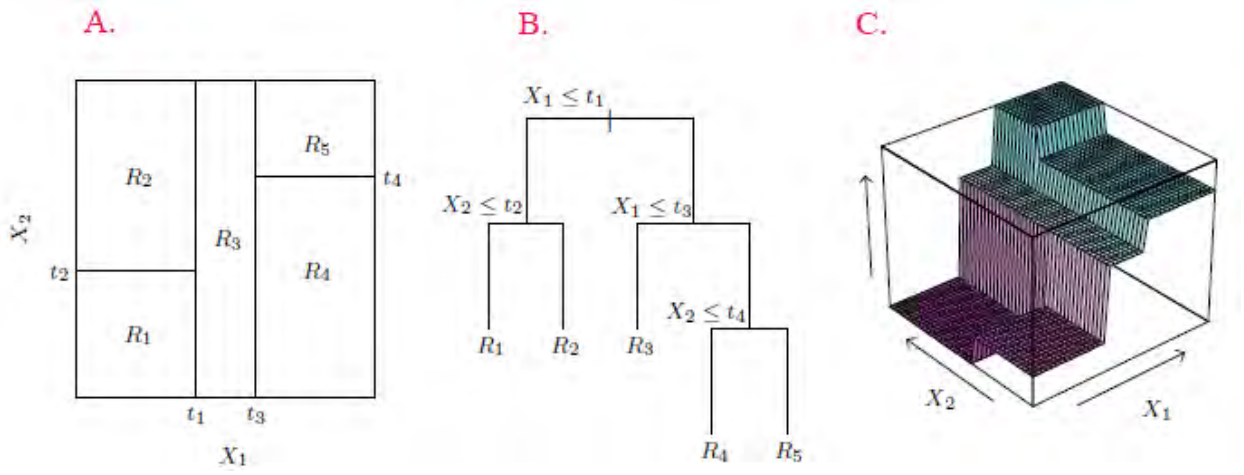
4.15. Arboles de regresión

Sea $\boldsymbol{Z} = (z_1, z_2, \dots, z_N)$ la data de entrenamiento con $z_i = (x_i\ , y_i)$ donde x_i contiene las p covariables $\mathrm{x_i} = \mathrm{x_{i1}}, \mathrm{x_{i2}}, \dots, \mathrm{x_{ip}}$ para cada i = 1, …, N, e y contiene a la variable de predicción, un árbol de regresión divide sucesivamente la data de entrenamiento en un conjunto de M regiones $R_1, \dots, R_M$ (figura 5). Según Hastie et al. (2009, p. 356) el árbol estima y como una contante γ_m en la región R_m, es decir $x \in R_M \Rightarrow \hat{g}(x) = \gamma_m$, por lo tanto un árbol de regresión se puede formalizar de la siguiente manera:

$$T\ (x, \Theta) = \sum_{m=1}^{M} \gamma_m I\{\ \mathrm{X}\ \in\ R_m\ \}$$

Ecuación 4

Figura 5*.Formación de un árbol de regresión*



Nota: A. Primero se divide $\mathrm{X_1}$ en $\mathrm{t_1}$, luego la región $\mathrm{X_1} \le \mathrm{t_1}$ se divide en $\mathrm{X_2} = \mathrm{t_2}$ y la región $\mathrm{X_1} > \mathrm{t_1}$ se divide en $\mathrm{X_1} \le \mathrm{t_3}$, finalmente $\mathrm{X_1} > \mathrm{t_3}$ se divide en $\mathrm{X_2} = \mathrm{t_4}$. El árbol de regresión

correspondiente a las particiones se observa en B, el resultado es la división en las regiones $R_1, \dots, R_5$. C muestra la gráfica de la función de regresión. Tomado de Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* [Elementos de aprendizaje estadístico]. Springer, New York.

Con parámetros $\Theta = \{R_m, \gamma_m\}$, donde I $\{X \in R_m\}$ es una función indicadora, que toma el valor $I = 1$ si $X \in R_m$ e igual a cero si $X \notin R_m$ y γ_m es una constante que debe ser estimada.

Según Kempen (2017) la división de la data de entrenamiento se realiza con el objetivo de agrupar en una región R_m los valores de la variable de respuesta más homogéneos posibles para lo cual se debe elegir la variable de división, el valor de tal variable donde ocurre la división y la topología o forma del árbol. Si el espacio dimensional se divide en M regiones $R_1, R_2, \dots, R_M$ y modelamos la variable de respuesta como una constante en cada región la función de regresión estimada es:

$$\hat{g}(\boldsymbol{x}) = \sum_{m=1}^{M} \hat{\gamma}_m \mathrm{I}\{X \in R_m\}$$

Ecuación 5

Se puede calcular γ_m mediante la minimización de la suma de cuadrados $\sum(y_i - \hat{g}(\mathbf{x_i}))^2$, es posible demostrar que el $\hat{\gamma}_m$ óptimo es simplemente el promedio de y_i en la región R_m:

$$\hat{\gamma}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

Ecuación 6

Donde $N_m = \#\{x_i \in R_m\}$ es el número de observaciones dentro de la región R_m. Por lo tanto, el árbol estima la variable de respuesta como la media de y dentro de cada región R.

Cada rama del árbol termina en un nodo terminal, comúnmente denominado como una "hoja" que representa una región R_m. Un conjunto de observaciones caen en una y exactamente una región (hoja), y cada hoja se define de forma única por un conjunto de reglas de división que la producen. Para un árbol de regresión, el nodo terminal es un único valor.

De los muchos tipos de algoritmos que implementan arboles de regresión, posiblemente el más popular en *machine learning* sea el algoritmo CART (*Classification and Regression Trees*) propuesto por Breiman et al (1998).

4.16. Random forest

Random Forest o bosques aleatorios (Breiman, 2001) es una colección de árboles de predicción (figura 7), es una modificación sustancial del método de *bagging* (Hastie et al., 2009, p. 282) que construye un ensamble de árboles de-correlacionados y luego los promedia. Matemáticamente tiene la siguiente forma:

$$g(\boldsymbol{x}, T, \Theta_b), \quad b = 1, \dots, B$$

Ecuación 7

La idea principal en *random forest* es la reducción de la varianza mediante *bagging* al reducir la correlación entre los árboles. Esto se alcanza en el proceso de construcción de los árboles a través de la selección aleatoria de las covariables para ajustar el árbol, de esta forma se introducen dos niveles de aleatoriedad en el modelo (Hastie et al., 2009).

Cuando se genera un árbol en una submuestra de *bootstrap*, antes de cada división se seleccionan $m \leq M$ covariables de forma aleatoria como posibles candidatas para realizar la división, donde M es el número total de covariables. Usualmente *m* puede tomar la forma $m = M/3$, pero en general es un parámetro que debe optimizarse.

Random forest se puede implementar a través del siguiente algoritmo: Para *b = 1* hasta *B* extrae una submuestra de *bootstrap* $\boldsymbol{Z}^{*b}$ de tamaño N de la data de entrenamiento. Ajusta un árbol de *random forest* T_b a la submuestra $\boldsymbol{Z}^{*b}$**,** al repetir recursivamente los siguientes pasos para cada nodo terminal del árbol, hasta que un tamaño mínimo de nodo n_{min} es alcanzado: i) Selecciona *m* covariables aleatoriamente de las *M* covariables de la data de entrenamiento, ii) selecciona la mejor variable y punto de división entre las m alternativas, y iii) divide el nodo en nodos secundarios o ramificaciones. Finalmente se genera el ensamble de árboles $\{T_b\}_1^B$.

Después de que B árboles $\{T(x;\ \Theta_b)\}_1^B$ se han generado mediante *bootstraping*, el predictor de *random forest* $\hat{g}_{rf}^B(x)$ en el punto *x* se halla mediante *bagging*

$$\hat{g}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\ \Theta_b).$$

Ecuación 8

Donde Θ_b es un vector que caracteriza el b-ésimo árbol en términos de sus parámetros: variables de decisión, número de nodos y valores de los nodos terminales; b es una submuestra de *bootstrap*, B es el número total de árboles, y $T(x;\ \Theta_b)$ es el árbol de regresión ajustado a la submuestra b.

Una propiedad del parámetro *m* es que si este disminuye, también disminuye la correlación entre cualquier par de árboles en el ensamble y por lo tanto se reduce la varianza del promedio de todos los árboles.

Las principales ventajas del *random forest* son:

- *Random forest* es una técnica muy versátil y flexible que no requiere la asunción de normalidad de la data (Hastie et al., 2009), adicionalmente tolera muy bien un gran número de covariables y multicolinealidad entre ellas.
- A diferencia de muchos otros estimadores no lineales, un modelo de *random forest* se puede ajustar en una sola fase, la validación del modelo se realiza mientras este se ajusta. Esto es posible gracias a las "muestras fuera de la bolsa" (Breiman & Cutler, 2003).
- La capacidad de las covariables para explicar la varianza del modelo se puede determinar mientras el modelo se ajusta, lo que permite mayor capacidad de interpretación del modelo (Hastie et al., 2009, p. 593)

La principal desventaja de *random forest* es:

- Que ignora por completo la ubicación de las observaciones y por lo tanto no considera el fenómeno de autocorrelación espacial de los datos, lo cual puede llevar a graves errores si esta ocurre con alta intensidad en la zona de estudio, sin embargo existen implementaciones que consideran el componente espacial, por ejemplo Hengl et al (2018) proponen usar diferentes métricas de distancia y ubicación para considerar el fenómeno de autocorrelación.

4.17. Evaluación de modelos de aprendizaje automático.

Para Hastie et al. (2009) la generalización de un método de aprendizaje se relaciona con su capacidad para predecir nueva información. La evaluación de la generalización de un modelo es extremadamente importante en la práctica, ya que guía la selección del método de aprendizaje o modelo, y nos da una medida de la calidad del modelo elegido.

4.17.1. Validación cruzada.

El proceso de validación cruzada de *K* iteraciones divide aleatoriamente la data de entrenamiento $\boldsymbol{\mathcal{T}}$ en *K* regiones de igual o aproximado tamaño, de las cuales *K*-1 regiones se utilizan para ajustar el modelo, simultáneamente calcula el error de predicción del modelo al predecir la data de la región *k*-ésima no utilizada para ajustarlo. Este procedimiento se repite *k = 1,2,...K* veces y finalmente combina las K estimaciones del error de predicción (Hastie et al., 2009, p. 242).

Sea $k = \{1, \dots, N\} \rightarrow \{1, \dots, K\}$ una función que determina la región a la cual la observación i-ésima es asignada aleatoriamente y $\hat{\mathrm{g}}^{-\mathrm{k}}(\mathrm{x})$ la función de regresión estimada con la *k*-ésima región de la data removida. La estimación mediante validación cruzada del error de predicción es:

$$\mathrm{CV}\,(\hat{\mathrm{g}}) = \frac{1}{N}\sum_{i=1}^{N} L\,(\,y_i, \hat{\mathrm{g}}^{-\mathrm{k(i)}}(\mathrm{x}_i)\,)$$

Ecuación 9

La validación cruzada estima efectivamente el error medio de predicción del modelo, K puede tomar cualquier valor dentro de los reales, valores típicos son K = 5 o 10 (Breiman et al., 1998). El caso donde K = N se denomina “validación cruzada con uno fuera”.

4.17.2. Explicaciones interpretables locales – LIME.

LIME fue introducido por (Ribeiro et al., 2016). El objetivo general de *LIME* es ajustar un modelo interpretable en la localidad de un punto de predicción para entender los factores que afectan una predicción (van Heumen, 2019). La idea es que un modelo complejo puede ser aproximado localmente por un modelo más simple, generalmente lineal como regresión LASSO (Tibshirani, 1996) o árboles de regresión.

Queremos encontrar un modelo que localmente aproxima un modelo más complejo, denominado de "caja negra" (Biecek & Burzykowski, 2021) f(x), cerca de un punto de interés $\underline{x}_*$.

Si G es la clase de modelos simples interpretables como un modelo lineal o un árbol de decisión. Para encontrar la aproximación adecuada, se minimiza la siguiente funcion de perdida:

$$\hat{g} = \arg min_{g \in G} L\left\{f, g, v(\underline{x}_*)\right\} + \Omega(g)$$

Ecuación 10

Donde el modelo g() pertenece a la clase de modelos G, $v(\underline{x}_*)$ define un vecindario local cerca a $v(\underline{x}_*)$, en el cual se hará la aproximación al modelo f(), L() es una funcion que mide la discrepancia entre el modelo f() y el modelo g() en el vecindario $v(\underline{x}_*)$ y $\Omega(g)$ es una penalidad a la complejidad del modelo g(). $\Omega(g)$ tiene la finalidad de limitar el tipo de modelos usados, si usan modelos con el mismo número de coeficientes $\Omega(g)$ puede ser omitido.

Es fundamental notar que los modelos g () y f () operan en diferentes espacios de covariables. El modelo complejo $f(x)$: $\chi \rightarrow \mathcal{R}$ se define en un espacio p- dimensional, donde χ corresponde a las p covariables usadas en el modelo. El modelo interpretable $g(x)$: $\tilde{\chi} \rightarrow \mathcal{R}$ esta definido en espacio q-dimensional $\tilde{\chi}$ donde q << p, denominado el espacio de "representación interpretable" (Ribeiro et al., 2016).

LIME será implementado mediante la librería *lime* de R (R Core Team, 2013), Según (Hvitfeldt et al., 2022) *LIME* sigue la siguiente aproximación:

- Para explicar cada predicción, permuta la observación n veces.
- Mediante el modelo complejo que se desea explicar predice las n observaciones permutadas.
- Calcula la distancia entre todas las permutaciones a la observación original.
- Convierte las distancias a una métrica de *similaridad* estadística.
- Selecciona m covariables que mejor describen el resultado del modelo complejo en la data permutada.
- Ajusta un modelo simple a la data permutada, explicando los resultados del modelo complejo con las *m* covariables en cada punto de la data permutada ponderada con su respectiva *similaridad* respecto a la observación original.
- Genera los pesos de las *m* covariables en base al modelo simple, y usa estos como explicación del comportamiento local (cerca del punto de predicción) del modelo complejo.

Beneficios de *LIME:*

- *LIME* es agnóstico respecto al modelo, no sigue ningún presupuesto sobre el tipo de estructura del modelo a aproximar.
- *LIME* permite obtener una representación interpretable, ya que el espacio de las covariables es transformado a una forma más interpretable y un espacio dimensional más pequeño.

- *LIME* ofrece fidelidad local, las explicaciones están localmente ajustadas al modelo más complejo.

4.17.1. Análisis de componentes principales.

El análisis de componentes principales (PCA, por sus siglas en inglés) aplica una transformación lineal que transforma un conjunto de variables correlacionadas en factores no correlacionados, que a su vez son capaces de explicar de forma máxima la varianza total de las variables (Wackernagel, 2010).

Según Kuhn y Johnson (2013) el método de PCA busca descubrir combinaciones lineales de las variables bajo estudio, conocidas como componentes principales (PC), los cuales capturan la mayor varianza posible de la data. El primer PC se define como la combinación lineal de los predictores que capturan la mayor variabilidad de todas las posibles combinaciones lineales. Los siguientes PC se calculan de tal manera que sus combinaciones lineales capturen la mayor variabilidad restante mientras asegura la no correlación con los demás PC. Matemáticamente el j-ésimo PC puede formularse de la siguiente manera:

$$\text{PC}_j = \left(a_{j1} \cdot \text{Predictor 1}\right) + \left(a_{j2} \cdot \text{Predictor 2}\right) + \cdots + \left(a_{jP} \cdot \text{Predictor P}\right)$$

Ecuación 11

Donde P es el número total de variables predictoras. Los coeficientes $a_{j1}, a_{j2}, \ldots, a_{jP}$ se denominan pesos de los componentes o carga del predictor (Kuhn & Johnson, 2013) y permiten entender qué variables predictoras son las que mayor influencia tienen en cada PC.

El PCA tiene múltiples usos en estadística multivariada, entre ellas:

- Interpretación de matrices de correlación.
- Estimación de la importancia de las variables (Kuhn & Johnson, 2013, p. 40)

El análisis de componentes principales es el método más utilizado de análisis de datos multivariantes debido a la simplicidad de su álgebra y a su interpretación directa (Wackernagel, 2010) y ha sido utilizado para estudiar la importancia de variables predictivas en el modelamiento de procesos edáficos..

4.18. Validación de estimaciones remotas de humedad del suelo.

La validación de estimaciones satelitales de la humedad del suelo tiene el objetivo de cuantificar el error de tales estimaciones a través de su comparación analítica con mediciones de campo (Gruber et al., 2020, p. 9), esta se realiza a través del establecimiento de sitios de validación (NASA, 2014) los cuales deben cumplir ciertos requisitos, descritos en la tabla 6.

Tabla 6. Requerimientos específicos de sitios de validación de estimaciones remotas de humedad del suelo.

Profundidad de medición	Mínima: 0- 5 cm Recomendada: 0 – 100 cm
Sensor	Calibración adecuada del sensor mediante comparaciones con el método gravimétrico (G. Topp & Ferré, 2018).
Cantidad de puntos de observación	Mínimo: 6 Recomendado: 15

Continuación tabla 6

Agregación espacial	Realizada mediante técnicas aceptables (Crow et al., 2012a).
Disponibilidad de información	Mínimo: 1 a 4 semanas.
Información adicional	Suelos, vegetación y meteorología.

Fuente Basado en NASA (2014, p. 128)

Si el contenido volumétrico de humedad del suelo real es Θ (en realidad la media espacial del sitio de validación) y Θ_E es la estimación remota de la humedad del suelo del satélite SMAP, el cuadrado medio del error es

$$RMSE = \sqrt{E[(\Theta_E - \Theta)^2]}$$

Ecuación 12

Donde E es la expectación. Este estadístico penaliza las desviaciones o diferencias cuadráticas de las estimaciones remotas respecto a la humedad real en campo y es una medida sencilla y de fácil comprensión de la precisión de estimación.

Un sitio de validación puede estar compuesto por uno o varios píxeles del producto de estimación satelital. En cada sitio se pueden estimar los estadísticos de validación, el requisito de precisión implica que para el sitio de validación N_i para el que están disponibles observaciones *in situ*, los productos de humedad del suelo SMAP deben satisfacer (NASA, 2014):

$$\frac{1}{N_i}\sum_{i=1}^{N_i}[RMSE_i] \leq 0.04\ \mathrm{cm^3 cm^{-3}}$$

Ecuación 13

El problema más común de la validación es la representatividad de las mediciones *in situ* usadas para la validación. Las propias mediciones en campo solo pueden proporcionar una estimación de la verdadera humedad del suelo. Puede haber sesgo y errores en las estimaciones debido a errores del sensor utilizado para medir la humedad en campo, pero también, y tal vez lo más importante, debido a que no se mide la humedad del suelo en toda el área sino solo en un número finito de puntos de esta, por lo tanto, se introducen errores estadísticos en la estimación de la media espacial de la humedad del suelo (NASA, 2014).

Das et al. (2019a) y ONeill et al. (2020) demostraron la precisión del producto SMAP-L2-E y la capacidad de este para lograr los objetivos científicos de la misión mediante la producción de la humedad del suelo de alta resolución espacial (9 km) con una precisión promedio estimada mediante un RMSE menor a 0.05 cm^3cm^{-3}.

Bai et al. (2019) compararon el producto SMAP-L3-E con mediciones *in situ* en el norte de China, obteniendo un RMSE de 0.074 cm^3cm^{-3}, para tal estudio se usaron solamente 4 estaciones de monitoreo de humedad.

Singh et al. (2019) a través de mediciones *in situ* con el sensor *ThetaProbe ML3* validaron el producto SMAP-L3-E en una región de India desde el 2017 al 2018, encontrando diferencias en el RMSE durante épocas secas (0.089 a 0.104 cm^3cm^{-3}) y épocas húmedas (0.017 a 0.051 cm^3cm^{-3}) concluyendo que el producto SMAP-L3-E

cumple con la precisión recomendada por la NASA, pero esto depende del nivel de saturación de humedad del suelo.

En el contexto latinoamericano el único estudio reportado es el de Hernandez-Sanchez et al. (2020) que como parte del Experimento de Hidrología Terrestre en México (THExMEX-18) evaluaron un producto del SMAP en áreas de cultivo de maíz en México durante el año 2019 y 2020 comparando las estimaciones remotas con data *in situ* de sensores de humedad, concluyendo que el producto obtiene un RMSE menor a 0.04 cm^3cm^{-3}, es decir un error menor al 4% de humedad volumétrica.

Además del RMSE, otro estadístico usado comúnmente en la validación de productos satelitales de humedad del suelo (Beck et al., 2021; Chan et al., 2016) es el coeficiente de correlación de Pearson (σ), el cual se define de la siguiente forma en el caso de la humedad del suelo:

$$\sigma = \frac{n\sum \theta_E\,\theta - \sum \theta \sum \theta_E}{\left[n\sum \theta^2 - (\sum \theta)^2\right]\left[n\sum {\theta_E}^2 - (\sum \theta_E)^2\right]}$$

Ecuación 14

4.18.1. Métodos de capacitancia para la medición de la humedad del suelo.

El *ThetaProbe* es un sensor de capacitancia inventado por Gaskin y Miller (1996) y comercializado por *Delta-T Devices* (2013). El *ThetaProbe* se basa en la teoría de líneas de transmisión para medir el voltaje de frecuencia fija que resulta de la discrepancia de impedancia entre un cable coaxial y el suelo, lo cual permite la estimación de la permisividad, ε, del medio bajo medición.

Se han producido tres versiones del *Theta Probe* que se han designado como ML1x, ML2x y ML3. Hay una pequeña diferencia en la respuesta a la permitividad y la conductividad eléctrica entre los tres diseños (Miller & Gaskin, 1999; Delta-T, 2013). *Delta-T* (2013) afirma que el modelo ML3 tiene menos sensibilidad a la temperatura y conductividad eléctrica del que los modelos ML1x y ML2x, las principales características técnicas del *ThetaProbe* ML3 se muestran en la tabla 7.

. ***Tabla 7****. Propiedades técnicas del sensor ThetaProbe*

Variable de medición	Contenido volumétrico de humedad del suelo.
Precisión	0.01 cm^3cm^{-3} (con calibración específica).
Rango de medición	Completamente preciso entre 0 a 0.5 cm^3cm^{-3}
	Rango completo de medición 0 a 1.0 cm^3cm^{-3}
Rango de salinidad	50 a 500 $m \cdot Sm^{-1}$
Rango de temperatura	Precisión completa entre 0 a 40°C.
Señal de salida	0 a 1.0 V correspondiente a 0 a 0.6 cm^3cm^{-3} .
Requerimiento de poder	5 a 14 V, aproximadamente 18 mA por 1 segundo.
Volumen de medición	60·30 mm de diámetro.

Elaborado por Marcelo Bueno Dueñas.

4.18.1. Monitoreo de la humedad del suelo.

De Gruijter et al. (2006) definen monitoreo como el muestreo continuo en el tiempo, con o sin un soporte espacial realizado con el objetivo de describir el comportamiento dinámico de una variable o un proceso.

Se ha sugerido que la cantidad de puntos de monitoreo en un sitio de validación de estimaciones remotas de la humedad del suelo depende principalmente de la heterogeneidad de las propiedades físicas, químicas y biológicas del suelo, la topografía del terreno y de la vegetación; Famiglietti et al. (2008) encontraron que 30 puntos de muestreo de humedad del suelo son adecuados para una resolución espacial de 50 Km; Crow et al., (2012) resumen los resultados de estudios de validación en la tabla 8.

Tabla 8. *Resultado de estudios de validación de estimaciones remotas de humedad del suelo respecto a la cantidad de puntos de monitoreo.*

Estudio	Precisión (Error absoluto)	Cantidad de puntos de monitoreo
Brocca et al. (2010)	0.02 cm^3 cm^{-3}	4 - 15
Brocca et al. (2007)	0.02 cm^3 cm^{-3}	15-35
Jacobs (2004)	0.02 cm^3 cm^{-3}	3-32
Wang et al. (2008)	0.05 cm^3 cm^{-3}	41

Continuación tabla 8

Famiglietti et al. (2008)	0.03 cm^3 cm^{-3}	7-17
Famiglietti et al. (2008)	0.02 cm^3 cm^{-3}	34
Western et al. (2004)	0.02 cm^3 cm^{-3}	14

Fuente: Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D., & Walker, J. P. (2012). Upscaling sparse ground-based soil moisture observations for the validation of coarse resolution satellite soil moisture products (Agregación de mediciones de campo de la humedad del suelo con fines de validación de productos satelitales de humedad del suelo). Reviews of Geophysics.

En un estudio más reciente Das et al. (2019b) validaron del producto L2-SMS-P de la misión SMAP mediante monitoreo *in situ* con mediciones distribuidas en redes de sensores de poca densidad espacial, desde abril del 2015 hasta octubre del 2018; en el mencionado estudio se utilizó una estación de medición de humedad del suelo por cada pixel del producto SMAPL3E; concluyendo que a pesar de los posibles errores por la baja representatividad espacial de la humedad del suelo medida por una sola estación de monitoreo, la relación entre la humedad del suelo *in situ* y la estimación remota es aceptable.

4.18.2. Coeficiente de correlación cuantílico multiescala (MQCC).

Previos estudios de *downscaling* espacial se han enfocado principalmente en el coeficiente de correlación σ como principal medida de validación (Atkinson, 2013; Q. Chen et al., 2020; S. Chen et al., 2019).

El coeficiente de correlación σ permite medir la correspondencia entre series de tiempo observadas *in situ* y las predichas por modelos de superficie o estimadas por productos satelitales en términos de su variabilidad temporal, y por lo tanto evalúa el aspecto más importante de las series de tiempo de humedad del suelo para la mayoría de aplicaciones practicas (Entekhabi et al., 2010; Gruber et al., 2020), además es una de las dos métricas recomendadas para la validación del SMAP (Beck et al., 2021; Entekhabi et al., 2010; Jackson et al., 2010).

Sin embargo, la principal dificultad al aplicar esta métrica para la validación de la desagregación en este estudio fue que es necesario tener una gran cantidad de estaciones de monitoreo para obtener resultados confiables. Por ejemplo, en estudios de validación previos a nivel global se usaron aproximadamente 800 estaciones de monitoreo de humedad del suelo (Beck et coal., 2021) ubicadas principalmente en USA y Europa.

El análisis de correlación tradicional se enfoca en la media e implícitamente se basa en el análisis de regresión lineal ordinario, el cual requiere que la variable dependiente se distribuya normalmente, y es proclive a *outliers*, heterocedasticidad y otras desviaciones del modelo lineal ordinario. Adicionalmente el coeficiente de correlación global no permite analizar la dependencia en regiones específicas de las distribuciones de ambas variables.

El coeficiente de correlación cuantílico ${\rho_\tau}^{X,Y}$, derivado de la combinación de regresión cuantílica y del coeficiente de correlación, pude ser empleado para analizar diferencias de correlación entre series de tiempo a diferentes niveles cuantílicos (en diferentes regiones de la distribución de las variables)(Choi & Shin, 2022).

En años recientes, el análisis multi-escala ha recibido mayor atención en el estudio de series de tiempo. Los estudios multi-escala permiten revelar fluctuaciones características existentes a diferentes agregaciones temporales. Los métodos de análisis de granularidad gruesa, análisis de wavelet y descomposición de moda empírica son métodos populares de análisis multi-escala (Xu et al., 2020). El método de granularidad gruesa (*course graining*) es el más popular debido a su facilidad de implementación y generalización (Xu et al., 2020). El método de granularidad gruesa consiste en agrupar "bloques" de datos consecutivos y promediarlos para generar una nueva serie de tiempo agregada a diferente escala temporal que la original.

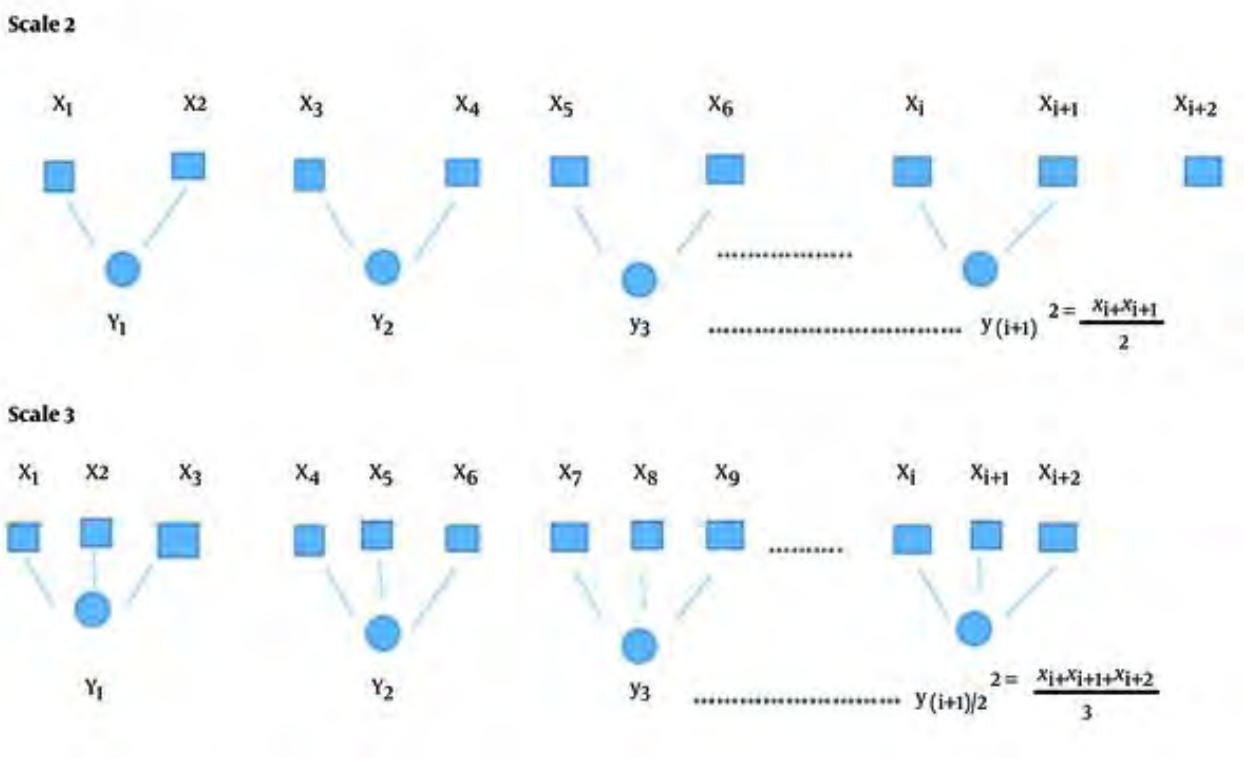
Dada una serie de tiempo con N datos, el procedimiento del método de granularidad gruesa es el siguiente: para la primera escala temporal, no hay diferencias entre la serie de tiempo original; para las escalas (n ≥2), la serie de tiempo es divida en bloques consecutivos no sobrepuestos, y cada bloque contiene n observaciones. Luego se calcula el promedio de cada bloque, con el cual se construye una serie de tiempo, este proceso puede ser descrito de la siguiente forma:

$$X_j^n = \frac{1}{n}\sum_{(j-1)n+1}^{jn} x_i, \;\; 1 \le j \le \frac{N}{n}$$

Ecuación 15

Donde j representa el bloque, n el número observaciones por bloque y X_j^n es el promedio de n observaciones del bloque j-ésimo (figura 6).

Figura 6.*Metodología de granularidad gruesa aplicada en el estudio*



Nota: Se muestra el método para las primeras 3 escalas temporales.

El método MQCC consiste en evaluar la correlación entre dos series de tiempo (humedad del suelo predicha por un modelo y la observada en campo) mediante los cuantiles condicionales a múltiples escalas temporales.

Dada una serie de tiempo con *N* datos el método de granularidad gruesa agrupa *n* datos y los promedia por bloques para generar una nueva serie de tiempo agregada, la serie de tiempo a la nueva escala será más corta que la serie de tiempo original (Xu et al., 2020).

En el análisis de correlación tradicional, (X, Y) son consideradas variables aleatorias que poseen momento de segundo orden. Si $\rho_X = Var(X), \rho_Y = Var(Y), y\ \rho_{XY} = Cov(X,Y)$, los coeficientes de regresión $\beta_{X,Y} = \frac{\rho_{XY}}{\rho_X}$ y $\beta_{Y,X} = \frac{\rho_{XY}}{\rho_Y}$ donde $\beta_{X,Y}$ y $\beta_{Y,X}$ se obtienen minimizando la suma de cuadrados del error. El coeficiente de correlación $\rho = \frac{\rho_{XY}}{\sqrt{\rho_X \rho_Y}} = \text{sign}(\beta_{X,Y})\sqrt{\beta_{X,Y}\beta_{Y,X}}$ es la media geométrica de los dos coeficientes de correlación.

El coeficiente de correlación ρ_τ en el cuantil τ se define a su vez como la media geométrica de los dos coeficientes de regresión en el cuantil τ $\beta_{X,Y}(\tau)$ y $\beta_{Y,X}(\tau)$ y se expresa como:

$$\rho_\tau{}^{X,Y} = \text{sign}\left(\beta_{X,Y}(\tau)\right)\sqrt{\beta_{X,Y}(\tau)\beta_{Y,X}(\tau)}\,, \tau \in (1,0)$$

Ecuación 16

Donde $\left(\alpha_{2,1}(\tau), \beta_{2,1}(\tau)\right) = argmin_{\alpha,\beta} L_\tau^{X,Y}(\alpha,\beta)$ y $\left(\alpha_{1,2}(\tau), \beta_{1,2}(\tau)\right) = argmin_{\alpha,\beta} L_\tau^{X,Y}(\alpha,\beta)$ y L es una funcion de perdida, generalmente cuadrática.

Para cada $\tau \in (1,0)$ el valor absoluto de ρ_τ representa la sensibilidad en el cuantil τ de una variable aleatoria a cambios en otra variable. Las estimaciones de los coeficientes de regresión en diferentes cuantiles son generalmente diferentes, lo que significa que dos variables están relacionadas con diferente magnitud a diferentes niveles. Por lo tanto, es necesario comparar las diferencias entre ρ_τ a diferentes valores de τ. Por ejemplo $\rho_{0.05} > \rho_{0.95}$ significa que el cuantil 0.95 de la variable es menos sensitiva que el cuantil 0.05 a cambios en otra variable. Si ocurre que, por ejemplo $\rho_{0.05} > \rho_{0.95}$ la correlación entre X y Y puede ser descrita como heterogénea (Xu et al., 2020).

4.18.3. Gráficos de dispersión y Gráfico Q-Q.

Una gráfica de dispersión permite representar la relación entre dos variables en el plano cartesiano, este tipo de grafica permite ver el tipo de relación entre dos variables, si su relación es lineal o no lineal o positivo o negativa, es un método muy útil de análisis exploratorio. Un gráfico cuantil-cuantil (*q-q plot*) es una técnica de análisis gráfico que permite determinar si dos conjuntos de datos provienen de poblaciones con distribuciones similares. Es un método no paramétrico de comparación de dos muestras. Un *q-q plot* es un gráfico que muestra los cuantiles del primer conjunto de datos respecto a los cuantiles del segundo grupo de datos.

Generalmente una línea de 45 grados también se grafica**.** Si los dos conjuntos de datos provienen de una población con la misma distribución, los puntos deberían mantenerse cerca de la línea de 45°, a mayor la divergencia de los puntos respecto a la línea, mayor es la evidencia de que los dos conjuntos de datos provienen de una población con distribuciones distintas. cada punto en el q-q plot corresponde a valores del mismo cuantil de los dos *datasets*. Un *q-q plot* que sigue una tendencia lineal refleja que las dos poblaciones están correlacionadas, y un *q-q plot* no lineal indica que las distribuciones siguen tendencias diferentes (Andersen & Dennison, 2019)**.**

V. DISEÑO DE INVESTIGACIÓN

5.1.1. Tipo de investigación

La presente tesis se enmarca en el tipo de investigación correlacional no experimental.

5.1.2. Ubicación temporal

Las actividades de obtención de información satelital, monitoreo, análisis y redacción se llevaron a cabo desde Julio del 2021 hasta Agoto del 2022.

La data satelital utilizada abarca observaciones desde el 2015 hasta el 2022.

5.1.3. Ubicación política

El área de recopilación de información satelital (figura 8) queda dentro del departamento del Cusco. Abarca completamente las provincias de Calca y Canchis, y parcialmente las provincias de Canas, Acomayo, Cusco, Anta y Urubamba.

El área de monitoreo queda dentro del departamento del Cusco, Provincia de Cusco y Distrito de San Jerónimo.

El área de validación queda dentro del departamento del Cusco, Provincia de Cusco y Distritos de San Jerónimo.

5.1.4. Ubicación geográfica

5.1.4.1. Ubicación del área de estudio.

El área de estudio seleccionada abarcó una superficie de aproximadamente 8 328 km^2, que va desde 72.30°O a 70.83° O y desde 13.13°S a 14.68° S; esta área es lo suficientemente grande para cubrir un número representativo de píxeles del producto

SMAP-L3-E de 9 km de resolución espacial (aproximadamente 400 píxeles) y por lo tanto permite tener una cantidad adecuada de observaciones satelitales.

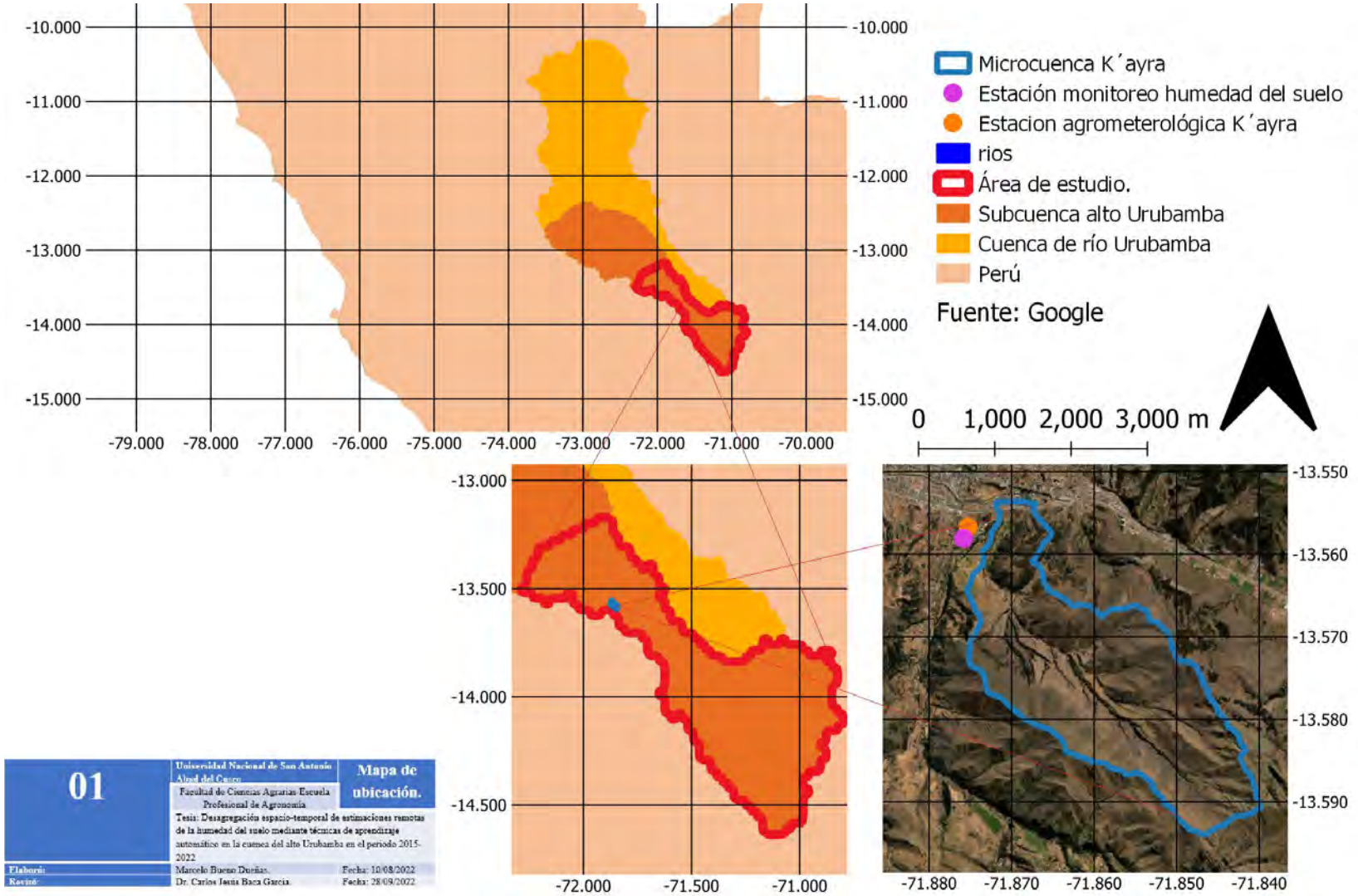
5.1.4.2. Ubicación de la estación de monitoreo de la humedad del suelo y área de validación.

La estación de monitoreo diario de la humedad del suelo (punto púrpura en la figura 7) se seleccionó con el criterio de accesibilidad y representatividad de un píxel del producto SMAP-L3-E del SMAP; se ubica a 160 metros de la estación meteorológica Granja K´ayra (punto naranja en la figura 7) a una elevación de 3216 m, con las siguientes coordenadas geográficas:

- Latitud: 13.558° S
- Longitud:71.876° O

Cabe mencionar que el área de monitoreo es un punto representativo del área de validación, (polígono azul en la figura 7)

***Figura** 7.Ubicación del área de estudio*



Nota: EPSG:4326 - WGS 84. Coordenadas geográficas.

5.1.5. Ubicación hidrográfica.

El área de recopilación de información satelital se localiza hidrográficamente dentro de la Cuenca Hidrográfica del Rio Urubamba-Vilcanota (figura 7).

Específicamente el área de monitoreo se ubica en la subcuenca Huatanay.

El área de validación fue la extensión completa de la microcuenca K´ayra.

5.1.6. Topografía.

La elevación media del área de estudio según el DEM SRTM3 (Modelo Digital de Elevación SRTM3) es de 3746.95 m con una desviación estándar de 399.48 m.

5.1.7. Uso del suelo y cobertura vegetal.

En el área de estudio predominan los matorrales con un 46.78 % de su superficie total, seguidos por pasturas con un 34.92%, áreas de cultivo con un 13.67%, bosques con un 2.83 %, y superficies artificiales como pueblos y ciudades con 1%.

5.1.8. Climatología.

La precipitación en la zona de estudio exhibe un fuerte ciclo estacional como ha sido demostrado por Imfeld, et al. (2021), la precipitación media diaria historia (1981-2016) mensual en los meses de diciembre, enero y febrero es 4.85 mm $día^{-1}$, para los meses de marzo, abril y mayo es de 1.94 mm $día^{-1}$, para los meses de junio, julio y agosto es de 0.21 mm $día^{-1}$ y para los meses de septiembre, octubre y noviembre es de 1.66 mm $día^{-1}$.

5.2. Materiales y métodos

5.2.1. Materiales

5.2.1.1. Equipos

- Sensor de humedad del suelo *ThetaProbe* ML3.
- GPS multi banda Garmin tr66.
- Cámara fotográfica digital.
- Computadora personal.

5.2.1.2. Información y datos.

- Datos del producto SMAP-L3-E del SMAP para el área de estudio desde abril del 2015 hasta julio del 2022.
- Datos diarios de precipitación grillada desde 1981 hasta el 2016 del producto PISCO del SENAMHI.
- Datos diarios de precipitación del producto satelital CHIRPS a 5 Km de resolución espacial (https://climateserv.servirglobal.net/).
- Información espacial de la base de datos *SoilGrids* 2.0 (Hengl et al., 2017)
- Estimaciones de conductividad hidráulica insaturada global a 1 km de resolución espacial (Gupta et al., 2021)
- Modelo Digital de Elevación (DEM) MERIT (Yamazaki et al., 2019) para el área de estudio.
- Datos meteorológicos del SENAMHI de la estación Granja K´ayra.
- Data diaria de monitoreo de la humedad del suelo mediante el sensor *ThetaProbe*.

5.2.1.3. Software

- QGIS.
- SAGA GIS (Sistema para Análisis Automatizados Geocientíficos).
- Lenguaje de programación R.

5.3. Descripción de los métodos.

5.3.1. Procesamiento geo-espacial de los datos.

Para procesar las covariables se usaron una combinación de software GIS libre, principalmente SAGA GIS y GRASS GIS (Neteler & Mitasova, 2008) y paquetes de R (R Core Team, 2013), principalmente los paquetes *raster* (Hijmans & van Etten, 2012), sp (Pebesma & Bivand, 2005) y rGDAL (Bivand et al., 2021) para realizar re-proyecciones, re-muestreo, conversión de formatos espaciales de datos *raster*, *stacking* y operaciones geoespaciales de uso común. Los métodos utilizados en la tesis se resumen en la figura 9.

5.3.1.1. Producto SMAP-L3-E.

La información remota de humedad del suelo fue el componente principal en el proceso de *downscaling.* En el presente estudio se usó el producto de nivel 3 SMAP-L3-E derivado del radiómetro de banda L del satélite de la NASA *Soil Moisture Active Passive* (SMAP) el cual se obtuvo mediante el Sistema de Información y Data de Observación de la Tierra (EOSDIS) de la NASA cuyas características principales se resumen en la tabla 9.

El producto SMAP-L3-E representa el contenido volumétrico de humedad medio del suelo aproximadamente a 5 centímetros de profundidad (Entekhabi et al., 2010).

Figura 8.*Diagrama de flujo*

.

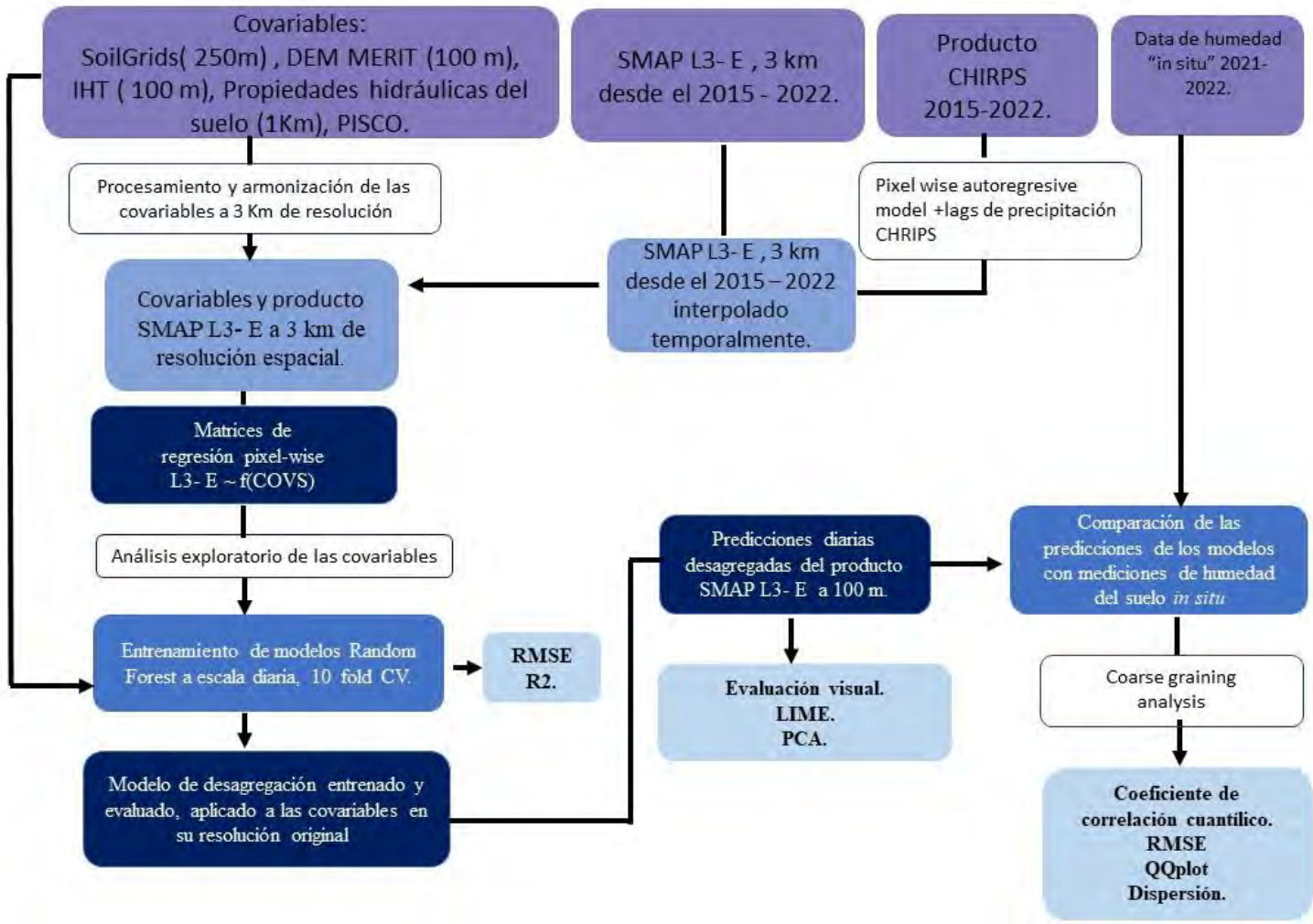


Tabla 9. *Detalles técnicos del producto SMAP-L3-E*

Variable_científica	Contenido volumétrico de humedad del suelo (cm^3cm^{-3}) aproximadamente a 5 centímetros de profundidad (Entekhabi et al., 2010)
Precisión ⋆	0.05 cm^3cm^{-3} (Das et al., 2019)
Algoritmo	Pasivo (O'Neill et al., 2020).
Formato	HDF5
Detalles de geolocalización y resolución	
Sistema de coordenadas geográficas	WGS 84
Unidades	metros
Código EPSG	4326
Resolución espacial	~ 9000 m (x)
	~ 9000 m (y)
Resolución temporal	De 3 a 4 días

⋆Precisión en lugares de validación bajo condiciones topográficas y de vegetación óptimas (S. K. Chan et al., 2016)

⋆⋆ El producto generado mediante el algoritmo SCA-V es el más preciso (ONeill et al., 2020b), por lo tanto será el usado en la investigación.

Fuente: Elaboración propia

La data del producto SMAP-L3-E se descargó de https://nsidc.org/data/spl3smp_e/versions/5 en formato *GTIFF* para cada fecha disponible. Se descargó la data desde el inicio de la misión SMAP (31 de marzo del 2015) hasta la fecha actual (julio del 2022) con el objetivo de utilizar toda la data disponible (aproximadamente 2000 *rasters*) para calibrar los modelos propuestos.

Cada imagen en formato *GTIFF* fue procesada de la siguiente manera:

- Se eliminaron los valores nulos de cada imagen, estos representan pixeles en los cuales no se pudo generar una estimación de la humedad del suelo.
- Se recortó cada imagen al área de estudio utilizando las herramientas geoespaciales de rGDAL.

Para este último proceso fue necesario realizar interpolación espacial mediante el método bilineal.

El siguiente paso consistió en unir todas las imágenes *GTIFF* para construir un objeto espacial denominado *rasterstack* (Hijmans & van Etten, 2012) el cual consiste en una colección de todas las imágenes *GTiff* para todas las fechas disponibles, este fue el objeto geoespacial básico para el procesamiento posterior porque permite ejecutar procesos y algoritmos geoespaciales de forma simultánea a toda la data en conjunto.

Existen errores en la adquisición de la temperatura de brillo y posterior estimación de la humedad del suelo ocasionados principalmente por las condiciones ambientales como la

precipitación o propiedades ópticas del terreno, estos errores se expresan como pixeles con valores anómalos y fueron removidos para los análisis posteriores.

5.3.1.2. Obtención y preprocesamiento de las covariables.

En el esquema de desagregación propuesto, fue importante establecer la relación entre las estimaciones remotas de humedad del suelo del SMAP con las variables predictivas que mejor representen los procesos e interrelaciones complejas entre la humedad del suelo y factores edáficos, topográficos y meteorológicos (Crow et al., 2012; Mohanty & Skaggs, 2001). Las variables geoespaciales generalmente utilizadas en estudios de desagregación tienen varios grados de fiabilidad científica, variabilidad espacial y correlación con la humedad de suelo (Peng et al., 2017)

5.3.1.3. Propiedades del suelo.

La disponibilidad de información analítica de propiedades del suelo a nivel de subcuenca es escasa y no suficiente para cumplir con los objetivos planteados por la investigación, por lo tanto, se usó información geoespacial de suelos derivada de la base de datos *SoilGrids.*

SoilGrids proporciona predicciones en formato *raster* para las propiedades del suelo mostradas en la tabla 10.

Tabla 10. Propiedades del suelo de SoilGrids

Propiedad del suelo⋆	Símbolo	Unidad	Descripción
Contenido de carbono orgánico	OC	$g\ kg^{-1}$	Contenido gravimétrico de carbono presente en la materia orgánica del suelo.
Densidad aparente	BD	$cg\ cm^{-3}$	Masa por unidad de volumen de suelo (Grossman & Reinsch, 2002).
Capacidad de intercambio catiónico	CIC	$cmol_{+}Kg^{-1}$	Suma total de las cationes intercambiables que un suelo puede absorber
Contenido de arcillas	CC	$g\ Kg^{-1}$	Contenido gravimétrico de minerales de tamaño menor a 1 μm.

⋆Todas las variables a 250 m de resolución espacial.

Fuente: Elaboración propia.

Recientemente Gupta et al. (2021, 2022) utilizando la misma base de datos y *SoilGrids*, derivaron la distribución global de las propiedades hidráulicas del suelo mediante *random forest* a un kilómetro de resolución espacial. Estas predicciones fueron utilizadas en el presente estudio. Un resumen de las propiedades hidráulicas usadas se aprecia en la tabla 11.

Tabla 11. Propiedades hidráulicas del suelo de Gupta et al. (Gupta et al., 2021, 2022)

Propiedad del suelo⋆	Símbolo	Unidad	Descripción
Conductividad hidráulica saturada del suelo	*KSat*	$cm\ día^{-1}$	Máxima cantidad de flujo de agua en un suelo en condiciones de saturación.
Contenido de agua en saturación.	θ_S	$cm\ cm^{-3}$	Contenido volumétrico de agua en el suelo cuando este llega a saturación.
Contenido de agua residual.	θ_r	$cm\ cm^{-3}$	Contenido volumétrico de agua mínimo posible en un suelo determinado.
Parámetros de la función de retención de humedad de van Genuchten.	$\alpha\ y\ n$	Adimensionales	Parámetros de ajuste de la función de van Genuchten, (1980) (Vereecken et al., 2019)
Alpha y n.			

⋆Todas las variables a 1 Km de resolución espacial. La data se puede acceder libremente mediante 10.5281/zenodo.5547338.

Fuente: Elaboración propia.

Se obtuvo información geoespacial de las propiedades físicas y químicas del suelo a 5 cm de profundidad a 250 metros de resolución espacial para el área de estudio usando el sistema de predicción espacial de propiedades y clases del suelo *SoilGrids* (Hengl et al., 2017) desarrollado por ISRIC (Centro de Información Internacional del Suelo) mediante la

el enlace siguiente: https://soilgrids.org/. Adicionalmente las propiedades hidráulicas del suelo fueron descargadas en formato GTiff de la dirección 10.5281/zenodo.5547338 del repositorio de zenodo (Chue, 2019)

Los *GTiffs* de propiedades del suelo fueron agregados a la resolución espacial del SMAP-L3-E y posteriormente convertidas a *stacks* de la misma manera que se hizo con la data del SMAP.

5.3.1.4. Modelo digital de elevación (DEM) e índice de humedad topográfica (TWI)

Se descargó el modelo de digital de elevación (DEM) MERIT a 90 m de resolución espacial de http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/, se realizó un recorte y reproyección al área de estudio.

El DEM fue procesado utilizando las herramientas de análisis de terreno del Sistema para Análisis Automatizados Geocientíficos (SAGA-GIS).

El software libre SAGA GIS implementa los siguientes algoritmos de enrutamiento de flujo (Neteler & Mitasova, 2008) para el cálculo del índice de humedad topográfico resumidos en la tabla 12.

Tabla 12. *Algoritmos de enrutamiento de flujo para el cálculo del índice de humedad topográfico IHT, propuestos en este proyecto de investigación.*

Nombre del algoritmo	Tipo	Descripción	Referencia
Algoritmo determinístico de ocho vecinos, D8.	SFD	El algoritmo prorratea el flujo de cada píxel hacia a un solo píxel adyacente a través de la dirección descendente más empinada.	
Algoritmo de dirección de flujo simple aleatorio, Rho8.	SFD	El algoritmo distribuye aleatoriamente el flujo a un pixel adyacente con una probabilidad proporcional a la pendiente.	
Algoritmo de enrutamiento cinemático, KRA.	SFD	El algoritmo dirige el flujo a la esquina del píxel con el valor de elevación más bajo; el valor se calcula promediando las elevaciones de los píxeles adyacentes.	
Red de modelo digital de elevación, DEMON	MFD	El algoritmo distribuye el flujo en cuatro direcciones a través de las esquinas de un píxel y asigna el flujo al mejor plano coincidente.	
Algoritmo determinístico infinito D∞.	MFD	El algoritmo distribuye el flujo en función de ocho planos triangulares determinados por el gradiente de pendiente más empinado	

Continúa Tabla 12

Modelo de relieve de Braunschweiger, BR.	MFD	El algoritmo elige el píxel más cercano al aspecto del píxel de origen y sus dos píxeles vecinos según el gradiente topográfico local descendente.	
Método de flujo de masa, MFM.	MFD	El algoritmo divide cada píxel en cuatro cuartos de píxeles, cuyo plano está determinado por las elevaciones de los píxeles enteros y dos píxeles vecinas.	(Gruber & Peckham, 2009)
Algoritmo de dirección de flujo múltiple triangular, MD∞.	MFD	El algoritmo divide cada píxel en regiones triangulares y el flujo se divide hacia píxeles vecinos, proporcionalmente a la gradiente topográfica.	
FD8	MFD	El algoritmo conduce el flujo a todos los píxeles vecinos de elevación más baja a través de un exponente de partición de flujo.	(Quinn et al., 1995)
MFD-md	MFD	El algoritmo dirige el flujo hasta todos los píxeles vecinos de elevación más baja en función de la función lineal de la gradiente topográfica máxima.	(Qin et al., 2007)

Fuente: Elaboración propia.

El procesamiento del DEM consistió en corrección hidrológica mediante el método de Wan y Lu, se calculó el índice de humedad topográfica (TWI), también llamado índice topográfico o índice topográfico compuesto (Qin et al., 2007) en su resolución espacial original, para lo cual se utilizó SAGA GIS mediante la implementación de los algoritmos de dirección de flujo único (SFD) y dirección de flujo múltiple (MFD).

Los *GTiffs* del DEM y de los índices de humedad topográficos fueron armonizados a la resolución espacial del SMAP-L3-E y posteriormente convertidas a *stacks* de la misma manera que se hizo con la data del SMAP y las propiedades del suelo.

5.3.1.5. Producto PISCO.

El producto PISCO fue obtenido en formato NETCDF del repositorio del Instituto de Clima y Sociedad de la Universidad de Columbia (http://iridl.ldeo.columbia.edu/). El preprocesamiento consistió en estimar los promedios históricos de precipitación diaria para las cuatro estaciones hidrológicas siguiendo a Imfeld et al., (2021). Posteriormente los *rasters* fueron armonizados a la resolución espacial del SMAP-L3-E y posteriormente convertidas a *stacks* de la misma manera que se hizo con la data del SMAP, las propiedades del suelo y la topografía.

5.3.1.6. Producto CHIRPS.

Se descargaron datos del producto CHIRPS (Funk et al., 2015) de precipitación grillada diaria a 5 km de resolución espacial mediante https://climateserv.servirglobal.net/ entre marzo del 2015 a julio del 2022.

Según mostró el análisis posterior de la data, la precipitación de un día particular tiene poca correlación con la humedad del suelo del mismo día, por lo menos respecto a la humedad

estimada con el producto SMAP, de tal manera que se calculó el promedio aritmético de la precipitación antecedente de los últimos 3 días y no se usó la precipitación del mismo día como covariable.

Estos datos recibieron el mismo tratamiento que las demás covariables, que consistió en la armonización a la resolución espacial del SMAP-L3-E y la conversión de los 2000 *rasters* a un solo *rasterStack*.

5.3.1.7. Matriz de regresión.

Se construyó una matriz de regresión en base a los *stacks* del produto SMAP-L3 y las covariables a la misma resolución espacial. La variable de respuesta fue el contenido volumétrico de humedad del suelo del producto SMAP-L3-E para cada fecha de data disponible; de tal manera que cada fila de la matriz de regresión corresponde a un pixel y fecha específicos del producto SMAP-L3-E; si la covariable es estática se consideró solamente la coincidencia de pixeles, sin embargo, si la covariable fue dinámica se consideró la coincidencia de pixeles y de fechas respecto a la variable de respuesta.

Esta matriz tuvo un tamaño considerablemente grande (aproximadamente 9 millones de filas) ya que en este estudio se consideraron covariables dinámicas como la precipitación que varían de forma diaria. Esta matrices fueron construidas mediante operaciones geoespaciales con la librerías *raster*, *sp* (Pebesma & Bivand, 2005) y *dplyr* en R.

El objetivo de este paso fue organizar toda la data disponible para facilitar el desarrollo de los modelos propuestos.

Posteriormente la matriz de regresión fue dividida en dos matrices, una para la desagregación temporal (para cada pixel) y otra para la desagregación espacial (para cada fecha) respectivamente.

5.3.2. Evaluación de la capacidad de desagregación espacio-temporal mediante *random forest* del producto SMAP-L3-E en el área de estudio.

5.3.2.1. Análisis exploratorio de las covariables.

Se realizó un análisis descriptivo de las covariables y del producto SMAP-L3-E, para lo cual se calcularon los estadísticos descriptivos principales como la media, desviación estándar, mediana y *skewness* entre otros.

Posteriormente se construyó una matriz de correlaciones con el objetivo de examinar las posibles relaciones estadísticas entre las covariables para detectar relaciones lineales y posible multicolinealidad de las covariables.

5.3.2.2. Construcción de los modelos de desagregación espacio-temporal.

En la presente investigación se propuso un método de desagregación espacio-temporal de las estimaciones remotas de la humedad del suelo del producto SMAP-L3-E basado en *random forest* descrito recientemente por Hengl et al. (2018), Heung et al. (2016), Zhao et al. (2018).

La distribución espacial de la humedad del suelo en el area de estudio fue estimada a una resolucion espacial aproximada de 100 metros de forma diaria. Para lo cual se utilizó la implementación *ranger* (Wright & Ziegler, 2017) dentro del ambiente de modelamiento de aprendizaje automático *mlr* (Schratz et al., 2021). *Ranger* es una implementación de alta

eficiencia de *random forest* (Breiman, 1999). Muchos estudios han probado que *random forest* es una de las mejores técnicas de aprendizaje automático en la actualidad (Hengl et al., 2018), y ha sido utilizado en la desagregación de data remota de humedad del suelo anteriormente (Abbaszadeh et al., 2019; Chen et al., 2019; Zhao et al., 2018) incluida data del SMAP (Hu et al., 2020; Rao et al., 2021; Zappa et al., 2019).

Khaledian y Miller (2020) describen en la tabla 13 y 14 algunos estudios relevantes y el número de parámetros que deben estimarse al entrenar un modelo de *random forest* respectivamente.

Tabla 13. Modelo de regresión aplicado como método de desagregación espacio-temporal de estimaciones remotas de la humedad del suelo en el presente proyecto de investigación

Nombre del modelo	Símbolo	Tipo de modelo	Estudios de desagregación
Random forest	Rf	Aprendizaje automático	(Bai et al., 2019a; S. Chen et al., 2019a; Im et al., 2016; Qu et al., 2021; Zhao et al., 2018a).

Elaboración por Marcelo Bueno Dueñas.

Tabla 14. *Parámetros del modelo random forest*

Nombre del modelo	Símbolo	Cantidad de parámetros	Parámetros
Random forest	Rf	4	• Número de árboles. • Número de variables predictoras elegidas en cada nodo. • Tamaño mínimo de los nodos de cada árbol. • Profundidad de cada árbol.

Fuente: Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping (Selección de métodos de aprendizaje automático para mapeo digital de suelos). Applied Mathematical Modelling, 81, 401-418.

5.3.2.3. Entrenamiento y parametrización del random forest.

Se realizó el entrenamiento de los modelos *random forest* en el soporte original del SMAP-L3-E (9 km) tanto para la desagregación temporal (~ 1300 modelos) como para la desagregación espacial para cada fecha entre el 2015 y el 2022 (~ 4000 modelos).

Uno de los parámetros fundamentales del algoritmo random forest es *mtry*, el cual se define como el número de variables elegidas aleatoriamente para realizar un partición de un árbol (Probst et al., 2019). Valores bajos de *mtry* producen arboles con menor

correlación, generando mejor estabilidad, en general también valores bajos de *mtry* generan peores predicciones. En general p/3 es bastante robusto y estable, aunque en algunos casos puede ser optimizado. Resultados empíricos han demostrado que, para problemas de regresión de baja dimensionalidad, $\sqrt{p}$ es por lo general mejor que $^{p}/_{3}$. En este estudio se usó *mtr* = 7. Además el tiempo de computación disminuye linealmente a medida que *mtry* baja (Wright & Ziegler, 2017).

El número de árboles del random forest debe ser suficientemente grande para evitar sesgo y *subajuste*. Para estimadores del error basados en perdidas cuadráticas medias como el error cuadrado medio (RMSE), a mayor número de árboles menor error de generalización (Probst et al., 2019). En este estudio se utilizaron 100 árboles para todos los modelos por motivos computacionales.

En general se utilizaron los parámetros recomendados por Probst et al. (2019) y los valores por defecto en *ranger* (Wright & Ziegler, 2017).

5.3.2.4. Modelos de desagregación temporal.

Para la reconstrucción de las series de tiempo de humedad del suelo del SMAP-L3-E a resolución diaria, se entrenaron aproximadamente 1200 *random forest*, uno para cada pixel dentro del area de estudio usando la serie de tiempo incompleta de humedad del suelo del SMAP-L3-E de cada pixel como data de entrenamiento, adicionalmente como covariables se usaron la precipitación y la humedad del suelo SMAP-L3-E antecedentes al día de predicción. Los parámetros usados fueron los mismos descritos anteriormente.

Como estrategia de validacion se usó la validacion temporal con origen móvil .

Mediante la aplicación de los modelos de desagregación temporal para cada pixel fue posible reconstruir todas las series de tiempo del producto SMAP-L3-E de forma diaria.

5.3.2.5. Modelos de desagregación espacial.

En este estudio se usó el algoritmo *random forest* (Q. Chen et al., 2020) como modelo de desagregación espacial de estimaciones remotas del contenido de humedad del suelo del producto SMAP-L3-E θ_{SMAP} . En total se utilizaron aproximadamente cuatro mil *rasters* de θ_{SMAP} para el entrenamiento de los modelos, la mitad de los cuales fueron obtenidos del SMAP y la mitad reconstruidos mediante el modelo de desagregación temporal.

5.3.2.6. Evaluación visual de la desagregación espacial.

Una desagregación exitosa debería ser capaz de reproducir la estructura espacial de la variable original, en este caso θ_{SMAP}. La evaluación generalmente se basa en una comparación visual entre la $\theta_{DWS}\ y\ la\ \theta_{SMAP}$ (Wakigari & Leconte, 2022; Zappa et al., 2019). Por lo tanto, para evaluar efectivamente la capacidad de desagregación de los modelos se comparó visualmente la correspondencia espacial entre el contenido volumétrico de agua en el suelo desagregada (θ_{DWS}) respecto a al contenido volumétrico de agua en el suelo del producto SMAPL3E (θ_{SMAP}) para dos fechas que representan condiciones típicas de humedad del suelo en la zona de estudio.

5.3.2.7. Evaluación estadística de modelos de desagregación espacio-temporal.

Antes de aplicar los modelos para la predicción de la humedad del suelo a altas resoluciones, se evaluó la performance de los modelos en el soporte espacial de los píxeles del θ_{SMAP} (~ 9km).

Ya que el area de estudio se considera pequeña en comparación con otros estudios (Bai et al., 2019; Rao et al., 2021). La evaluación del error de generalización de los modelos se realizó mediante validación cruzada de 10 *folds repetida con búsqueda grillada implementada según* Krstajic et al. (2014, p. 3) en *mlr* (Schratz et al., 2021).

Los estadísticos más comunes para evaluar el desempeño de un modelo de regresión se resumen en la tabla 15.

De esta forma fue posible evaluar el error medio de generalización de los modelos sin la necesidad de dividir la data en datos de entrenamiento y de evaluación.

Tabla 15. *Medidas comunes de evaluación del desempeño de un modelo de regresión*

Símbolo	Nombre	Ecuación	Explicación
MAE.	Error absoluto medio.	$\frac{1}{n}\sum_{j=1}^{n}\lvert y_j - \hat{y}_j\rvert$	MAE se calcula promediando los valores absolutos de los residuos.
RMSE.	Raíz del error cuadrático medio	$\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$	RMSE se calcula mediante la raíz cuadrada de la suma de cuadrados de los residuos.

Continúa tabla 15.

R^2	Coeficiente de determinación.	$\frac{\sum_{j=1}^{n}(x_j y_j - \overline{xy})}{\left(\sum_{j=1}^{n} x^2{}_j - \bar{x}^2\right)\left(\sum_{j=1}^{n} y^2{}_j - \bar{y}^2\right)}$ $= 1 - \frac{SSRE}{SST}$	El R^2 indica la proporción de la varianza de la variable de predicción que el modelo es capaz de explicar. R^2 indica qué tanto el modelo mejora la predicción de la variable en comparación a usar la media de los valores observados.

En MAE y RMSE $n,\ y_j, \widehat{y}_j$ son el tamaño muestral, los valores observados y los valores predichos respectivamente. R^2, x_j, y_j son los valores observados y predichos, $\bar{x}, \bar{y}$ son las medias respectivas; ρ es el coeficiente de correlación de Pearson, σ_x, σ_y son las respectivas varianzas de los valores observados y predichos y μ_x, μ_y son las medias de los valores observados y predichos respectivamente.

Para evaluar la capacidad predictiva de los modelos de desagregación de forma cuantitativa se usó el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2) (Colliander et al., 2017; Entekhabi et al., 2010).

Estos estadísticos se calcularon sobre los residuales de los modelos entre los valores observados con los valores predichos por los modelos para cada *fold* de validacion (CV). Mediante la validacion cruzada se estimó el error de predicción de los modelos (RMSE y R^2).

Ya que el objetivo del estudio es cambiar el soporte espacial del producto SMAP-L3-E de 9 km a ~100 m no es posible generalizar las estimaciones de error de los modelos de desagregación a las predicciones de alta resolucion (~ 100 m)

Para lo cual es necesario realizar validacion externa, es decir comparar las predicciones con data independiente agregada a la misma resolución espacial que el producto desagregado (Crow et al., 2012).

5.3.2.8. Interpretación de los modelos de desagregación.

Generalmente los modelos de aprendizaje automático como *random forest* no suelen ser fáciles de interpretar, por el gran número de parámetros y su capacidad de capturar relaciones no lineales entre las covariables y la variable de respuesta.

Con el fin de obtener *insights* y dilucidar la complejidad del *random forest* como modelo de desagregación se aplicaron dos métodos para entender su funcionamiento y lógica interna: el algoritmo *LIME* y la interpretación de árboles de regresión extraídos de dos modelos *random forest* para dos fechas representativas del periodo de monitoreo.

5.3.2.9. Explicaciones interpretables locales – LIME.

Con el fin de entender mejor los resultados de los modelos y comprender qué covariables afectan las predicciones en esta investigación se implementó el método *LIME*.

El algoritmo *LIME* fue ejecutado para dos modelos de desagregación (que representan la época seca y época húmeda en el periodo de monitoreo, 16 de agosto del 2021 y 9 de febrero del 2022 respectivamente). Se seleccionaron aleatoriamente 6 píxeles distribuidos

en el área de estudio, estos se eligieron aleatoriamente con el objetivo de que se distribuyan en la mayor parte de rango de ocurrencia de θ_{SMAP} (0.05 a 0.60 cm^3 cm^{-3}).

El modelo de aproximación local elegido fue LASSO (Hvitfeldt et al., 2022; van Heumen, 2019) . Se usó *LIME* para aproximar localmente los modelos de *downscaling* para cada fecha. El número de covariables analizadas fue 20 (ya que en este estudio se usaron un número relativamente pequeño de covariables, se optó por mantener la dimensionalidad original de las covariables, aproximadamente ~ 20 covariables).

Los gráficos generados muestran la importancia de cada covariable en la predicción de cada pixel seleccionado, particularmente cómo la covariable modula su importancia en el modelo para cada pixel mediante su variación a través de los casos predichos.

Como análisis adicional y para aislar la influencia de la precipitación también se implementó *LIME* sin las covariables de PISCO (medias mensuales de precipitación diaria).

5.3.2.10. Aproximación mediante árboles de regresión.

Durante la construcción del modelo, *random forest* entrena *n* arboles de regresión simultáneamente. Cada árbol de regresión tiene una estructura diferente dada por los parámetros, *mtry* y *maxnode*. Ya que se entrenan simultáneamente cientos de árboles y finalmente se promedian las predicciones de cada uno, *random forest* pierde interpretabilidad (Khaledian & Miller, 2020).

Sin embargo, cada árbol entrenado puede proporcionar valiosa información sobre las covariables y su relación con la variable de respuesta.

Se entrenaron dos árboles de regresión como aproximación a los *random forest* para dos fechas que representan la época seca y época húmeda en el periodo de monitoreo, 16 de agosto del 2021 y 9 de febrero del 2022 respectivamente.

De la misma manera que con *LIME*, para aislar la influencia de la precipitación se entrenaron dos árboles de regresión para las mismas fechas sin PISCO.

5.3.2.11. Generación de mapas de humedad del suelo a alta resolucion.

La evaluación visual de los modelos de desagregación a ambas escalas espaciales (9 km y 100 m) y la evaluación estadística sugirieron la capacidad de los modelos de desagregar el producto SMAP-L3-E con adecuada precisión; de tal manera que, asegurándonos de su capacidad de capturar las relaciones no lineales entre las covariables y la humead del suelo estos fueron aplicados en la desagregación de θ_{SMAP} (~ 9km) para predecir el contenido de agua del suelo θ_{DWS} a altas resoluciones espaciales $(\sim 100m)$.

Predecir diariamente la humedad del suelo a 100 metros de resolucion espacial en el área completa de estudio fue inalcanzable con los recursos computacionales que se disponían, por lo tanto, las predicciones se realizaron solo para la superficie abarcada por la micro- cuenca K´ayra, esto se justifica porque en ella se encuentra la estación de monitoreo.

Las predicciones se realizaron bajo el supuesto de que los modelos *random forest* construidos a bajas resoluciones espaciales también son válidos para predecir la humedad del suelo a más altas resoluciones espaciales usando las covariables a la resolucion deseada. En otras palabras, se asume que los modelos de desagregación son invariantes respecto a la escala espacial.

5.3.3. Determinación de la influencia de la topografía, las propiedades del suelo y la precipitación en la dinámica espacial del producto SMAP-L3-E desagregado mediante random forest en el área de estudio.

5.3.3.1. Análisis espacial del producto SMAP-L3-E desagregado mediante random forest.

La data desagregada de humedad del suelo reveló patrones a diferentes escalas que emergen por las interacciones entre la hidrología, topografía y propiedades del suelo a través del paisaje (Vergopolan et al., 2022), y por lo tanto nos permitió estudiar la variabilidad espacial de la humedad del suelo.

Para cuantificar la variabilidad de la humedad del suelo a escalas locales o de campo se muestrearon 80 polígonos de aproximadamente 1 Km^2 de superficie, mediante muestreo grillado para dos fechas hidrológicas representativas (época seca y época húmeda). Se procedió a calcular la media, desviación estándar y coeficiente de variación espaciales de la humedad del suelo desagregada a 100 m (μDWS, σDWS y C.V. DWS) para cada polígono

En total se obtuvieron 80 observaciones de cada variable a analizar. Dos polígonos ocurrieron en zonas impermeables o cuerpos de agua y fueron excluidos del análisis.

También se calcularon tanto la media como la desviación estándar de las covariables dentro de cada polígono seleccionado, conformando en total 40 variables para el análisis de componentes principales (PCA).

5.3.3.2. Evaluación de los factores relacionados con la distribución espacial del producto SMAP-L3-E desagregado mediante random forest.

En el presente estudio se llevó a cabo un análisis de componentes principales (PCA) para identificar y caracterizar la relación entre la variabilidad espacial de la humedad del suelo y las características físicas del paisaje (*drivers* de la variabilidad de la humedad del suelo caracterizado por las covariales). El análisis de componentes principales PCA permitió identificar los modos dominantes de variación en la data. Y cuantificar cómo co-varian diferentes variables y su influencia en la media y variabilidad del producto SMAP-L3-E desagregado. En específico, el PCA se utilizó para comparar la media y la desviación estándar de la humedad del suelo desagregada $\sigma_{\mathcal{D}}$ con la media y desviación estándar de covariables a alta resolucion que modulan la humedad del suelo en el paisaje, como las propiedades del suelo, topográficas e hidrológicas.

Antes de aplicar el PCA las covariables fueron estandarizadas, con el fin de reducir la influencia de ciertas variables por diferencias en la escala de medición (por ejemplo, la magnitud de la elevación es cientos de veces más grande que la de conductividad hidráulica satura del suelo).

El análisis se realizó en la librería *FactoMineR* (Husson et al., 2008) mediante el algoritmo de descomposición de valores singulares, SVD (Husson et al., 2017) sobre la matriz de correlaciones de las medias y desviaciones estándar de las covariables.

Los resultados se interpretaron mediante un *biplot*, el cual indica cómo las propiedades espaciales de las covariables covarían con la humedad del suelo desagregada respecto al primer y segundo componente principal. Mas específicamente, cada punto representó la

media y desviación estándar espacial de la humedad del suelo desagregada, en polígonos de 1 km^2 de superficie, y su posición en el grafico representa su asociación con los dos principales modos de variación de la data (componentes principales). Cada flecha representa la carga de una covariable, con su longitud y dirección indicando qué tanto una covariable influencia un componente principal. Los ángulos entre las flechas indican correlación entre las covariables, adicionalmente el PCA será calculado para la media $\mu_{\mathcal{D}}$ y el coeficiente de variación espacial $CV_{\mathcal{D}}$ de la humedad del suelo desagregada.

5.3.4. Análisis de la relación entre el producto SMAP-L3-E desagregado mediante *random forest* con la humedad del suelo medida *in situ* en el área bajo estudio.

5.3.4.1. Monitoreo de la humedad del suelo.

Considerando los objetivos del presente proyecto de investigación, los resultados de investigaciones previas y la dificultad y costos de implementación de estrategias de monitoreo espacial aleatorio, se propuso una sola estación de monitoreo diario de la humedad del suelo para un pixel del producto SMAP-L3-E, el esquema propuesto se resume en la tabla 16.

Tabla 16. *Esquema de monitoreo de la humedad del suelo propuesto con fines de validación del producto SMAP-L3-E* ⋆

Universo objetivo	Horizonte superficial del suelo del sitio de validación (un píxel del producto SMAP-L3-E de humedad del suelo), desde mayo del 2021 hasta julio del 2022.

Continuación tabla 16.

Variable objetivo	Contenido volumétrico de humedad del suelo (cm^3cm^{-3}) a cinco centímetros de profundidad.
Frecuencia de medición	Diaria
Soporte físico de medición.	Volumen de medición del sensor de capacitancia *ThetaProb*e ML3 (60·30 mm de diámetro).
Método medición.	Medición electromagnética de la humedad del suelo mediante el sensor de capacitancia *ThetaProbe* ML3 (Cooper, 2016; Gaskin & Miller, 1996).
Hora de medición.	Aproximadamente a las 7:00 de la mañana.
Densidad de puntos de medición.	Un punto de muestreo por píxel del producto SMAP-L3-E.
Protocolos de recolección de data.	Geolocalización de la ubicación de muestreo con GPS *Garmin GPSMAP 66sr* (precisión de 1.5-2 m), observación de las condiciones de precipitación y registro en libreta de notas impresa previamente.

⋆Terminología basada en de Gruijter et al. (2006)
Fuente: Elaboración propia

Se monitoreó la humedad del suelo de forma diaria desde mayo del 2021 hasta julio del 2022 mediante el siguiente protocolo:

Para cada fecha de medición se ubicó el punto de monitoreo dentro de un pixel del producto SMAP-L3-E desagregado con un GPS multibanda modelo GPSMAP marca *Garmin*, con precisión de aproximadamente 1.5 metros; una vez ubicadas las coordenadas del punto de muestreo se procedió a tomar tres a cuatro mediciones con el sensor *ThetaProbe* ML3 (figura 9) distribuidas a una distancia aproximada de 1.5 metros entre ellos con la ayuda de una cinta métrica, posteriormente las mediciones se promediarán (esto es permisible debido a la exactitud del GPS y tiene el objetivo de reducir la variabilidad de la humedad del suelo (Cooper, 2016)) la medición se hizo a una profundidad de 5 cm desde la superficie del suelo (Babaeian et al., 2019, p. 20); la única condición para que la medición sea considerada valida es que el volumen de suelo medido sea homogéneo y no tenga predominancia de restos orgánicos, rocas o grietas muy grandes (NASA, 2014).

Figura 9.*Sensor de capacitancia de la humedad del suelo ThetaProbe ML3*

Nota: Adaptado de la página web del fabricante, 2021 (https://delta-t.co.uk/product/ml3/)

Adicionalmente se realizó la calibración del sensor de capacitancia *ThetaProbe* ML3 en el sitio de validación mediante la metodología de (Topp & Ferré, 2018), el cual consistió en estimar los parámetros de la ecuación de Gaskin y Miller (1996) mediante la comparación entre las mediciones del sensor *ThetaProbe* ML3 y mediciones gravimétricas.

Las mediciones diarias fueron utilizadas para la comparación estadística con las predicciones de humedad del suelo de los modelos de desagregación.

5.3.4.2. Validación del producto desagregado del SMAP-3L-E.

En el presente estudio se usó el coeficiente de correlación de Pearson calculado entre las series de tiempo del contenido de agua del suelo observado *in situ* y el producido por la desagregación del SMAP-L3-E en la zona de monitoreo.

Estudios previos de validación (Bai et al., 2019; Colliander et al., 2017; Liu et al., 2020; Singh et al., 2019a; Y. Xu, 2019) utilizaron el coeficiente de correlación σ para la comparación analítica entre las mediciones in situ y las estimaciones remotas de humedad del suelo:

$$\sigma = \frac{\frac{1}{M}\sum_{k=1}^{M}\left(\left({\theta^D}_j - \frac{1}{M}\sum_{k=1}^{M}{\theta^D}_j\right)\left({\theta^K}_j - \frac{1}{M}\sum_{j=1}^{K}{\theta^K}_j\right)\right)}{\sigma_D \sigma_K}$$

Ecuación 17

Donde ${\theta^D}_j$ representa la estimación remota de la humedad del suelo para un sitio de validación (píxel) determinado el día de muestreo j, ${\theta^K}_j$ representa la humedad del suelo medida en campo para el mismo sitio de validación en el día muestreo j, σ_D representa la desviación estándar de todas las estimaciones remotas de humedad del suelo y σ_K

representa la desviación estándar de todas las mediciones en campo de humedad del suelo, la ventana de tiempo *M* representa la cantidad total de datos que se corresponden con los eventos de muestreo o el periodo de monitoreo de la humedad del suelo en campo.

5.3.4.3. Coeficiente de correlación cuantílico multiescala (MQCC).

Se optó por el análisis de coeficiente de correlación cuantílico multiescala o MQCC (Xu et al., 2020). Un enfoque similar fue usado por Singh et al. (2019), Beck et al. (2021) usaron el coeficiente de regresión de la mediana (cuantil 0.5) en diferentes escalas temporales para evaluar diferentes productos satelitales de humedad del suelo.

Antes de analizar la correlación entre dos series de tiempo, se realizó un análisis multiescalar, que consistió en la transformación de las series de tiempo observadas y predichas mediante el método de *coarse graning,* se agregó las series de tiempo a diferentes agregaciones o granularidades temporales, desde la diaria (original) hasta la escala mensual (el promedio de 20 días).El análisis multiescalar nos permitió analizar la relación entre las dos variables a diferentes escalas, y por lo tanto fue posible encontrar patrones que no sería posible encontrar con las series de tiempo originales.

Finalmente se implementó el análisis cuantílico en cada agregación temporal mediante la librería *quantreg* (Koenker et al., 2017).

5.3.4.4. Gráficos de dispersión y Gráfico Q-Q.

Se construyeron gráficos de dispersión y *qq-plots* entre las series de tiempo de humedad del suelo observada y predicha por los modelos de desagregación.

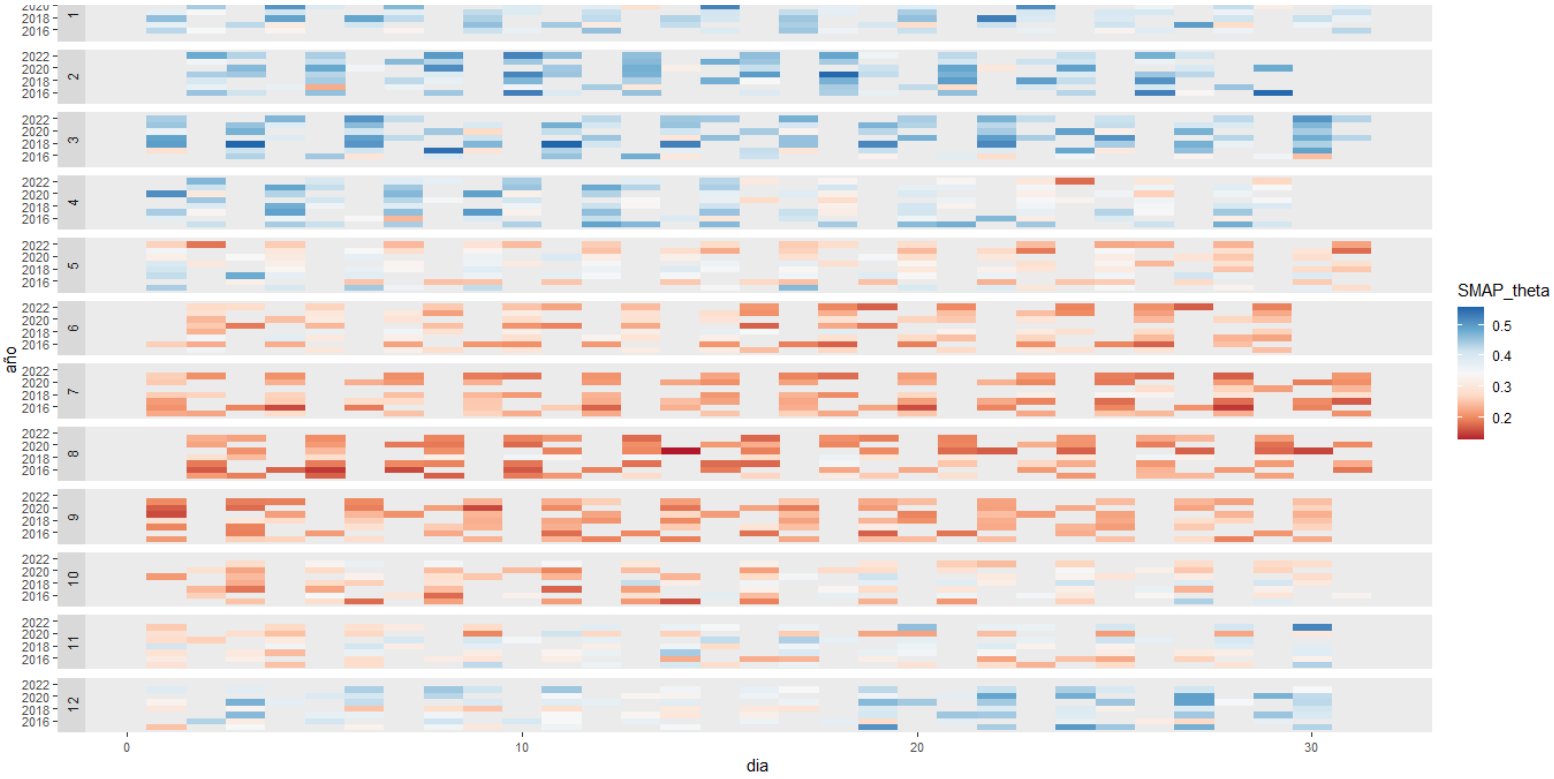
VI. RESULTADOS.

6.1. Evaluación de la capacidad de desagregación espacio-temporal mediante *random forest* del producto SMAP-L3-E en el área de estudio.

6.1.1. Producto SMAP-L3-E.

La figura 10 es un *levelplot* de series de tiempo, el *levelplot* muestra la información temporal del producto SMAP-L3-E para el pixel de monitoreo para los años 2015 a 2022. De forma general el producto es capaz demostrar la estacionalidad de la humedad en la región de estudio dada principalmente por el régimen de precipitaciones estacionales. Los meses de enero, marzo y abril muestran un contenido de agua del suelo de entre 0.4 a 0.5 cm3 cm-3, a finales de abril la humedad del suelo tiende a disminuirá a niveles cercanos a 0.3 cm3 cm-3, incluso con valores cercanos a 0.2 cm3 cm-3 en una ocasión en abril del 2022. Los meses de mayo, junio, agosto, septiembre y mediados de noviembre muestran contenidos de humedad del suelo entre 0.3 a 0.1 cm3 cm-3, a fines de noviembre la humedad del suelo empieza a aumentar de nuevo llegando a valores de 0.4 cm3 cm-3.

Figura 10*.Variación temporal del producto SMAP-L3-E para el pixel de monitoreo (2015-2022)*

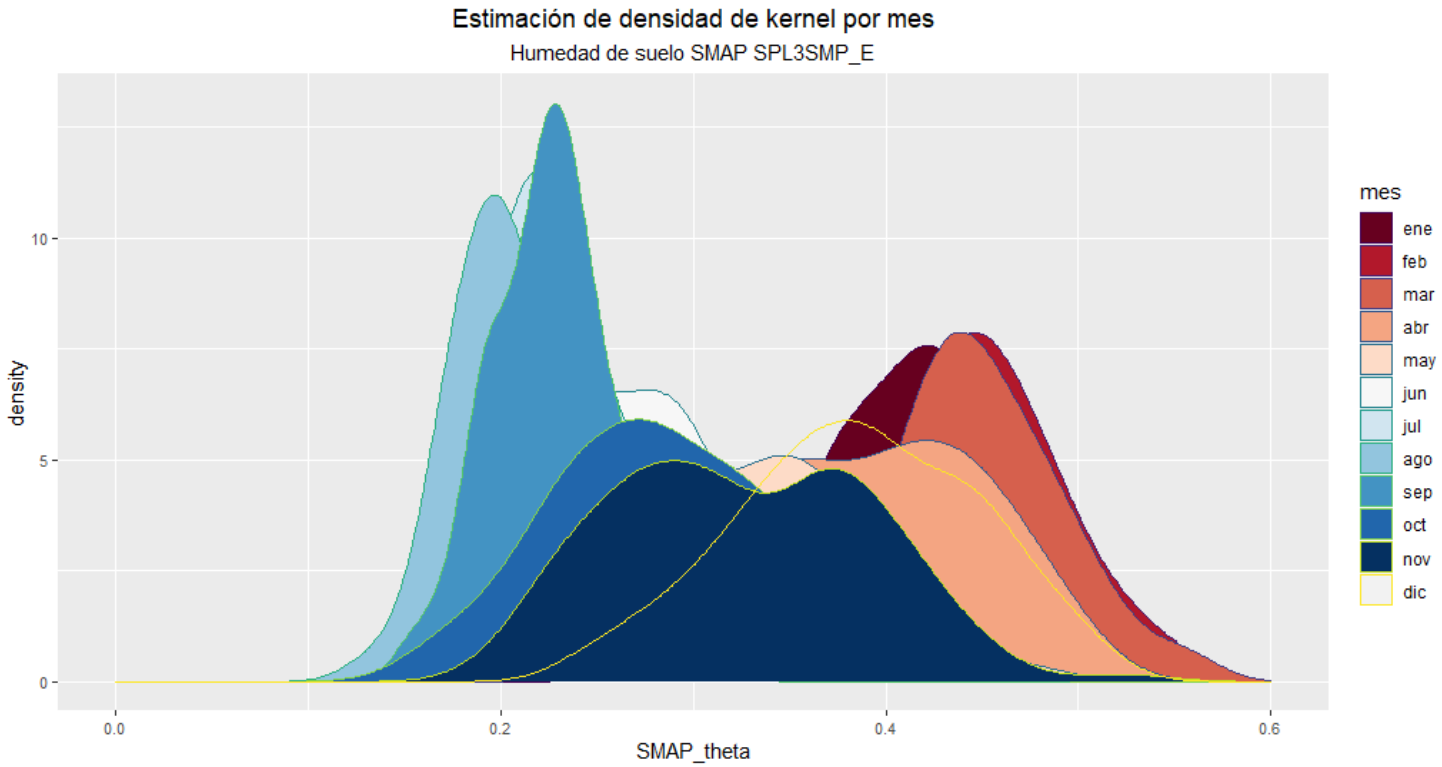


Elaborado por Marcelo Bueno Dueñas. SMAP theta es el contenido volumétrico de humedad del suelo del producto SMAP-L3-E en cm^3 cm^{-3}.

De igual manera la figura 11 muestra la distribución de la humedad del suelo del producto SMAP-L3-E agrupadas por mes del año entre el 2015 al 2022, esta grafica refuerza la idea de que el producto capta de forma adecuada la dinámica estacional de la humedad del suelo. En específico la imagen muestra estimaciones de densidad de *kernel* de la humedad del suelo, los meses de marzo y febrero muestran máxima densidad, aproximadamente a 0.5 cm^3 cm^{-3} de contenido de agua en el suelo, desde noviembre tiene una distribución más compleja, aproximadamente bimodal, con dos picos de máxima densidad cerca de 0.3 y 0.4 cm^3 cm^{-3} respectivamente. Finalmente, los meses de agosto,

septiembre y octubre muestran máximas densidades en 0.18, 0.22 y 0.28 cm^3 cm^{-3} respectivamente, cada uno progresivamente más húmedo que el anterior.

Figura 11.*Distribución de la humedad del suelo SMAP-L3-E por mes*



.

Elaborado por Marcelo Bueno Dueñas.

Adicionalmente la tabla 17 muestra la distribución estacional de la humedad del suelo estimada por el producto SMAP-L-E agrupada por estaciones hidrológicas, con el fin de utilizar la fecha de observación de cada imagen del producto de humedad del suelo del SMAP y para capturar su variabilidad estacional, se agruparon las observaciones según la estación definidas aquí de la siguiente manera: invierno Q1 (junio, julio y agosto),

primavera Q2 (septiembre, otubre y noviembre), verano Q3 (diciembre, enero y febrero) y otoño Q4 (marzo, abril y mayo)

Tabla 17.Distribución de la humedad suelo del producto SMAP-L3-E por estación hidrológica.

Q	Min	Max	Media	Mediana	Percentil25	Percentil75
Q1	0,066743	0,390584	0,157237	0,147677	0,121946	0,18112
Q2	0,055876	0,558597	0,212878	0,194508	0,138685	0,277061
Q3	0,091711	0,627069	0,400627	0,40772	0,354248	0,453941
Q4	0,103815	0,613292	0,353683	0,363628	0,277621	0,433757

Elaborado por Marcelo Bueno Dueñas.

Se puede observar que el agrupamiento de la humedad del suelo según la estación del año definida previamente explica la variabilidad estacional al igual que respalda la división del ciclo hidrológico en temporada seca (Q1 y Q2) y temporada lluviosa (Q3 y Q4) respectivamente en los andes tropicales.

6.1.2. Análisis exploratorio de las covariables.

En la tabla 18 se pueden apreciar los estadísticos principales de la distribución de las covariables armonizadas a la resolución del producto SMAP-L3-E (~ 9 km) usadas en esta tesis.

Tabla 18. Estadísticos descriptivos principales de las covariables usadas en el estudio

Covariables	Media.	Desviación estándar.	Mediana	Skewness	Kurtosis.	Error estándar.	Q0.25	Q0.75
Contenido de arcilla (g/Kg)	240	24.92	237	0.5109	-0.113	0.697	222	256
Contenido de arena (g/Kg)	446	29.63	451	-0.304	-0.743	0.828	423	470
Capacidad de intercambio catiónico (mmol/Kg)	230	36.13	233	0.055	-0.674	1.010	201	258
Densidad aparente (cg/cm3)	117	6.83	118	-0.26	-0.825	0.191	112	123
Contenido de limo (g/Kg)	312	26.19	312	-0.021	-0.283	0.732	295	330

Continúa tabla 18.

Contenido de carbono orgánico (dg/kg)	485	179.6	406	0.813	-0.699	5.024	343	629.5
Densidad de carbono orgánico (dg/kg)	368	34.96	359	1.169	1.1028	0.978	343	383.8
Stocks de carbono orgánico (dg/kg ha)	61.9	7.719	62	0.2932	-0.634	0.216	55.3	67.8
alpha_vG	0.452	0.0567	0.449	0.4593	0.1178	0.001	0.41	0.486
Conductividad hidráulica saturada. (log_{10} cm $día^{-1}$)	2.33	0.2605	2.278	1.9842	5.4213	0.007	2.16	2.424
n_vG_par	0.184	0.0123	0.183	0.1349	-0.553	0.00027	0.18	0.19
thetar_v (cm^3 cm^{-3})	0.0938	0.00874	0.092	1.1287	2.1843	0.00024	0.09	0.098
thetas_v (cm^3 cm^{-3})	0.559	0.0319	0.563	-0.612	0.3055	0.00089	0.54	0.58
DEM (m)	4166	469.23	4166	-0.238	-0.361	13.13	3846	4530
FD8	7.76	2.180	7.09	2.3631	8.1130	0.061	6.34	8.48
PISCO-DEF (mm $día^{-1}$)	4.86	1.160	4.96	-0.1603	0.7230	0.0325	4.14	5.61

Continúa tabla 18

PISCO-MAM (mm día^{-1})	1.94	0.442	2.01	-0.4798	0.1492	0.0124	1.66	2.22
PISCO-JJA (mm día^{-1})	0.206	0.111	0.17	1.9738	6.9759	0.0031	0.13	0.26
PISCO-SON (mm día^{-1})	1.66	0.3408	1.691	-0.1741	-0.140	0.0095	1.43	1.89

Elaborado por Marcelo Bueno Dueñas

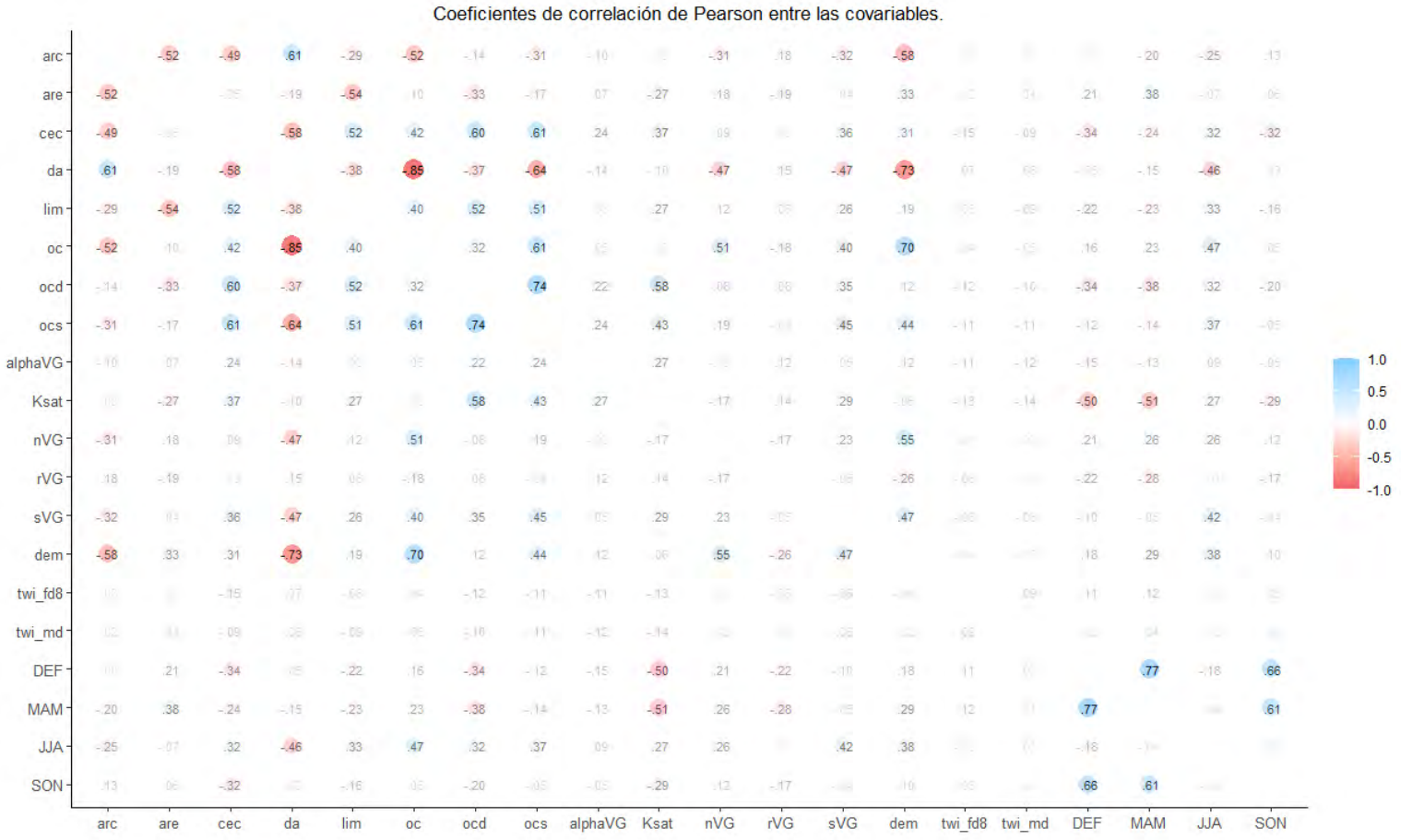
Un valor negativo de *skewness* indica que la media es menor a la mediana y la distribución es asimétrica a la izquierda, un valor positivo de *skewness* indica que la media es mayor a la mediana y que la data sigue una distribución asimétrica a la derecha, un valor de *skewness* próximo a cero indica que la data sigue aproximadamente una distribución normal.

Algunas variables mostraron aparente normalidad, por ejemplo, la capacidad de intercambio catiónico, el contenido de carbono orgánico y el DEM, mientras que otras mostraron distribución asimétrica a la derecha como, por ejemplo, la densidad de carbono orgánico, la conductividad hidráulica saturada y los índices de humedad topográficas como FD8 y MFD-md.

La figura 12 y la tabla 19 muestran los coeficientes de correlación de Pearson entre las covariables y los valores p para la correlación respectivamente. El contenido de arcilla muestra coeficientes de correlación negativos significativos con el contenido de arena, con

la capacidad de intercambio catiónico, el contenido de carbono orgánico, con la elevación de la superficie DEM y con la media de la precipitación en la temporada JJA con -0.49, -0.52, -0.31, -0.58 y -0.25 respectivamente. Así mismo muestra coeficientes de correlación positivos significativos con la densidad aparente 0.61 (tabla 19).

Figura 12*.Matriz de coeficientes de correlación entre las covariables.*



Elaboración propia. Cada correlación es representada con un círculo en el diagrama, el tamaño cada círculo representa el valor absoluto de la correlación, el color de cada círculo representa el signo de la correlación. El método usado fue el de Pearson

Tabla 19.Significancia estadística de los coeficientes de correlación de Pearson entre las covariables mediante valores p.

	arc	Are	cec	Da	lim	Oc	ocd	Ocs	alphaVG	Ksat	nVG	rVG	sVG	dem	twi_fd8	twi_md	DEF	MAM	JJA	SON
arc		0	0	0	0	0	0	0	.001	.259	0	0	0	0	.364	.43	.906	0	0	0
are	0		.087	0	0	0	0	0	.01	0	0	0	.117	0	.476	.115	0	0	.011	.049
cec	0	.087		0	0	0	0	0	0	0	.002	.31	0	0	0	.001	0	0	0	0
da	0	0	0		0	0	0	0	0	0	0	0	0	0	.008	.026	.076	0	0	.219
lim	0	0	0	0		0	0	0	.296	0	0	.026	0	0	.03	.001	0	0	0	0
oc	0	0	0	0	0		0	0	.097	.31	0	0	0	0	.172	.057	0	0	0	.068
ocd	0	0	0	0	0	0		0	0	0	.002	.003	0	0	0	0	0	0	0	0
ocs	0	0	0	0	0	0	0		0	0	0	.122	0	0	0	0	0	0	0	.05
alphaVG	.001	.01	0	0	.296	.097	0	0		0	.293	0	.054	0	0	0	0	0	.002	.05
Ksat	.259	0	0	0	0	.31	0	0	0		0	0	0	.033	0	0	0	0	0	0
nVG	0	0	.002	0	0	0	.002	0	.293	0		0	0	0	.61	.33	0	0	0	0
rVG	0	0	.31	0	.026	0	.003	.122	0	0	0		.09	0	.03	.956	0	0	.716	0
sVG	0	.117	0	0	0	0	0	0	.054	0	0	.09		0	.044	.023	0	.097	0	.193
dem	0	0	0	0	0	0	0	0	0	.033	0	0	0		.196	.394	0	0	0	.001
twi_fd8	.364	.476	0	.008	.03	.172	0	0	0	0	.61	.03	.044	.196		.001	0	0	.333	.101
twi_md	.43	.115	.001	.026	.001	.057	0	0	0	0	.33	.956	.023	.394	.001		.411	.184	.342	.881
DEF	.906	0	0	.076	0	0	0	0	0	0	0	0	0	0	0	.411		0	0	0
MAM	0	0	0	0	0	0	0	0	0	0	0	0	.097	0	0	.184	0		.206	0
JJA	0	.011	0	0	0	0	0	0	.002	0	0	.716	0	0	.333	.342	0	.206		.357
SON	0	.049	0	.219	0	.068	0	.05	.05	0	0	0	.193	.001	.101	.881	0	0	.357	

Elaboración propia. Valores p menores a 0.05 son significativos al 95% de confianza, valores p menores a 0.01 % son significativos al 99% de confianza.

Las abreviaturas al igual que la figura 14 son: arc es Contenido de arcilla, are es contenido de arena, cec es capacidad de intercambio catiónico, da es densidad aparente, lim es contenido de rcilla, oc es contenido de carbono orgánico, ocd es densidad de carbono orgánico, ocs es stocks de carbono orgánico, alphaVg, nVg, rVg, sVg parámetros de la función de van Genuchten, Ksat es la conductividad hidráulica saturada, twi_fd8 y twi_md son los índices de humedad topográficos calculados mediante los algoritmos de (Quinn et al.,1995) y (Qin, C. et al., 2007) respectivamente. DEF, MAM, JJA y SON la precipitación media histórica del producto PISCO para los meses de diciembre, enero y febrero; marzo, abril y mayo; julio, junio y agosto; y septiembre, octubre y noviembre respectivamente.

La capacidad de intercambio catiónico muestra coeficientes de correlación positivos significativos respecto al contenido de limo (0.52), el contenido de carbono orgánico, la densidad de carbono orgánico y el stock de carbono orgánico, 0.42, 0.60 y 0.61 respectivamente. También muestra coeficientes de correlación positivos respecto al contenido de humedad de saturación, al DEM y la precipitación media en la temporada JJA. Así mismo el coeficiente de correlación es negativo respecto a la densidad aparente (-0.58).

La densidad aparente muestra coeficientes de correlación negativos con la capacidad de intercambio catiónico, el contenido de carbono orgánico, los stocks de carbono y el DEM, -0.58, -0.85, -0.64 y -0.73 respectivamente.

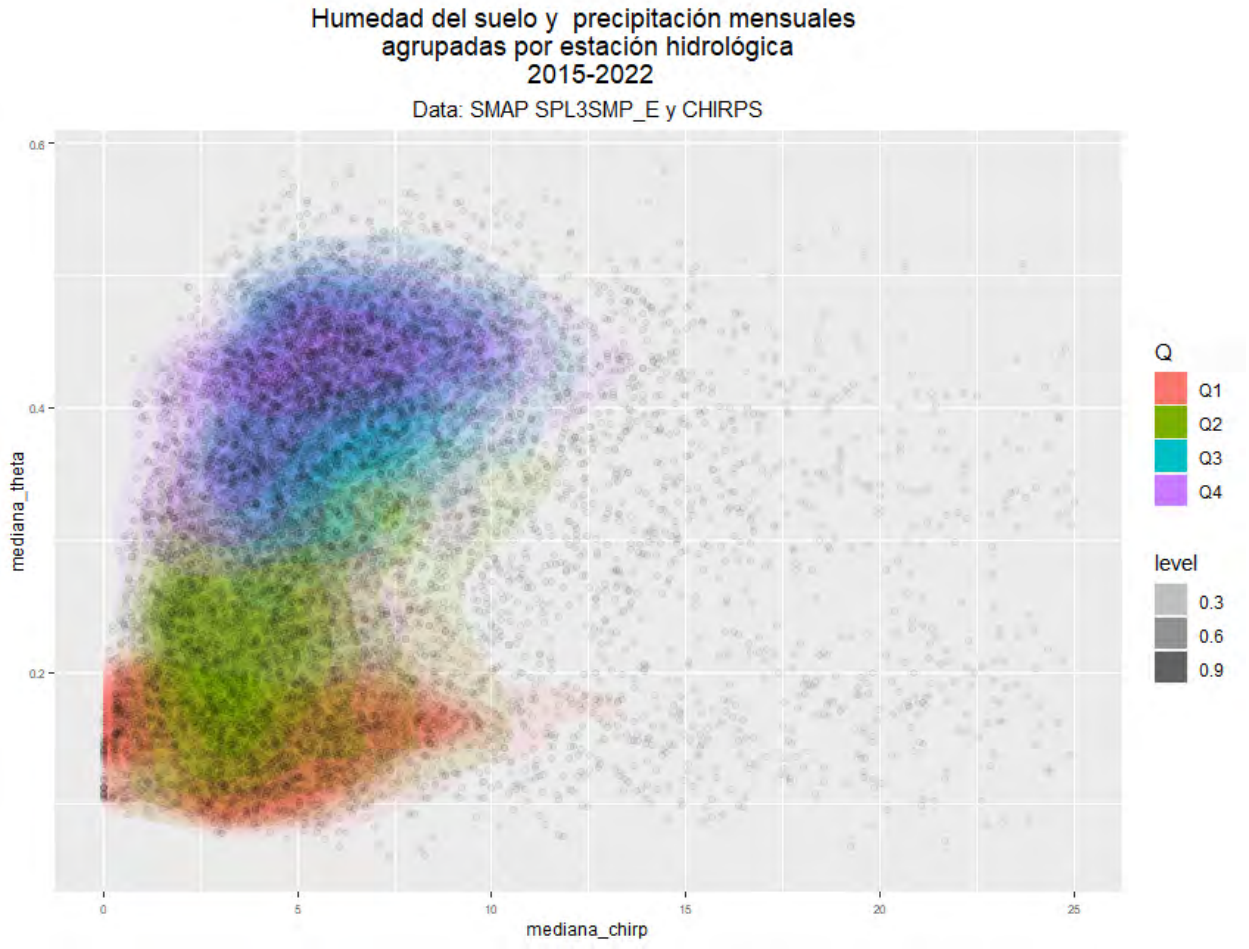
El coeficiente de correlación entre las medias históricas de precipitación DEF con MAM es 077, DEF con SON es 0.66, y MAM es SON 0.61.

Además, la conductividad hidráulica del suelo, tiene correlaciones positivas con la densidad de carbono orgánico y los stocks de carbono orgánico, con coeficientes de0.58 y 0.43 respectivamente, y correlaciones negativas con DEF y MAN, con coeficientes de -0.5 y -0.51 respectivamente.

Los índices de humedad topográficos FD8 y MD no muestran correlación entre sí ni con las demás covariables.

Para describir la relación dinámica entre el producto SMAP-L3-E y la precipitación del producto CHIRPS se compararon mediante una gráfica de dispersión agrupada por estaciones hidrológicas del año (figura 13).

Figura 13.*Diagrama de dispersión entre la humedad del suelo SMAP-L3-E y el producto CHIRPS.*



Elaboración por Marcelo Bueno Dueñas: La agrupación representa las estaciones hidrológicas, Q1 es invierno (junio, julio y agosto), Q2 primavera (septiembre, octubre y noviembre), Q3 verano (diciembre, enero y febrero) y Q4 otoño (marzo, abril y mayo).

Es fácil observar que la estructura de la relación entre la precipitación y la humedad depende del periodo del año considerado, en Q3 y Q4 la relación es casi lineal, mientras que en Q1 y Q2 la relación entre el producto SMAP-L3-E y el producto CHIRPS es menos lineal con una relación más compleja.

6.1.3. Entrenamiento y parametrización del *random forest*.

En la tabla 20 se pueden apreciar los parámetros utilizados tanto en la desagregación temporal como la desagregación espacial.

Tabla 20.*Resultado de estudios de validación de estimaciones remotas de humedad del suelo respecto a la cantidad de puntos de monitoreo.*

Parámetro.	Descripción.	Valores típicos.	Valores utilizados.
Número de variables utilizas en cada separación.	Numero de covariales utilizadas en el proceso de *spliting* en cada árbol.	p/3.	7.
Tamaño muestral	Número de observaciones utilizadas en cada árbol.	N.	1278.
Muestreo con reemplazo.	Usar muestreo con o sin reemplazo para entrenar cada árbol.	Sí.	Sí.

Continúa tabla 20

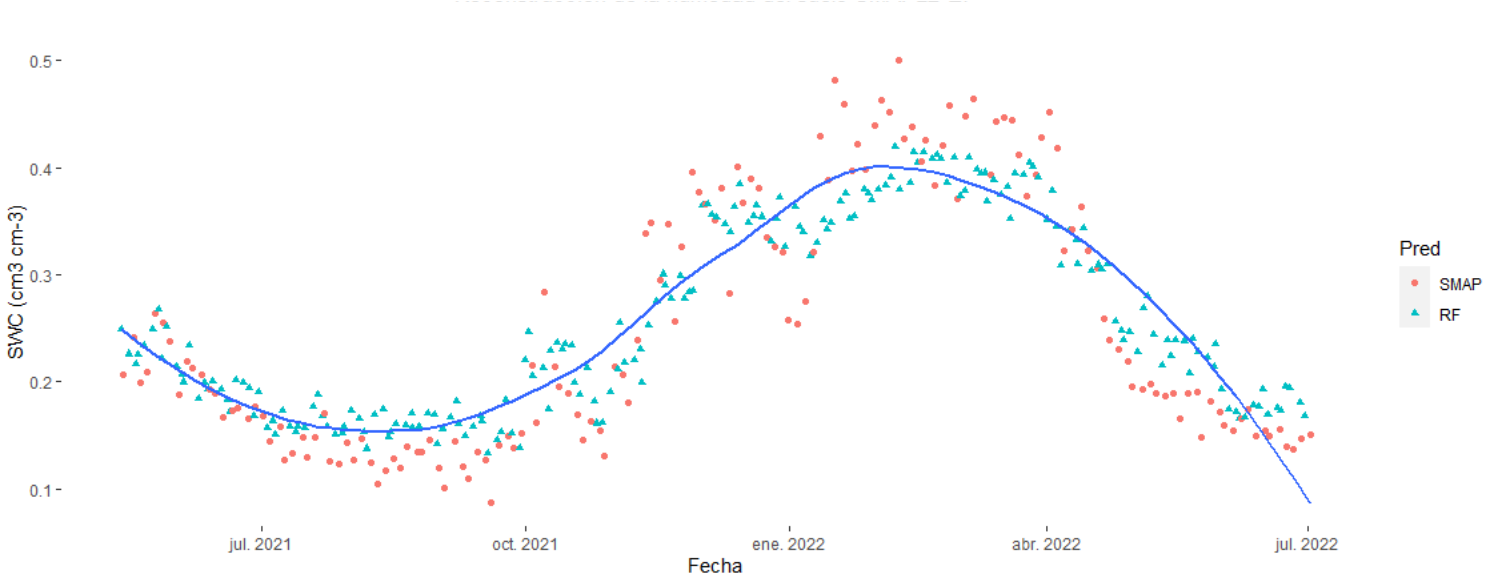
Tamaño de los nodos.	Número mínimo de observaciones en un nodo.	5.	5.
Número de árboles.	Número de árboles totales del *random forest.*	100.	100.
Criterio de separación.	Métrica que determina si un nodo se divide o no.	Varianza (para regresión). *Gini index* (para clasificación).	Varianza

Fuente: Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P.,

1.1.1. Modelos de desagregación temporal.

El proceso de reconstrucción temporal del producto SMAP-L3-E para el pixel de monitoreo puede observarse en la figura 14 (solo para el periodo de monitoreo entre mayo del 2021 hasta julio del 2022). En la figura se puede apreciar que la reconstrucción en general sigue adecuadamente la tendencia estacional del contenido de agua, este comportamiento se replica en todos los pixeles de area de estudio (la línea azul indica la línea de regresión local.)

Figura 14*.Serie de tiempo del producto SMAPL3E reconstruida mediante random forest para el pixel de monitoreo (-71.87449,-13.56040 EPSG:4326 - WGS 84)*



.

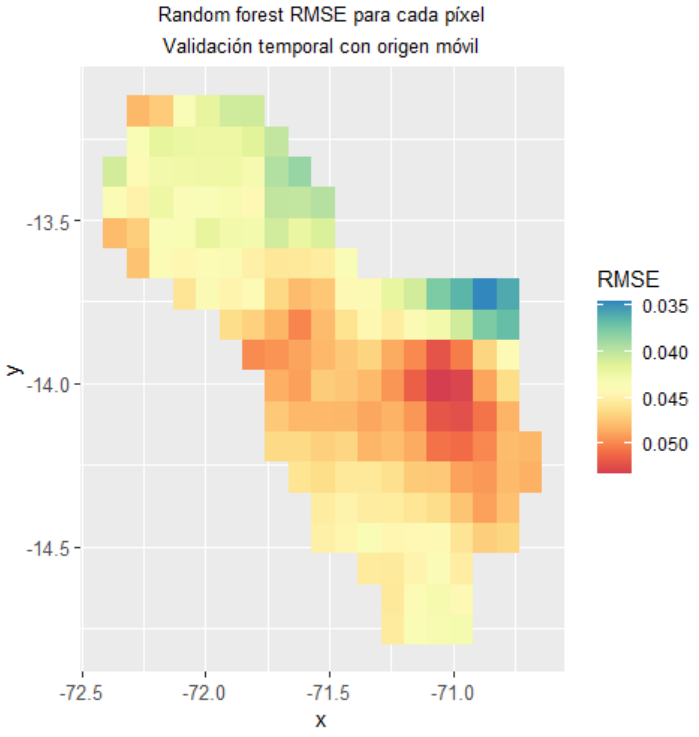
Elaborado por Marcelo Bueno Dueñas. SMAP es el contenido volumétrico de humedad del suelo en cm^3 cm^{-3} del producto SMAPL3E, con resolucion temporal de 3 a 4 días. RF es el contenido volumétrico de humedad del suelo en cm^3 cm^{-3} estimado mediante Random Forest para las fechas faltantes.

Adicionalmente en la figura 14 se puede apreciar que existe una sobreestimación de la humedad del suelo en el proceso de reconstrucción (la humedad reconstruida usualmente es mayor al contenido de agua en el suelo en días cercanos, sobre todo en épocas relativamente más secas, como entre julio a agosto) también se aprecia una aleatoriedad muy marcada.

La figura 15 muestra el RMSE (error cuadrático medio del error) de los modelos, se puede aprecia que valores más bajos de error de validacion (0.037 a 0.04 cm^3 cm^{-3}) en la región norte del area de estudio que, al este el RMSE oscila entre 0.035 a 0.04 cm^3 cm^{-3}

siendo en esta región donde el error es menor. El error aumenta de este a oeste y de norte a sur, en general el error máximo (0.05 cm^3 cm^{-3}) se encuentra próximo a las coordenadas geográficas -71.00 y -14.00 de longitud y latitud respectivamente.

Figura 15*.Distribución espacial del error de generalización de random forest por pixel*



.

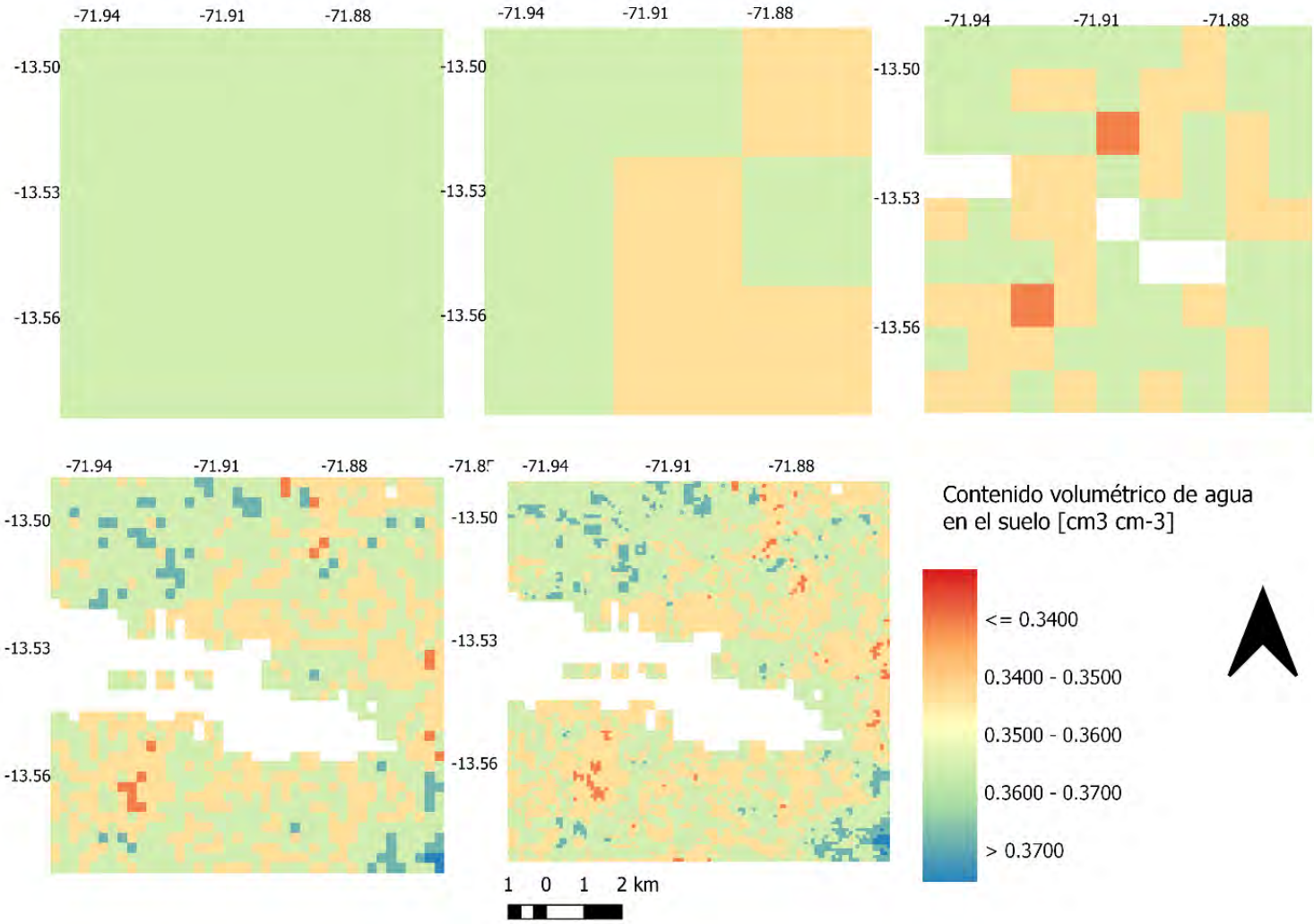
Elaborado por Marcelo Bueno Dueñas: RMSE es el error cuadrático medio expresado en las mismas unidades que la variable de estudio, cm3 cm^{-3}.

6.1.4. Modelos de desagregación espacial.

La figura 16 muestra el proceso de desagregación y la distribución espaciales de la humedad del suelo en un área alrededor del punto de monitoreo en la estación agrometeorológica K´ayra a través de diferentes soportes espaciales (3km a 1 km a 250 m a ~ 100 m) para una fecha dada, las áreas blancas indican superficies impermeables como áreas urbanas o cuerpos de agua superficiales que fueron retirados del análisis previamente.

La desagregación permite obtener información detallada de la distribución de la humedad del suelo (zonas de alta saturación y zonas de mayor homogeneidad).

Figura 16.*Proceso de desagregación espacial del producto SMAP-L3-E.*

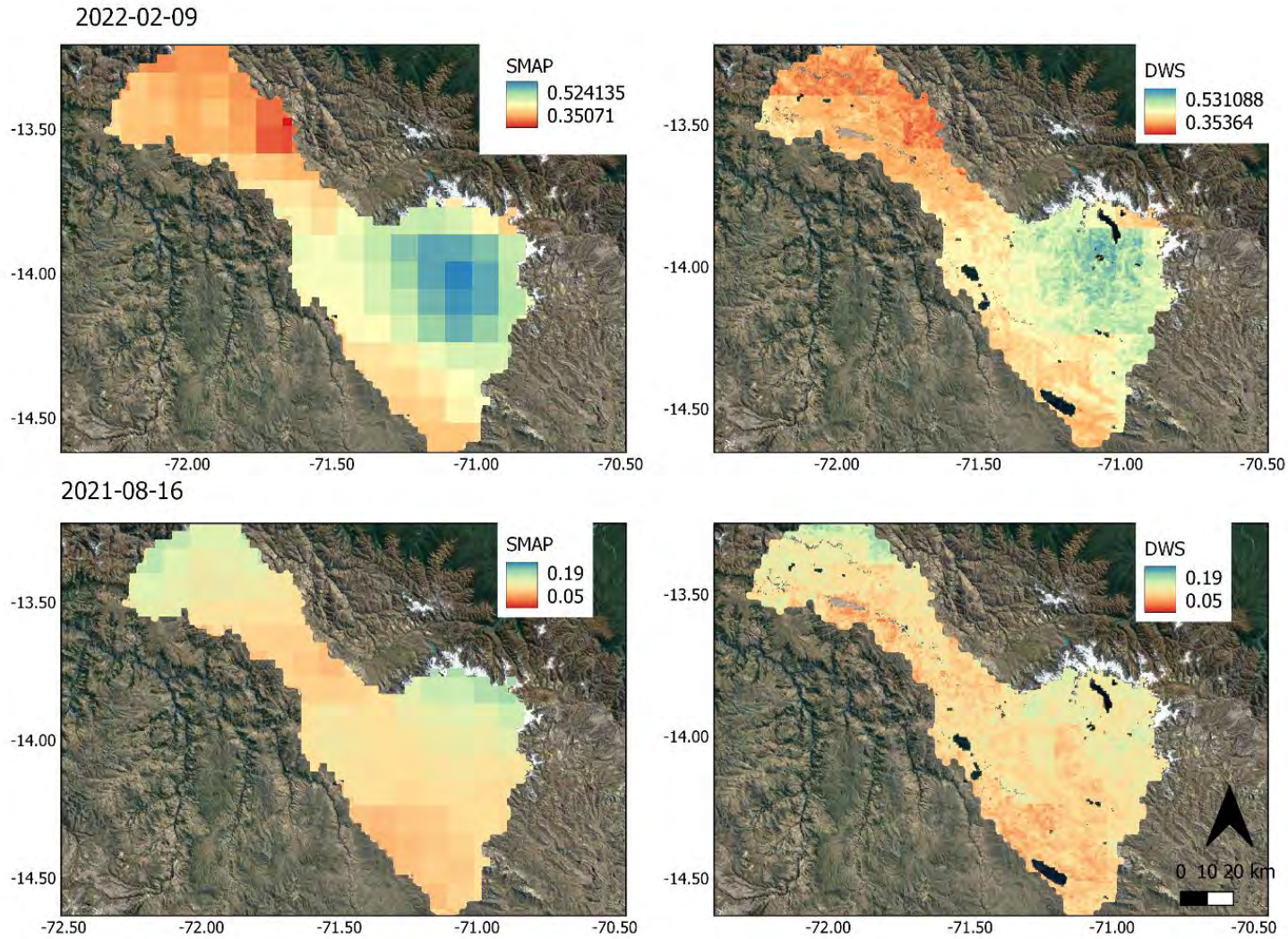


Elaborado por Marcelo Bueno Dueñas. Nota: De derecha a izquierda y de arriba abajo: pixel original del producto SMAP-L3-E a 9km de resolución espacial. Desagregación a 3km. Desagregación a 1 Km. Desagregación a 250 m y finalmente desagregación a 100m (producto final).

1.1.1. Evaluación visual de la desagregación espacial.

La figura 17 muestra la distribución original de la data SMAPL3E y la distribución desagregada mediante el método propuesto para el 2022-02-09, fecha representativa de la época húmeda, y para el 2021-08-16, fecha representativa de la época seca.

Figura 17.*Distribución espacial del producto original y del producto desagregado*



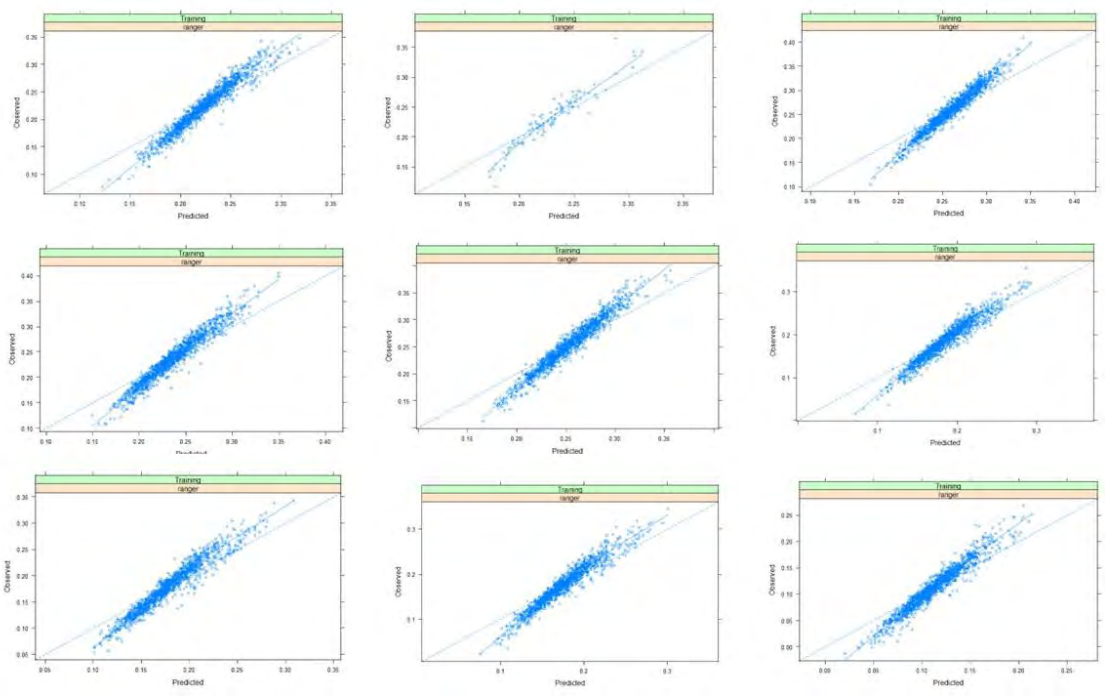
Elaborado por Marcelo Bueno Dueñas: SMAP es el contenido volumétrico de humedad del suelo en $cm^3\ cm^{-3}$ del producto SMAPL3E a 9 km de resolucion espacial. DWS es el contenido volumétrico de humedad del suelo en $cm^3\ cm^{-3}$ del producto SMAPL3E desagregado (*downscaled*) a 100 m de resolucion espacial.

La figura muestra que la distribución espacial del producto SMAPL3E antes de la desagregación es bastante consistente con la distribución del producto desagregado. La correspondencia entre las distribuciones espaciales de la humedad del suelo en época húmeda es óptima y se observa que los rangos de humedad se mantienen en ambos mapas (0.35 a 0.53 cm^3 cm^{-3}), también las propiedades de variación espacial se mantienen (clusterización y autocorrelación), en la época seca la desagregación sobreestima en las regiones de baja humedad, y sobre estima en regiones de alto contenido de agua en el suelo, en ambos casos, aproximadamente entre 0.03 a 0.04 cm^3 cm^{-3}. Sin embargo mantiene las propiedades de variación espacial, lo cual es especialmente valioso en muchos aplicaciones (Vergopolan et al., 2021, 2022).

6.1.5. Evaluación estadística de modelos de desagregación espacio-temporal.

La figura 18 muestra la dispersión entre observaciones y predicciones de los primeros 10 modelos (fechas del 02 de abril hasta el 11 de abril del 2015) al soporte original de la humedad del suelo (SMAP L2E ~9 km). En general se pudo apreciar la buena correlación entre los puntos, entre sin embargo los puntos no están distribuidos de forma homogénea a lo largo de la línea 1:1, en general se puede observar que lo modelos subestiman en condiciones de alta humedad o casi saturación del suelo (superiores a 0.35 cm^3 cm^{-3}), puntos distribuidos encima de la línea 1:1, y sobre estiman en condiciones de secamiento o baja humedad (menores a 0.15 cm^3 cm^{-3}), puntos distribuidos por debajo de la línea 1:1.

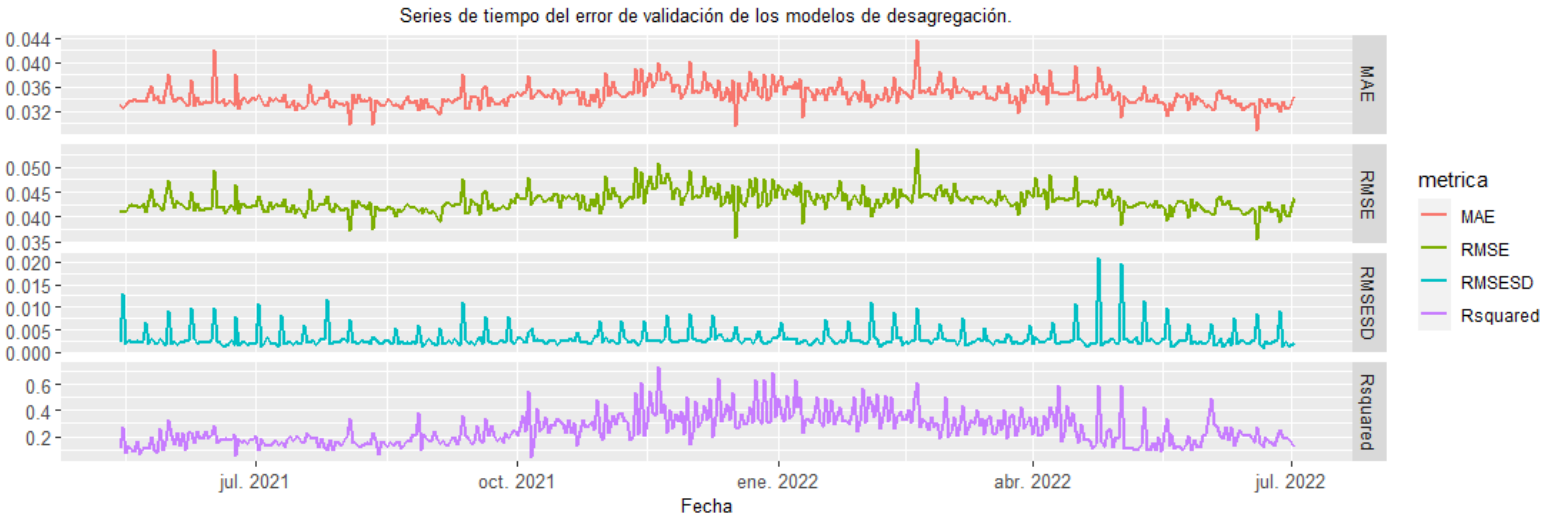
Figura 18 *Diagramas de dispersión de los nueve primeros modelos de desagregación espacial.*



En la tabla A-1 de los anexos se muestran las métricas de validacion para los 400 modelos diarios correspondientes con el periodo de monitoreo de la humedad del suelo en la estación K´ayra (14 de mayo del 2021 al 02 de julio del 2022).

En la figura 19 se puede observar la distribución de las métricas de error y su variación respecto al tiempo.

Figura 19.*Series de tiempo del MAE, RMSE, RMSE y coeficiente de determinación*



Elaborado por Marcelo Bueno Dueñas: MAE es error absoluto medio, RMSE es el error cuadrático medio, RMSESD es la desviación estándar del error cuadrático medio, Rsquared (R^2) es el coeficiente de determinación.

Todos los modelos tienen similar comportamiento en cuanto al RMSE, con oscilaciones alrededor de los 0.040 a 0.045 $cm^3\ cm^{-3}$ con picos de hasta 0.050 $cm^3\ cm^{-3}$, es claro que en promedio el RMSE es más alto en época húmeda (entre noviembre a marzo) y regresa a valores alrededor de 0.04 $cm^3\ cm^{-3}$ en los modelos entrenados en época seca (de mayo a septiembre), en general el RMSE está dentro de los limites esperados de exactitud del producto SMAP (0.04 a 0.06 $cm^3\ cm^{-3}$) (Chaubell, 2016). La desviación estándar del error cuadrático medio (RMSESD), muestra que la oscilación del error medio a lo largo del tiempo es aproximadamente aleatoria, pero con valores especialmente grandes en mayo y junio y mucho más pequeños en época húmeda. Esto significa las predicciones de los modelos entrenados en época húmeda tienen más sesgo respecto a las predicciones en época seca, algo que ha sido observado repetidas ocasiones en otros estudios.

Respecto al coeficiente de determinación R^2 este muestra un comportamiento similar, con valores bajos en época seca (0.20 a 0.30) y más altos en época húmeda (~ 0.40), este resultado es especialmente interesante y se ampliará en la sección de discusiones.

En la tabla 21 se pueden observar las métricas de error agrupadas por mes.

Tabla 21. Métricas de validación de los modelos de desagregación espacial por mes

Mes	RMSE.	Coeficiente de determinación R^2.	MAE.	RsquaredSD
Enero	0,044221	0,38345	0,035383	0,069279
Febrero	0,044143	0,360641	0,035293	0,081015
Marzo	0,04363	0,303492	0,034927	0,072434
Abril	0,044034	0,261313	0,03533	0,078903
Mayo	0,042264	0,155633	0,033871	0,074581
Junio	0,041907	0,199341	0,033637	0,073758
Julio	0,042098	0,157451	0,03366	0,06742
Agosto	0,041493	0,170123	0,03323	0,069134
Septiembre	0,042235	0,199677	0,033786	0,080795
Octubre	0,043515	0,279027	0,034653	0,07034
Noviembre	0,045433	0,359112	0,036262	0,070427
Continuación tabla 21				
Diciembre	0,044497	0,376604	0,035607	0,07403

Elaborado por Marcelo Bueno Dueñas: MAE es error absoluto medio, RMSE es el error cuadrático medio, RMSESD es la desviación estándar del error cuadrático medio, Rsquared (R^2) es el coeficiente de determinación

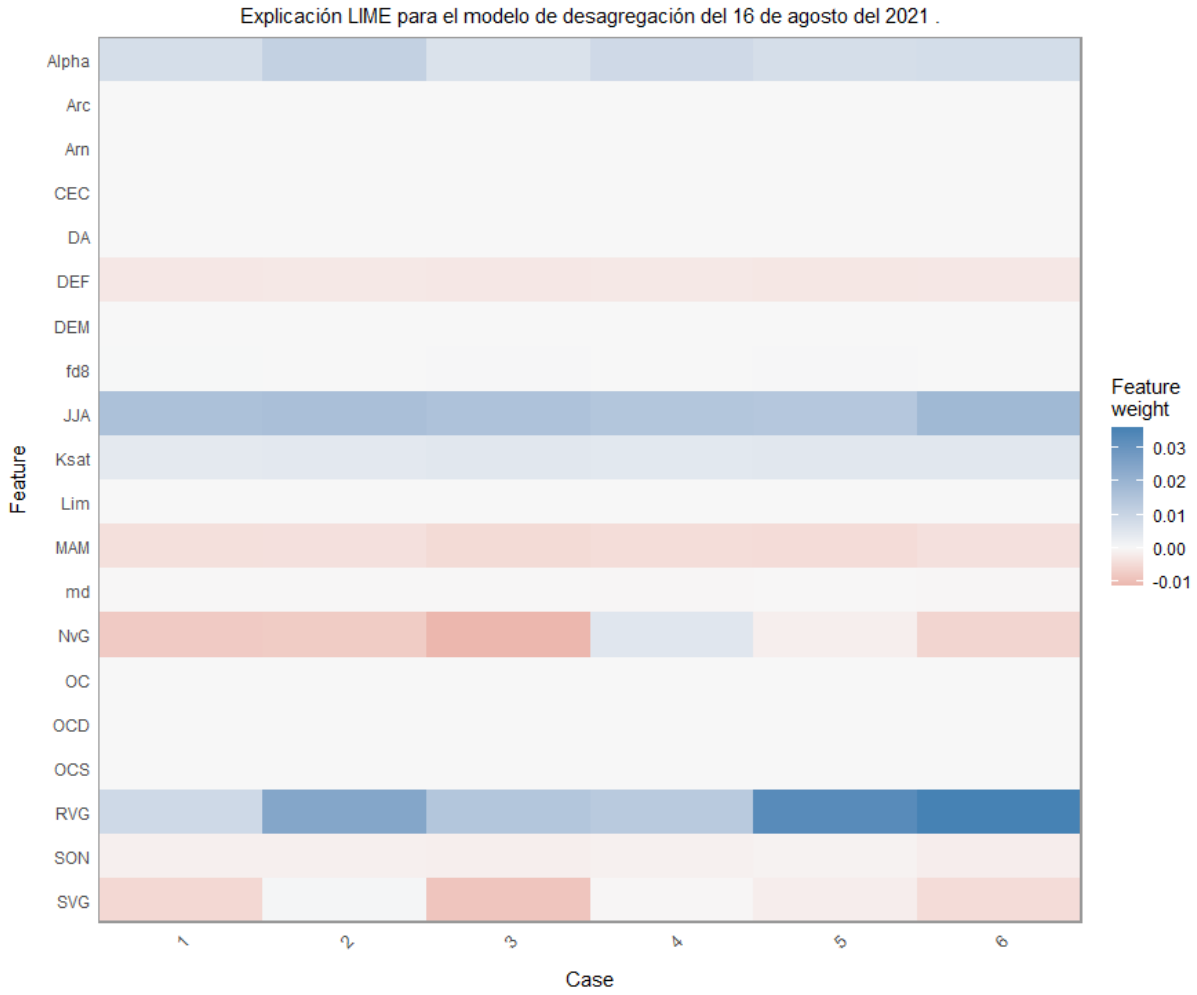
Al igual que la gráfica de series de tempo de RMSE, en el caso mensual el RMSE se distribuye de forma similar durante todos los meses.

6.1.6. Interpretación de los modelos de desagregación.

6.1.6.1. Explicaciones interpretables locales - LIME.

Para el análisis LIME del modelo del 16 de agosto del 2021 de la figura 20 los casos 1, 2, 3, 4, 5 y 6 son los valores para cada pixel de humedad del suelo θ_{SMAP} correspondientes a 0.126, 0.048, 0.167, 0.101, 0.115 y 0.131 $cm^3\ cm^{-3}$ respectivamente, y las predicciones del modelo fueron 0.108, 0.112, 0.161, 0.111, 0.107 y 0.110 $cm^3\ cm^{-3}$ respectivamente, con bondad de ajuste del modelo de aproximación LASSO entre 0.28 a 0.21. Podemos observar que en época seca las predicciones sobreestiman θ_{SMAP} , para el caso 2 en particular (θ_{SMAP} = 0.04 vs. $\theta_{SMAP-RF}$= 0.10, diferencia de 0.06 cm3 cm-3 o 6 % de contenido de humedad del suelo).

Figura 20. *Diagrama interpretativo LIME.*

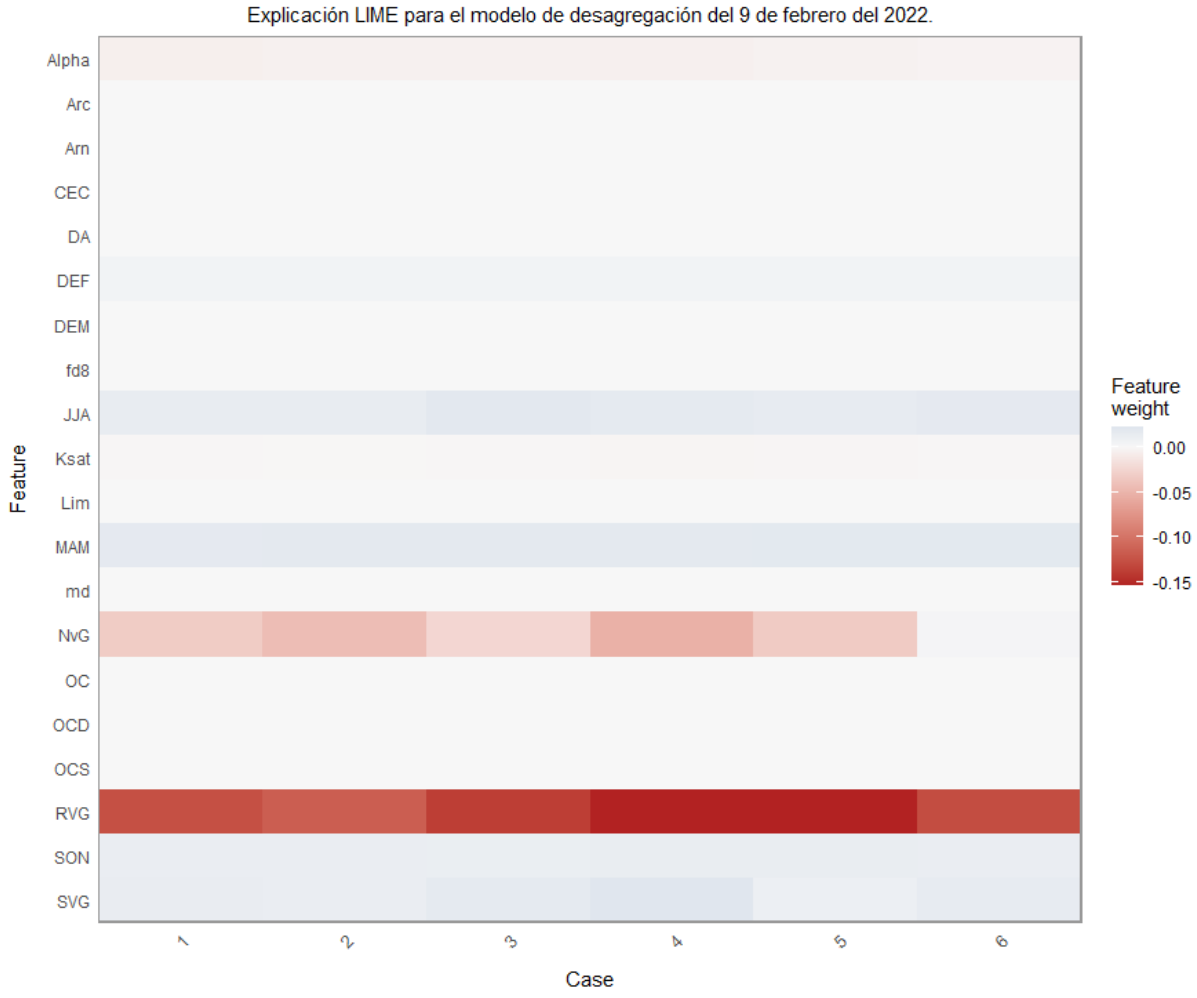


Elaborado por Marcelo Bueno Dueñas

Para el análisis LIME del modelo del 9 de febrero del 2022, cuyos resultados se aprecian en la figura 21 los casos 1, 2, 3, 4, 5 y 6 son los valores para cada pixel de humedad del suelo θ_{SMAP} correspondientes con 0.473, 0.509, 0.440, 0.515, 0.393 y 0.492 $cm^3\ cm^{-3}$ respectivamente, y las predicciones del modelo fueron 0.472, 0.452, 0.392, 0.443, 0.361 y 0.492 $cm^3\ cm^{-3}$ respectivamente, podemos observar que en época húmeda las predicciones son mucho mejores que en época seca. En general el ajuste de aproximación

del modelo LIME es mejor en el modelo entrenado en época húmeda (valores de ajuste de aproximación entre 0.50 a 0.53).

Figura 21*.Diagrama interpretativo LIME.*



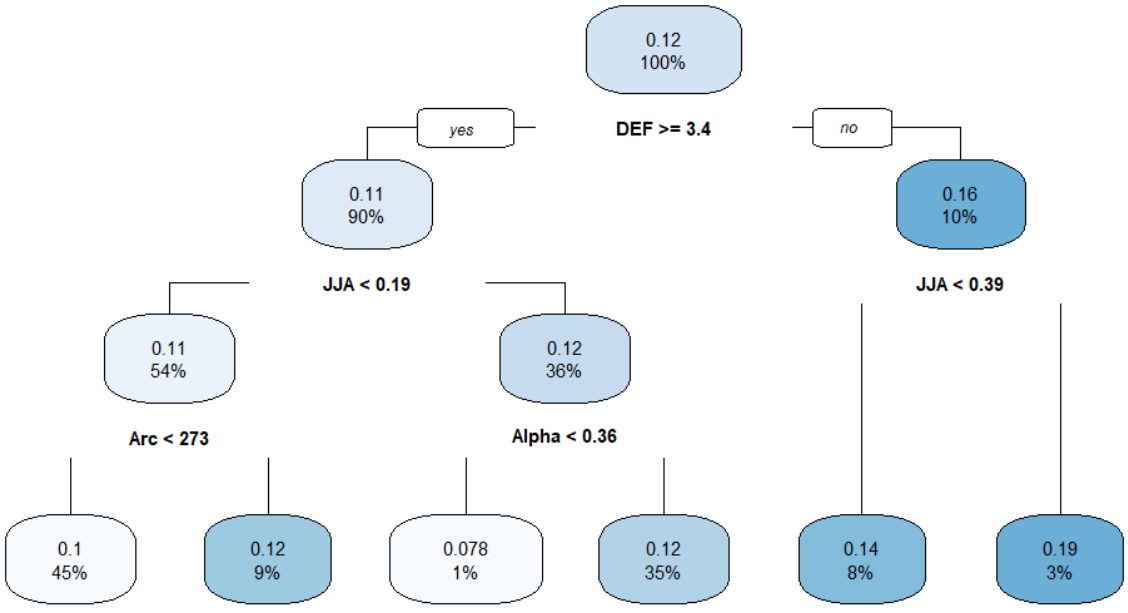
Elaborado por Marcelo Bueno Dueñas.

6.1.6.2. Aproximación mediante árboles de regresión.

En árbol de regresión entrenado en época seca (16 de agosto del 2021, figura 22) se aprecia que la θ_{SMAP} es mejor predicha por la media histórica de precipitación del producto PISCO (Imfeld et al., 2021) los meses de DEF (época lluviosa) y la media histórica de

precipitación del producto PISCO JJA (junio julio y agosto), cuando JJA es menor a 0.19 mm día^{-1} la humedad del suelo θ_{SMAP} por lo general oscila entre 0.1 a 0.12 cm^3 cm^{-3} y cuando JJA es mayor a 0.19 mm día^{-1} la humedad del suelo θ_{SMAP} tiene un rango entre 0.14 a 0.19 cm^3 cm^{-3}. Estos resultados demuestran la importancia de la época del año donde se hace la predicción y además la importancia de la resolucion espacial de las covariables, el producto pisco PISCO tiene una resolucion aproximada de 10 Km, similar a la de θ_{SMAP}, por lo tanto, la influencia es más directa. Además en la figura 24 se puede observar la influencia del contenido de arcilla del suelo, a menor contenido de arcilla 273 g Kg^{-1} > menor contenido de humedad (0.1 cm^3 cm^{-3}), mayor contenido de arcilla 273 g Kg^{-1} > mayor humedad del suelo (0.12 cm^3 cm^{-3}), pero la influencia de la precipitación es lo principal a mencionar (Brocca et al., 2007, 2016; Vergopolan et al., 2022).

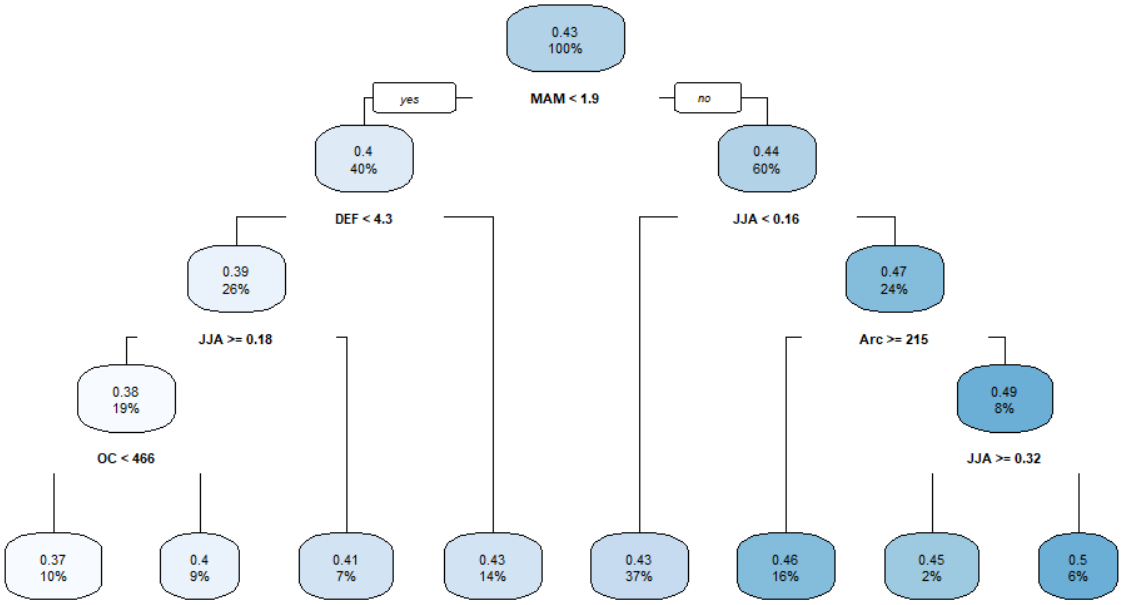
Figura 22*.Árbol de regresión, para el 16 de agosto del 2021*



Elaborado por Marcelo Bueno Dueñas.

Un análisis similar se muestra para el árbol de regresión entrenado el 9 de febrero del 2022 (figura 23). Al igual que en el modelo entrenado en época seca (18 de agosto), la precipitación media histórica de PISCO es la principal covariable que modula la distribución de humedad del suelo del modelo.

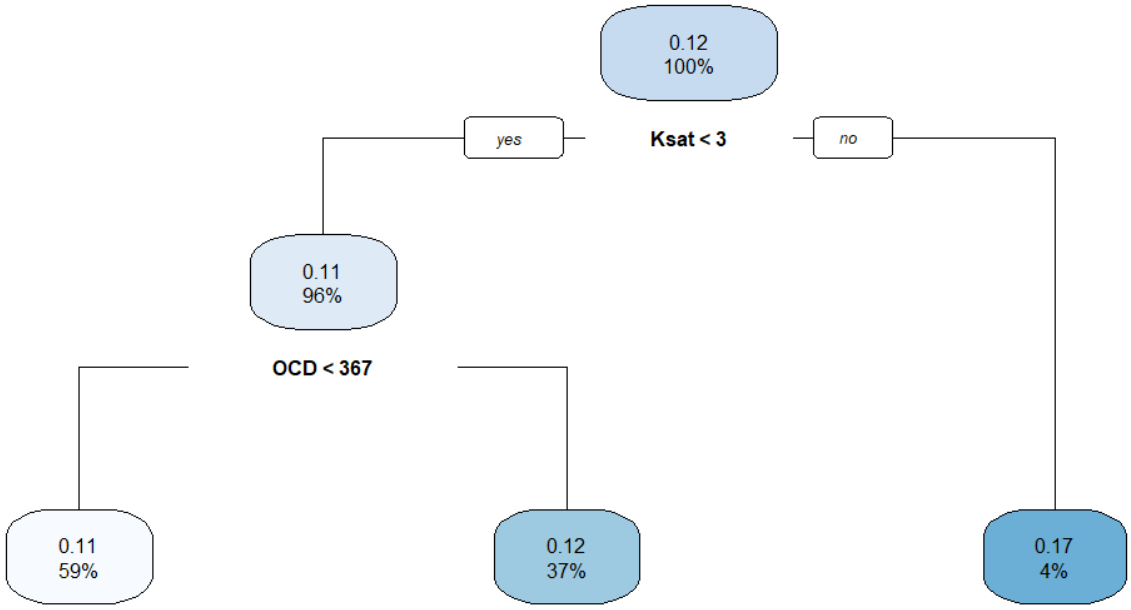
Figura 23*.Árbol de regresión, para el 9 de febrero del 2022.*



Elaborado por Marcelo Bueno Dueñas.

Para aislar la influencia de la precipitación (el mayor *driver* de la θ_{SMAP}) se entrenaron arboles de regresión sin PISCO. Los resultados se muestran en la figura 24.

Figura 24. *Árbol de regresión, para el 16 de agosto del 2021. Sin PISCO.*



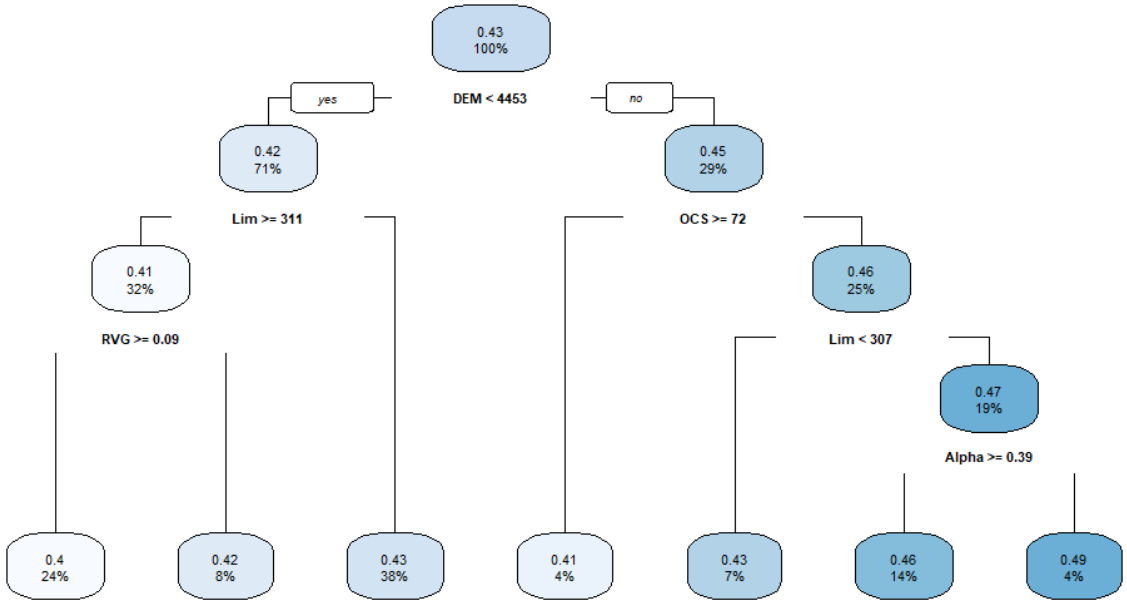
Elaborado por Marcelo Bueno Dueñas.

El árbol de regresión para el 16 de agosto del 2021 indica que *Ksat* [cm día^{-1}] determinar la distribución de θ_{SMAP} , *Ksat* menor a 3 produce θ_{SMAP} significativamente menores (0.11 a 012 cm^3 cm^{-3}) que *ksat* (Gupta et al., 2021) mayores a 3 que produce un θ_{SMAP} de 0.17, este comportamiento también esta modulado por la densidad de carbono orgánico [g dm^{-3}] OCD del *soilGrids* (de Sousa et al., 2020), si OCD es menor a 367 g dg^{-1} la humedad θ_{SMAP} será menor,

El árbol de regresión para el 9 de febrero del 2022 (figura 25) indica la influencia de la elevación de la superficie (DEM) en θ_{SMAP}, a elevaciones mayores a 4463 m la el contenido de agua de suelo de θ_{SMAP} es por lo general menor (0.4 a 0.43 cm^3 cm^{-3}),

mientras que en elevaciones menores la θ_{SMAP} es por lo general mayor (0.41 a 0.49 cm cm), indicando la influencia de la gradiente de elevación en la distribución de la humedad del suelo a lo largo del area de estudio.

***Figura 25**.Árbol de regresión, para el 9 de febrero del 2022. Sin PISCO.*



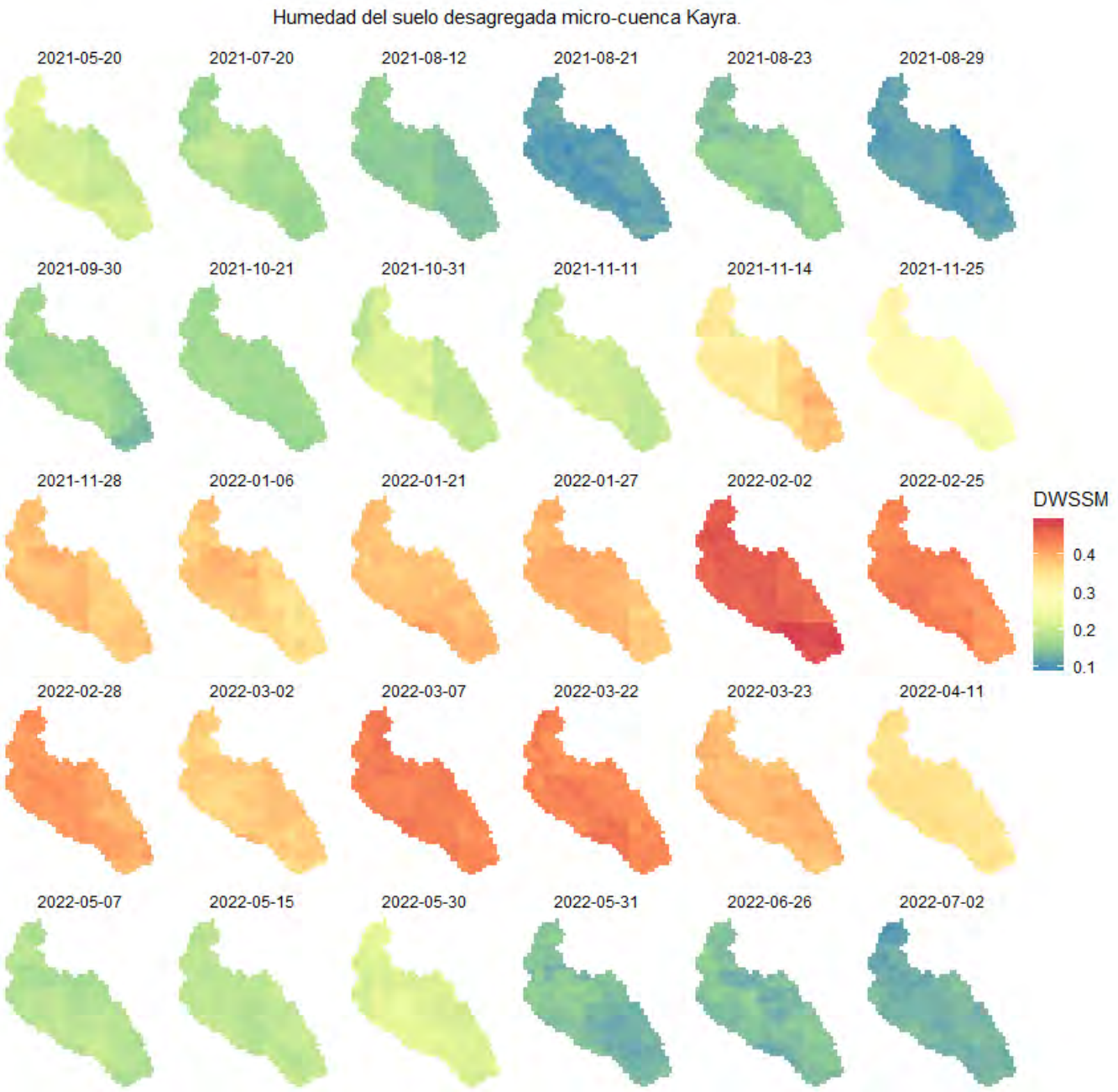
Elaborado por Marcelo Bueno Dueñas.

6.1.7. Generación de mapas de humedad del suelo a alta resolución.

Solo una vez evaluados los modelos y asegurándonos de su poder de capacidad de capturar las relaciones no lineales entre las covariables y la humedad del suelo, estos fueron aplicados en la desagregación de la humedad θ_{SMAP} (~ 9km) para predecir el contenido de agua del suelo θ_{DWS} a altas resoluciones espaciales $(\sim 100m)$.

En la figura 26 se muestran los resultados de la desagregación espacial para la microcuenca K´ayra, cada mapa fue construido aplicando un modelo de desagregación para cada fecha en particular. La imagen muestra el contenido volumétrico de humedad del suelo aproximadamente a 100 m de resolución para fechas en el periodo de monitoreo (El anexo A muestras los mapas para los 400 días de monitoreo). En la figura 26 es posible observar la distribución espacial de la humedad del suelo, determinada principalmente a esta resolucion por la hidro-topografía y las propiedades del suelo. También es posible observar la dinámica diaria de humedad del suelo, determinada principalmente por la precipitación. La capacidad del esquema de desagregación permite observar la variabilidad del proceso de flujo de agua en el suelo y su redistribución a altas escalas espacio-temporales.

Figura 26*.Mapas a alta resolución del producto SMAP-L3-E desagregado a 100 m, para diferentes fechas entre mayo del 2021 a julio del 2022 en la microcuenca K´ayra.*



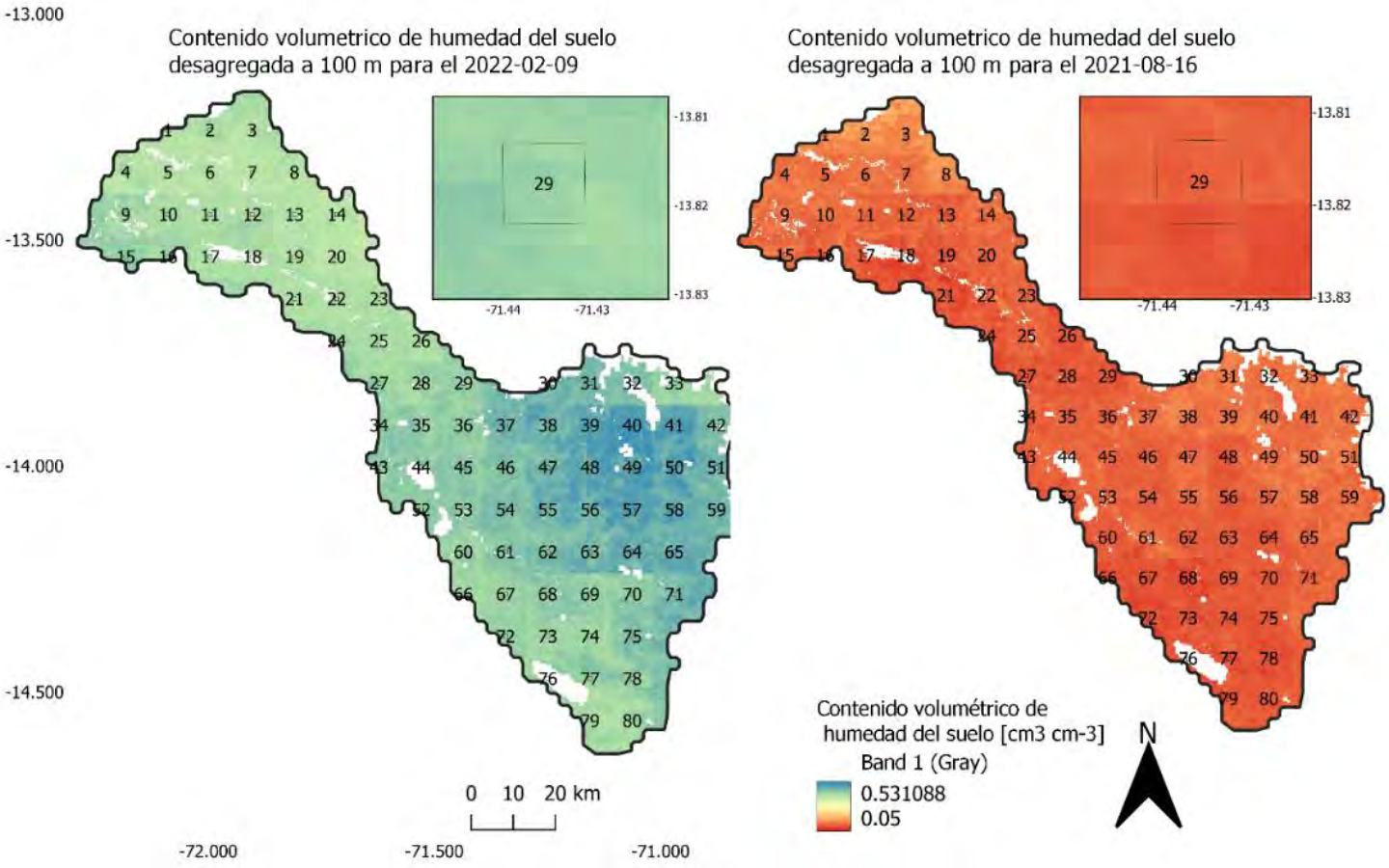
Elaborado por Marcelo Bueno Dueñas: DWSSM es el contenido volumétrico de humedad del suelo cm^3 cm^{-3} del producto SMAPL3E desagregado a 100 m de resolución espacial en la cuenca K´ayra.

6.2. Determinación de la influencia de la topografía, las propiedades del suelo y la precipitación en la dinámica espacial del producto SMAP-L3-E desagregado mediante *random forest* en el área de estudio.

6.2.1. Análisis espacial del producto SMAP-L3-E desagregado mediante *random forest*.

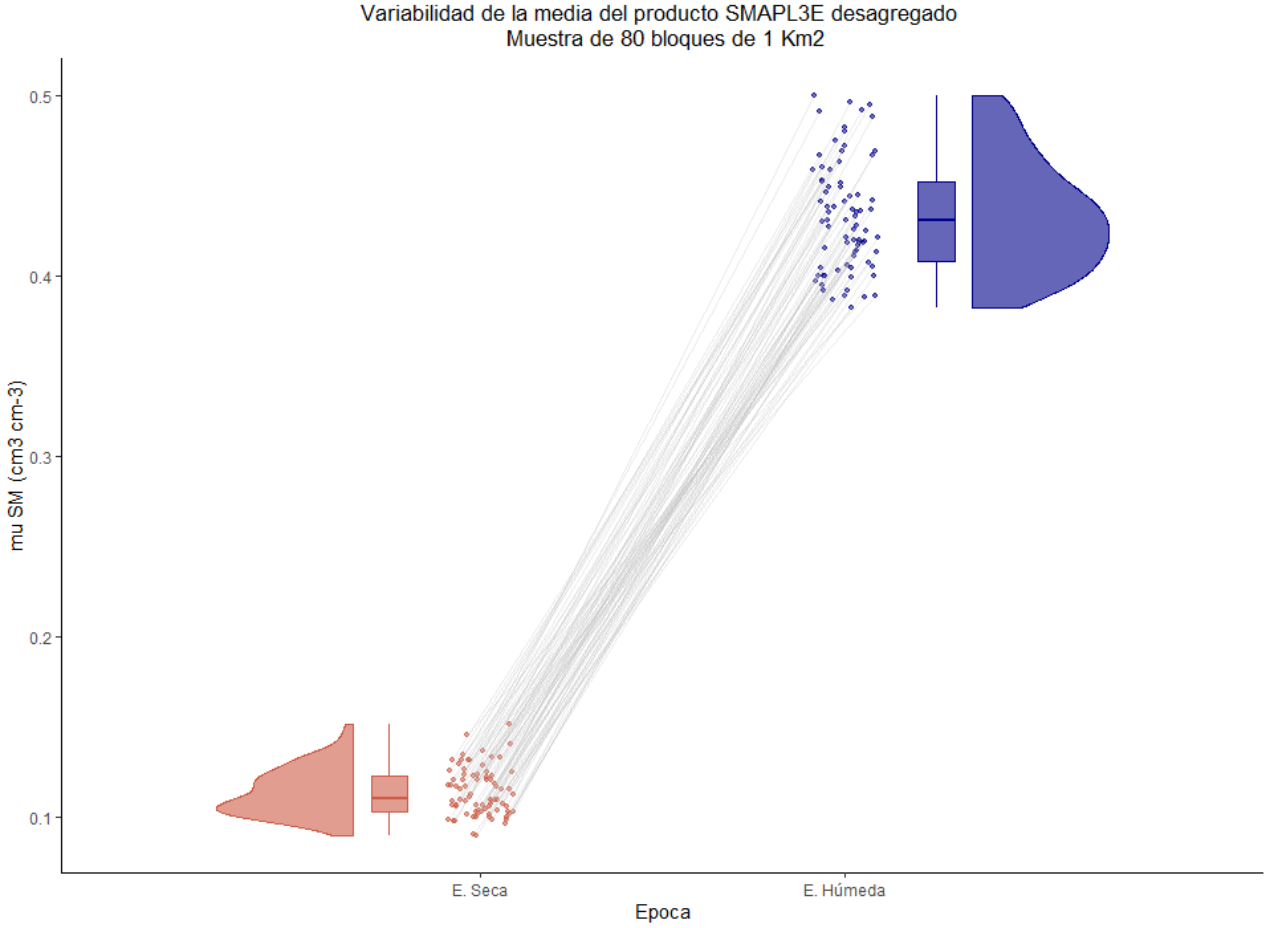
En la figura 27 se muestran las ubicaciones de los 80 polígonos muestreados para el análisis espacial, como ejemplo se presenta el polígono número 29 agrandado, cada polígono tuvo 1 km^2 de superficie.

Figura 27*. Esquema del análisis espacial con PCA.*



Así mismo la figura 28 muestra la media de la humedad desagregada a 100 m para 80 bloques de 1 km^2 cada uno, en la figura 31 se puede notar que en época húmeda la variabilidad de las medias del contenido de agua del suelo desagregada (0.38 a 0.52 cm^3 cm^{-3}) es mucho mayor que en época seca (0.08 a 0.14 cm^3 cm^{-3}).

Figura 28. *Diagramas de distribución de la media espacial de la humedad desagregada para 80 polígonos y dos fechas representativas de la estacionalidad hidrológica*

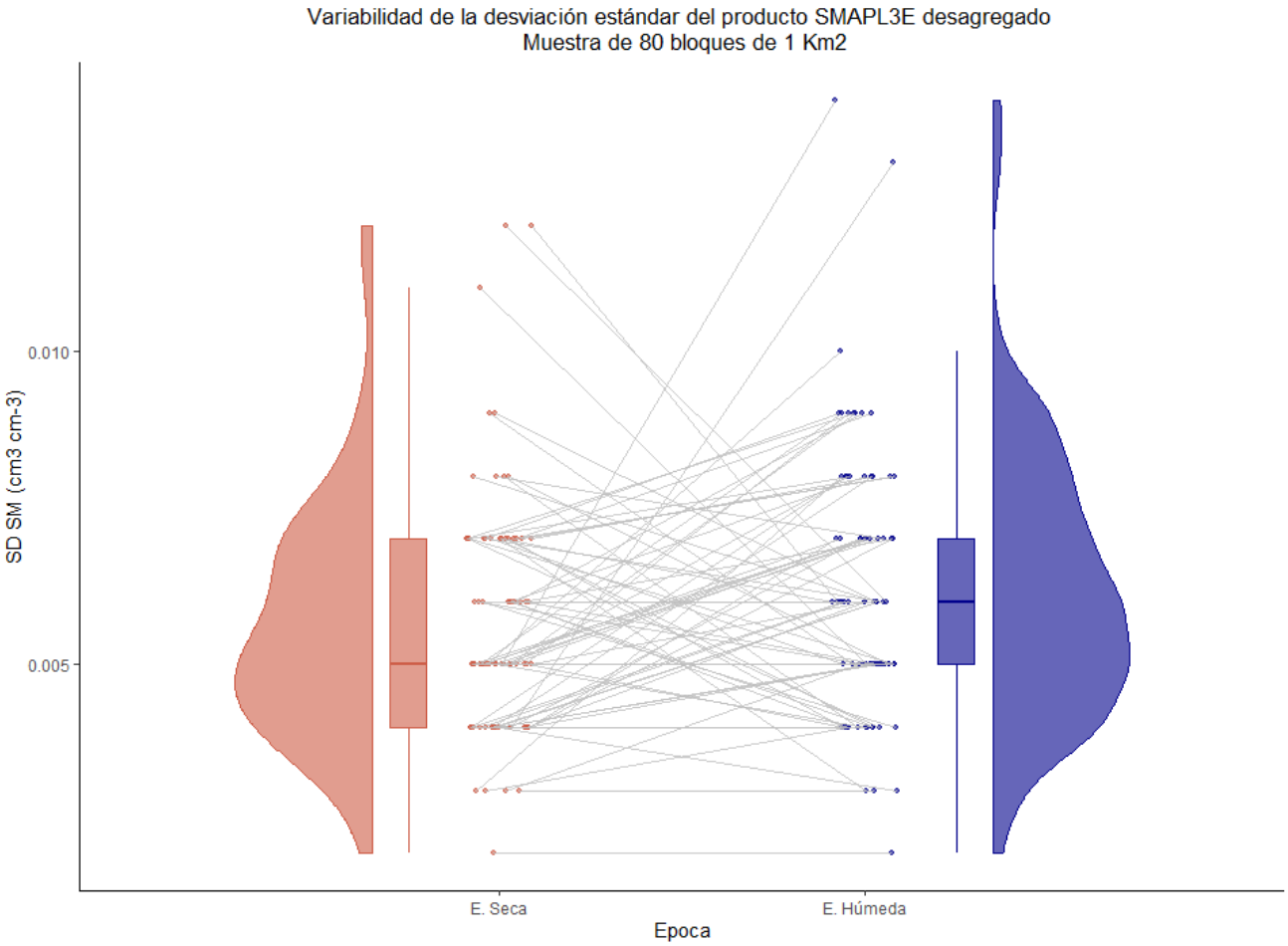


Elaborado por Marcelo Bueno Dueñas

De la figura 29 se puede señalar que el contenido de agua en el suelo -18 de agosto del 2021- en época seca en general posee la misma variabilidad espacial que en época húmeda

- 9 de febrero del 2022 - (desviaciones estándar de 0.05% en promedio), esto para los 80 bloques de 1 km2 cada uno.

Figura 29. *Diagramas de distribución de la desviación estándar espacial de la humedad desagregada para 80 polígonos y dos fechas representativas de la estacionalidad hidrológica*



Elaborado por Marcelo Bueno Dueñas

6.2.2. Evaluación de los factores relacionados con la distribución espacial del producto SMAP-L3-E desagregado mediante *random forest*.

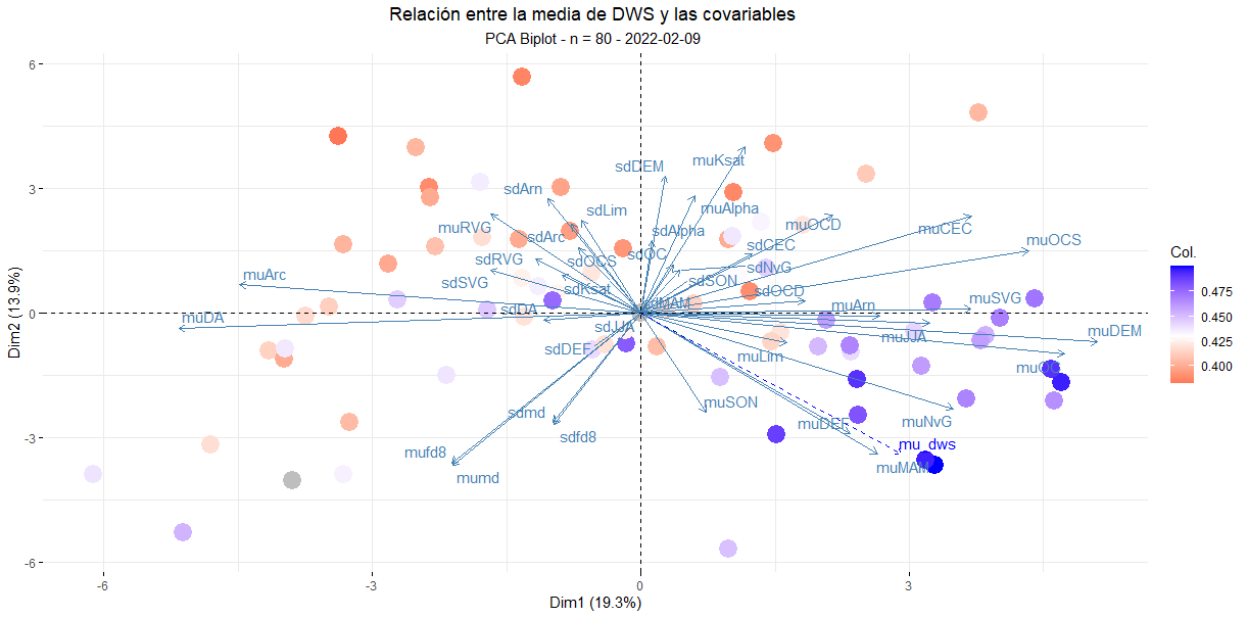
Las figuras 30 y 31 muestran los *biplot* de análisis de componentes principales de las covariables tanto para el 16 de agosto del 2021 como para el 9 de febrero de 2022 y su

relación con la media y desviación estándar espacial del producto SMAP-L3-E desagregado.

6.2.2.1. Análisis PCA para la época lluviosa.

En general se puede observar una tendencia en la media espacial de la humedad del suelo desagregada en época húmeda (9 de febrero del 2022) μDWS_H, esta tiende a seguir el primer componente principal (PC1), el cual está modulado positivamente por las variables μDEM, μOC, y negativamente por μDA y μARC (a mayor contenido de arcilla y densidad aparente μDWS_H tiende a ser menor y a mayor altura media, y contenido de carbono orgánico μDWS_H tiende a ser más alto). En cuanto a la desviación estándar de la desagregación ρDWS_H esta muestra una tendencia mucho menos organizada, pero también a lo largo del PC1. El PC2 parece no influir en la variabilidad espacial de ρDWS_H.

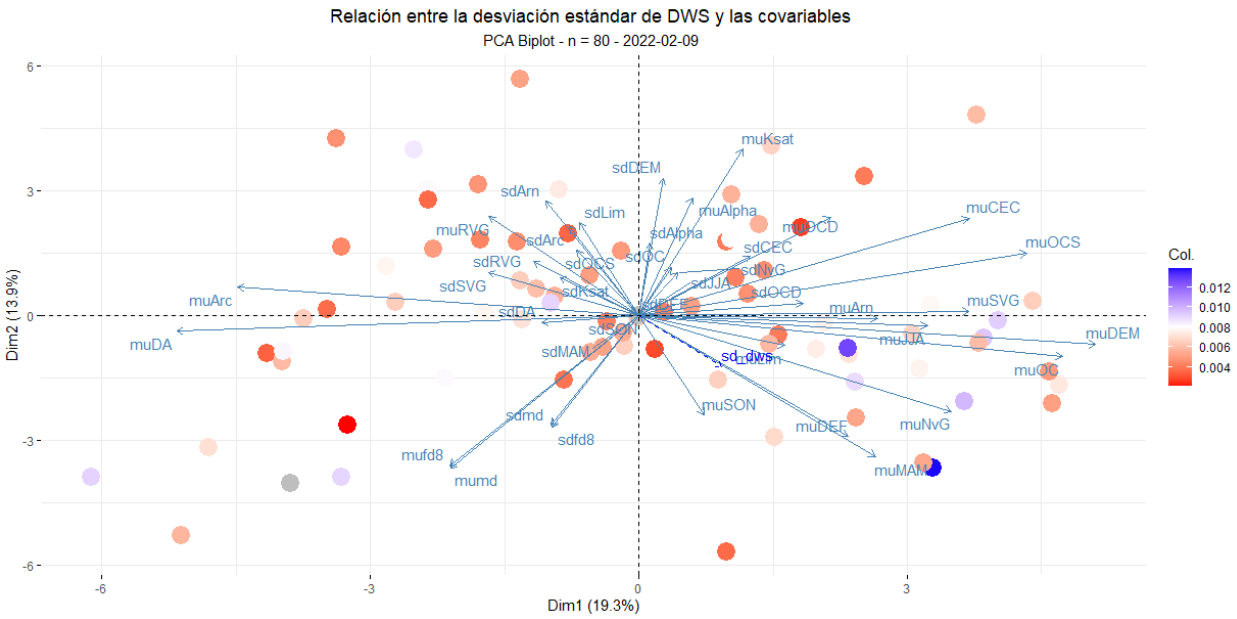
Figura 30. *Diagrama de biplot para las covariables para el 9 de febrero del 2022 y su relación con la media espacial de la humedad del suelo*



Elaborado por Marcelo Bueno Dueñas

Figura 31. *Diagrama de biplot para las covariables para el 9 de febrero del 2022 y su relación con la desviación estándar espacial de la humedad del suelo*

.



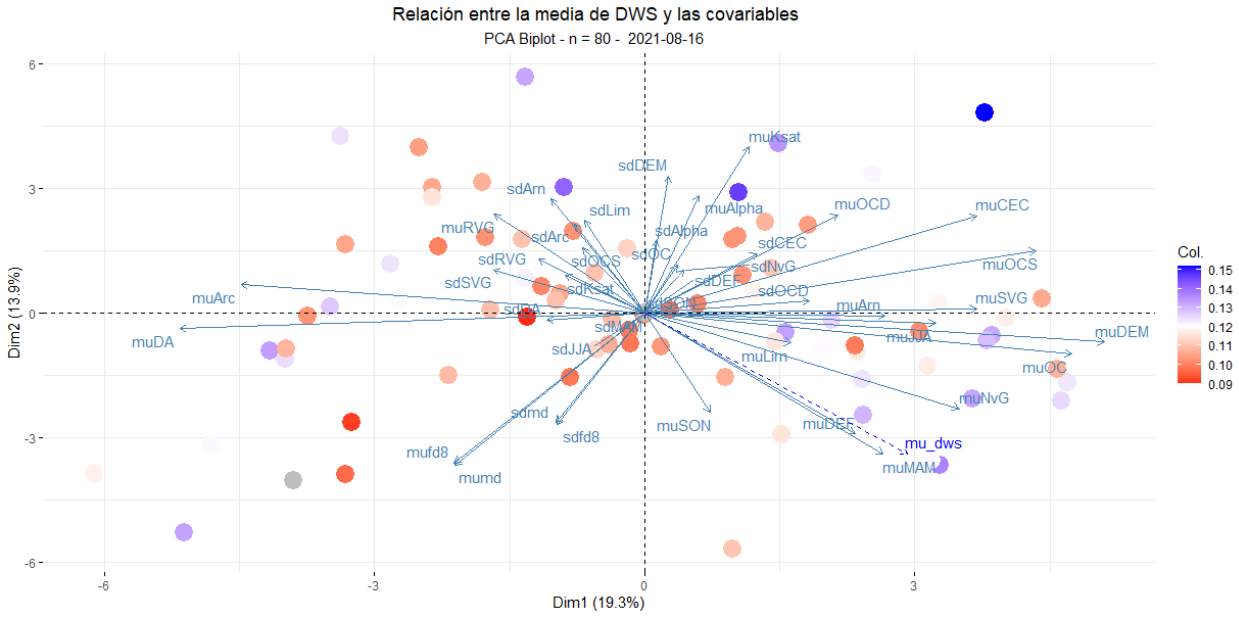
Elaborado por Marcelo Bueno Dueñas

6.2.2.2. Análisis PCA para la época seca.

En época seca los patrones de distribución de la humedad no son tan fuertes (figuras 32 y 33), y se aprecia una mayor variabilidad no explicada por las covariables de forma completa, sin embargo, aún se puede apreciar una tendencia a lo largo del PC1, las variables ya no explican tan bien la distribución de la humedad. Aún la ρDEM y el μKsat son responsables de valores altos de la humedad del suelo, lo mismo que μMAM, pero en general se aprecia mayor heterogeneidad o aleatoriedad en la distribución de la humedad del suelo. A su vez parece que el PC2 modula mejor la humedad del suelo media, particularmente la variación de la altura ρDEM, el contenido medio de arena del suelo μARN, μAlpha y la conductividad hidráulica saturada μKsat parecen modular condiciones

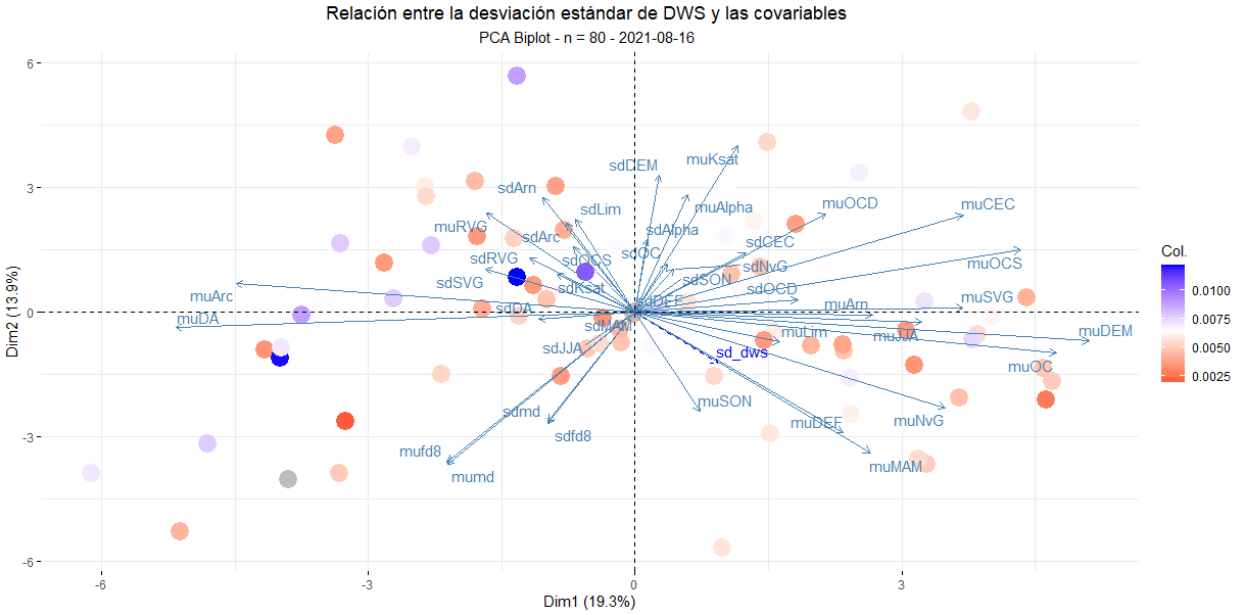
de humedad alta en épocas secas. En cuanto a la desviación estándar de la desagregación en época seca ρDWS_S esta parece no ser modulada significativamente por ninguna covariable, a lo largo de PC1 la ρDWS_H se distribuye de forma desorganizada con valores altos y bajos de humedad del suelo por igual (comparar con la distribución de $\mu\ DWS_H$).

Figura 32.*Diagrama de biplot para las covariables para el 18 de agosto del 2021 y su relación con la media espacial de la humedad del suelo*



Elaborado por Marcelo Bueno Dueñas

Figura 33*.Diagrama de biplot para las covariables para el 18 de agosto del 2021 y su relación con la desviación estándar espacial de la humedad del suelo*



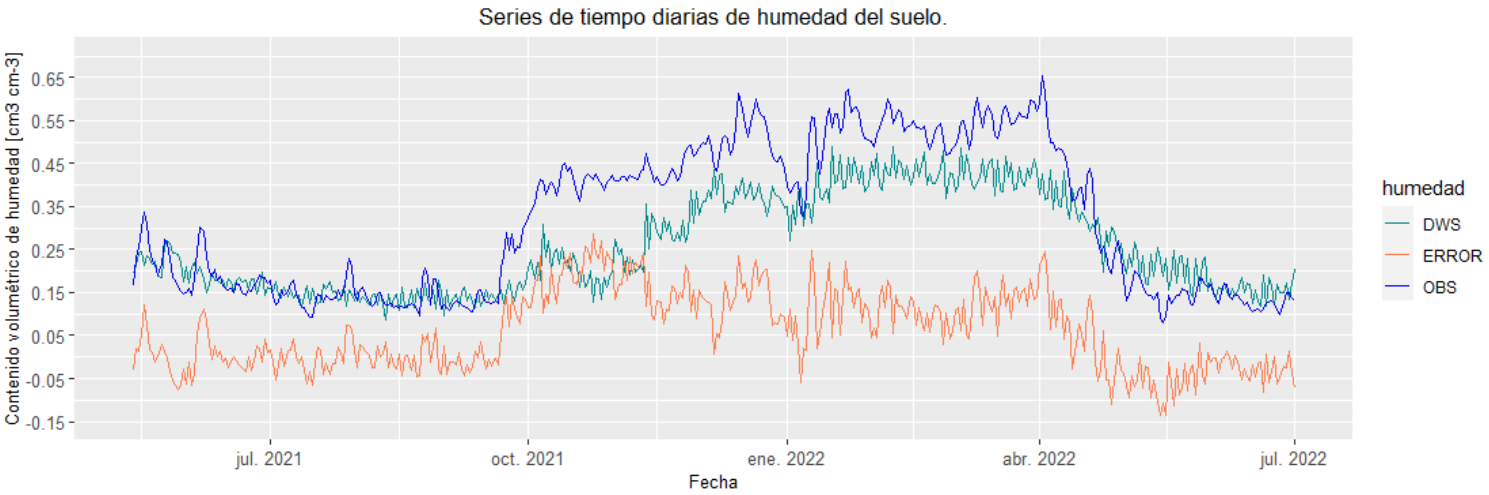
Elaborado por Marcelo Bueno Dueñas

6.3. Análisis de la relación entre el producto SMAP-L3-E desagregado mediante *random forest* con la humedad del suelo medida *in situ* en el área bajo estudio.

6.3.1. Monitoreo de la humedad del suelo.

La figura 34 muestra la gráfica de series de tiempo diarias de las mediciones *in situ*, del producto SMAP-L3-E desagregado y de la diferencia entre ambas series de tiempo. Se puede observar la buena concordancia entre las dos series de tiempo. La concordancia entre mayo a agosto es casi perfecta con diferencias de entre 0.01 a 0.05 cm^3 cm^{-3}.

Figura 34*.Series de tiempo de la humedad del suelo observada mediante monitoreo y de la humedad del suelo desagregada para el pixel de monitoreo*



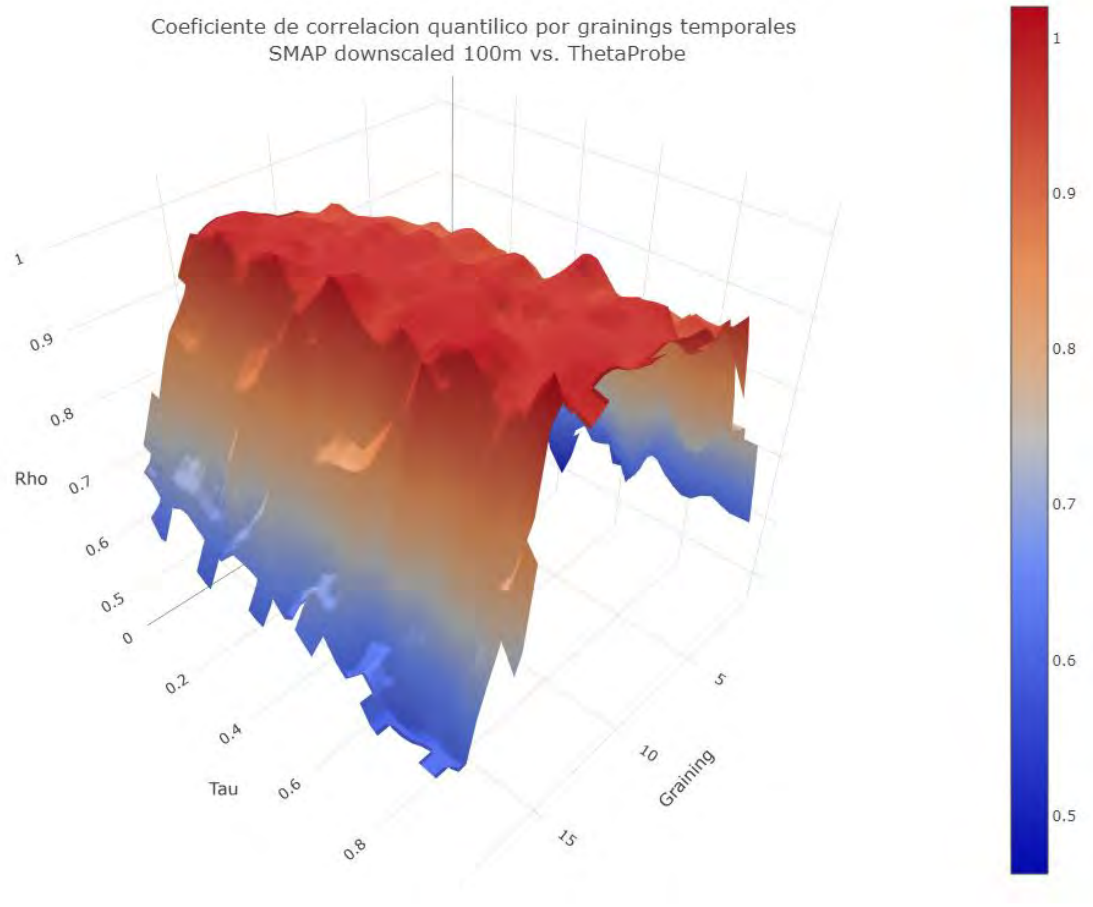
Elaborado por Marcelo Bueno Dueñas: DWS es el contenido volumétrico de humedad del suelo $cm^3\ cm^{-3}$ del producto SMAPL3E desagregado a 100 m de resolución espacial. OBS es el contenido volumétrico de humedad del suelo $cm^3\ cm^{-3}$ observado en campo mediante monitoreo diario con el sensor *ThetaProbe* ML3. ERROR = OBS – DWS.

6.3.2. Validación del producto desagregado del SMAP-3L-E.

6.3.2.1. Coeficiente de correlación cuantílico multiescala (MQCC).

La figura 35 muestra la correlación entre OBS y DWS a diferentes escalas y en diferentes cuantiles, en ella se puede observar la correlación entre la humedad del suelo desagregada (DWS) y las observaciones *in situ* de la humedad (OBS) con el sensor *ThetaProbe* ML3.

Figura 35.*Coeficiente de correlación cuantílico.*



Elaborado por Marcelo Bueno Dueñas.

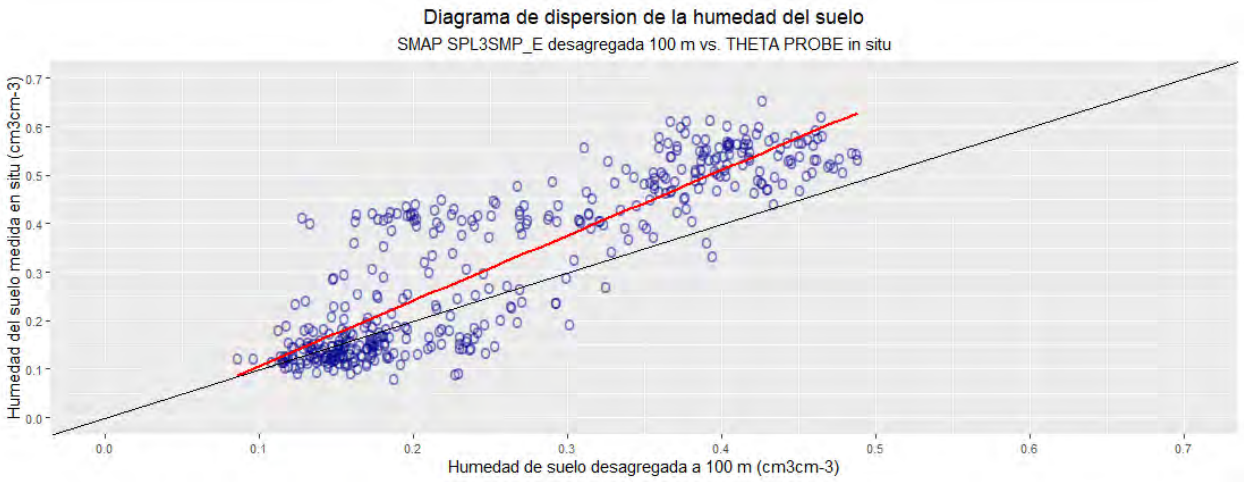
De la gráfica tridimensional de las correlaciones entre DWS y OBS en función de la agregación en la escala temporal (*granings*) y de los cuantiles considerados (tau), se puede observar claramente que la humedad del suelo desagregada muestra un coeficiente de correlación positivo con la humedad medida en campo, generalmente mayor al 0.5, y el coeficiente de correlación cuantílica (ρ_τ) aumenta con el aumento del nivel de agregación temporal. Al nivel de las series de tiempo originales (*graining* = 1, diarias), la correlación entre DWS y OBS fluctúa entre 0.54 a 0.62. En agregaciones de tiempo moderadas (6 días

≤ *grainings* ≤ 14 días) el coeficiente correlación fluctúa entre 0.93 a 0.99 independientemente de cuantil considerado.

6.3.2.2. Gráficos de dispersión y Gráfico Q-Q.

La figura 36 muestra el gráfico de dispersión entre las observaciones diarias del producto SMAP-3L-E y la humedad del suelo observada en campo para aproximadamente 400 días.

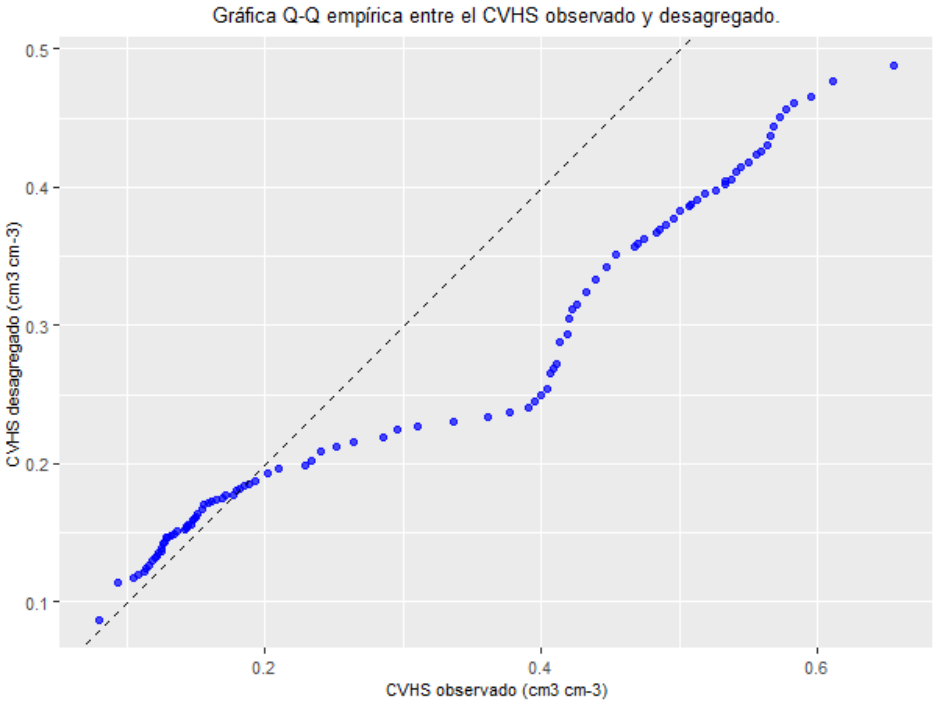
Figura 36*.Diagrama de dispersión entre la humedad observada in situ y la humedad desagregada*



Elaborado por Marcelo Bueno Dueñas.

En la figura 37 se muestra el *q-q plot* entre las series de tiempo de la humedad del suelo desagregada y medida i*n situ*.

Figura 37.*Diagrama q-q entre la humedad observada in situ y la humedad desagregada.*



Elaborado por Marcelo Bueno Dueñas: CVHS es contenido volumétrico de humedad del suelo.

VII. DISCUSIÓN DE RESULTADOS.

En este estudio se propuso un método de generación de información de humedad del suelo diaria y de alta resolución espacial, con la finalidad de disponer de información relevante en la toma de decisiones agrícolas a nivel de parcela (~ 1 hectárea). Con tal fin los resultados obtenidos son discutidos y posibles errores u omisiones son resaltadas de manera que permitan mejorar las estimaciones en posteriores estudios para el beneficio de la agricultura.

7.1. Evaluación de la capacidad de desagregación espacio-temporal mediante *random forest* del producto SMAP-L3-E en el área de estudio.

7.1.1. Producto SMAP-L3-E.

Las figuras 10 y 11 permiten concluir que el producto SMAP-L3-E es capaz de capturar la dinámica temporal del contenido de agua en el suelo de forma coherente a la división tradicional del año hidrológico en los andes peruanos (Imfeld et al., 2021), esta observación y los estadísticos por mes del producto SMAP-L3-E respaldan esta conclusión, sin embargo es necesario comparar las series de tiempo mediante análisis estadísticos que permitan considerar tanto la variabilidad espacial del producto como errores de estimación relacionados a los sensores y modelos dieléctricos aplicados en la estimación de la humedad del suelo del SMAP (Colliander et al., 2017).

7.1.2. Análisis exploratorio de las covariables.

Los coeficientes de correlación de la figura 12 en general muestran coherencia con el conocimiento edafológico, por ejemplo, la correlación positiva entre la conductividad

hidráulica del suelo y la densidad aparente ha ido reportado en múltiples estudios (Guevara & Vargas, 2019; Gupta et al., 2021). Sin embargo la falta de correlación de los índices de humedad topográficos es un resultado no esperado, recientes estudios han mostrado que el índice de humedad topográfico es un excelente predictor de propiedades del suelo por el efecto del flujo de agua en la escorrentía y erosión (Li et al., 2020; Quinn et al., 1995; Raduła et al., 2018).

Cabe mencionar que a la escala espacial a la que se analizaron las covariables, la variabilidad a pequeña escala generalmente se pierde por el proceso de agregación espacial. El cambio de escala o agregación espacial ha sido demostrado que disminuye la variabilidad intrínseca de las propiedades del suelo y de los parámetros topográficos (Famiglietti et al., 2008; Heuvelink et al., 2021; Vergopolan et al., 2022), por lo que es de esperar que algunas correlaciones significativas a escalas más grandes se pierdan.

La relación entre la precipitación descrita por el producto CHIRPS y la humedad del suelo estimada del producto SMAP-L3-E (figura 13) mostró un comportamiento no lineal modulado por la estación hidrológica, en general a valores altos de precipitación no hay un cambio significativo en la humedad del suelo en las estaciones Q1 y Q2 (estaciones secas), en cambio en Q3 y Q4 la relación es mucho más directa y positivamente proporcional. Esto puede ser explicado por la evapotranspiración, en las estaciones Q1 y Q2 la evapotranspiración supera significativamente el aporte de la precipitación en el balance hídrico del suelo, además al ser el producto SMAP-L3-E una estimación no diaria, es incapaz de detectar la dinámica intra-diaria de la precipitación (Lu et al., 2015).

7.1.3. Construcción de los modelos de desagregación espacio-temporal.

7.1.3.1. Entrenamiento y parametrización del random forest.

En este estudio se parametrizó un *random forest* diario, a diferencia de estudios como el de (Abbaszadeh et al., 2019; Wakigari & Leconte, 2022) que parametrizaron un *random forest* con toda la data disponible para estaciones del año enteras debido al tamaño del pixel del SMAPL3E y su area de estudio pequeña, disponían solamente de 130 pixeles del SMAP, lo cual generaba una cantidad pequeña de data de entrenamiento. En nuestro caso se parametrizó un *random forest* diario con aproximadamente 1300 pixeles del SMAPL3E dentro del area de estudio, lo cual impuso restricciones computacionales por un lado, pero permitió captar la relación entre la humedad y las covariables de forma dinámica (por ejemplo, la relación entre la humedad del suelo y el contenido de arcilla varía no linealmente dependiendo del estado de saturación)(Famiglietti et al., 2008). Sin embargo esta estrategia reduce el tamaño muestral usado para entrenar los modelos diarios, en ese caso se corre el riesgo de que los modelos no dispongan de suficientes observaciones para poder captar las relaciones entre las covariables y la humedad del suelo del producto SMAP-L3-E, este es un problema que ya se ha observado anteriormente y es común en modelos de aprendizaje automático (Adab et al., 2020; Heuvelink et al., 2021).

En este este estudio se observó un comportamiento sub-óptimo de los modelos de desagregación (figura 15). Esto debido principalmente a que los modelos no fueron optimizados en el espacio de los parámetros (tabla 20).

Los motivos de esta decisión fueron principalmente la falta de poder de computación, ya que se optó por ajustar un random forest de regresión para cada fecha, aproximadamente

se entrenaron 4000 modelos desagregar el producto SMAP-L3-E de forma diaria, y el poder de cómputo necesario para hallar los valores óptimos de los parámetros mediante estrategias de parametrización y validacion más complejas (Krstajic et al., 2014) estaba fuera del alcance de la capacidad de cómputo disponible, adicionalmente aunque se hubiera realizado la calibración esta seguiría siendo una estimación mediocre del poder de desagregación de los modelos, de tal manera que se decidió estimar el error exterior (Brus, 2019) de las predicciones desagregadas mediante monitoreo de la humedad del suelo (aproximadamente 400 días) dentro de un pixel de predicción. Esta forma de validacion es superior a la estimación interna del error de los modelos mediante validacion cruzada (Guevara & Vargas, 2019; Heuvelink et al., 2021) y es la más usada en la práctica, cuando el objetivo es la desagregación espacial (Abbaszadeh et al., 2019; Bai et al., 2019).

En general, el peligro más importante al no optimizar los modelos de aprendizaje automático es el sobreajuste (Schratz et al., 2021), sin embargo, generalmente es considerado que *random forest* no genera sobreajuste (Breiman, 1999).

7.1.4. Modelos de desagregación temporal.

Los error de generalización de los modelos de reconstrucción mostrado en la figura 14 y 15 son considerados admisibles y se han reportado en múltiples investigaciones previas (Colliander et al., 2017a; Gruber et al., 2020). Sin embargo, esta distribución del error nos sugiere que debemos tomar mayores precauciones en las zonas donde el error es mayor, principalmente el error máximo (0.05 cm^3 cm^{-3}) se encuentra próximo a las coordenadas geográficas -71.00 y -14.00 de longitud y latitud respectivamente, es aquí donde debemos ajustar mejor los modelos y generar mejores predicciones.

Adicionalmente en la figura 15 se puede apreciar que existe una sobreestimación de la humedad del suelo en el proceso de reconstrucción (la humedad reconstruida usualmente es mayor al contenido de agua en el suelo en días cercanos, sobre todo en épocas relativamente más secas, como entre julio a agosto) también se aprecia una aleatoriedad muy marcada. Esto podría obedecer a la dinámica de la precipitación, pero se requieren estudios posteriores para verificar tal relación.

7.1.5. Modelos de desagregación espacial.

En este estudio el producto de humedad del suelo SMAP-L3-E (~ 9 Km) fue usado como variable base de desagregación. Un total de 2000 imágenes del SMAP y 2000 adicionales reconstruidas mediante random forest fueron usadas con tal objetivo (figura 16) que abarcaron aproximadamente 7 años de data continua (toda la data disponible de la misión SMAP). Todas las imágenes fueron utilizadas para el entrenamiento de los modelos y la estimación de los errores de generalización se hizo mediante validación cruzada de 10 iteraciones.

La metodología de desagregación basada en *random forest* permitió hacer uso de la correlación entre las covariables y el producto SMAP-L3-E a la resolucion espacial de 9 km, de tal forma que permitió estimar el contenido de agua del suelo a 100 m de resolución espacial lo cual permite representar mejor la distribución espacial de la humedad del suelo.

7.1.6. Evaluación visual de la desagregación espacial.

Una desagregación exitosa debe reproducir los patrones espaciales del producto original del que deriva (Wakigari & Leconte, 2022). La figura 16 y 17 se usó como medio para probar la correspondencia entre el producto SMAP-L3-E original y el desagregado,

dicho análisis fue realizado para dos fechas que representan dos modelos entrenados en épocas secas y época húmeda (18 de agosto y 9 de febrero respectivamente).

La inspección visual mostro una alta congruencia en la distribución espacial de la humedad, también se puede señalar que las propiedades de variación espacial se mantuvieron (*clusterización* y autocorrelación), en la época seca la desagregación sobreestima en las regiones de baja humedad, y sobre estima en regiones de alto contenido de agua en el suelo, en ambos casos, aproximadamente entre 0.03 a 0.04 cm^3 cm^{-3}. Sin embargo mantiene las propiedades de variación espacial, lo cual es especialmente valioso en muchos aplicaciones (Vergopolan et al., 2021b, 2022).

En los mapas de la figura 17, y en específico para el mapa del 2 de febrero se muestra un gradiente de humedad del suelo de norte a sur, con valores más altos en la zona sur-este, al sur la humedad es menor, en el rango de 0.35 a 0.40 cm^3 cm^{-3} y se aprecia que el producto desagregado mantiene el mismo gradiente mejorando la variabilidad a escalas más grandes, una propiedad de los campos de humedad del suelo demostrado en múltiples estudios (Famiglietti et al., 2008; Western & Blöschl, 1999).

Un resultado alentador es que la humedad del suelo desagregada estimada mediante r*andom forest* es capaz de aumentar la información espacial de la que se disponía originalmente mediante uso de las covariables. Es de esperar que a resoluciones altas la modulación de la humedad siga exactamente la distribución espacial de las covariables utilizadas en el estudio, particularmente la topografía. Este resultado es prometedor pero es necesario implementar actividades de medición de la humedad del suelo con alta densidad de muestreo (Gruber et al., 2020; Huang et al., 2020) para poder analizar si la variabilidad

indicada por la figura 17 es en realidad una señal observada en campo o un artefacto ocasionado por las covariables o el modelo.

Otros estudios han demostrado que la incorporación de covariables relacionadas a la fenología y estructura poblacional de la vegetación mejoran los resultados de la desagregación espacial, principalmente covariables a alta resolucion espacial (Vergopolan et al., 2022) son poderosas predictoras de la humedad del suelo y pueden implementarse en próximos estudios.

7.1.7. Evaluación estadística de modelos de desagregación espacio-temporal.

Las gráficas de dispersión (figura 18) y los estadísticos de error indican que los métodos de desagregación mediante *random forest* capturaron satisfactoriamente la relación no lineal entre el contenido de agua del suelo y las covariables utilizadas con relativo éxito a escalas pequeñas (bajas resoluciones espaciales).

Todos los modelos tienen similar comportamiento en cuanto al RMSE, con oscilaciones alrededor de los 0.040 a 0.045 $cm^3\ cm^{-3}$ con picos de hasta 0.050 $cm^3\ cm^{-3}$, es claro que en promedio el RMSE es más alto en época húmeda (entre noviembre a marzo) y regresa a valores alrededor de 0.04 $cm^3\ cm^{-3}$ en los modelos entrenados en época seca (de mayo a septiembre), en general el RMSE está dentro de los limites esperados de exactitud del producto SMAP (0.04 a 0.06 $cm^3\ cm^{-3}$) (Chaubell, 2016)

En cuanto al coeficiente de determinación R^2, se encuentra en promedio en un rango entre de 0.19 a 0.38 lo que indica por lo general baja performance de cada modelo para predecir la data del SMAP basado en el conjunto de covariables a la misma resolucion espacial. Es más interesante fijarse en la dinámica temporal de R^2 que se muestra en la

figura 21, se aprecian valores bajos de R^2 en época seca (0.20 a 0.30) y más altos en época húmeda (~ 0.40), este comportamiento es un resultado no esperado, por lo general se espera que el modelo actúe mejor en época seca como se demostró en multitud de estudio previos (Wakigari & Leconte, 2022). Una posible explicación en este caso y que es reforzado por resultados posteriores es que la humedad del suelo se distribuye de forma más compleja en periodos secos, donde su distribución no depende tanto de la precipitación sino de flujos subterráneos y depende más de las propiedades del suelo que son mucho más variables a escalas grandes que las propiedades topográficas o hidrológicas en el paisaje (Famiglietti et al., 2008).

Este resultado pude parecer desalentador a primera vista, sin embargo solo es ligeramente inferior a estudios previos (Beck et al., 2021) y se pude conjeturar que es debido a dos motivos en específico: Primero los modelos de desagregación no fueron optimizados como fue explicado en la sección previa, y segundo el coeficiente de determinación es calculado al soporte original del SMAP-L3-E y por lo tanto no es una métrica propia del error de desagregación sino del error de construcción de los modelos a esas escalas espaciales. Si bien es cierto que los modelos deben entrenarse con el objetivo de tener los parámetros que reduzcan el error y maximicen la varianza explicada (R^2), en realidad la calidad de las covariables y el número de observaciones determinaran la capacidad real de los modelos de predecir adecuadamente.

Siguiendo esta lógica, una de las principales debilidades del presente estudio fue utilizar como covariables bases de datos globales, que si bien han demostrado ser eficientes

en otros estudios (Vergopolan et al., 2021; Xing et al., 2017) su fiabilidad es menor en nuestro país por la propia naturaleza de los datos que se usaron para construir tales estimaciones (de Sousa et al., 2020). En particular SoilGrids fue construida con bases de datos de muestras de suelo y observaciones de perfiles globales que en general no son representativas de la zona de estudio. Esta incertidumbre asociada al uso de covariables no validadas es común en estudios que hacen uso de funciones de edafo-transsferencia en general (Gupta et al., 2021).

Además cabe recalcar que los resultados mostrados en la figura 18 y en la tabla 21 corresponden a la validacion denominada interna (Brus, 2019) , y que por lo general subestima el error real. Además ya que los modelos fueron entrenados en un soporte espacial mucho más grande del soporte al que se desean las predicciones se introducen errores adicionales (Brocca et al., 2007; Wakigari & Leconte, 2022) que no pueden ser estimados en el proceso de validacion cruzada inicial y que requieren de una validacion con data de campo.

7.1.8. Interpretación de los modelos de desagregación.

7.1.8.1. Explicaciones interpretables locales – LIME.

En general la figura 20 permite concluir que en época húmeda ninguna variable en especial influye en las predicciones θ_{SMAP} de forma específica, esto se explica porque el rango de θ_{SMAP} en esta fecha es relativamente pequeño (0.47 a 0.51 $cm^3\ cm^{-3}$), por lo que las covariables modulan la predicción de forma similar (colores similares a lo largo de todos los casos examinados). Sin embargo es importante notar la marcada influencia de las covariables NVG y RVG que representan los parámetros de la funcion de van Genuchten

(1980), NVG es un parámetro relacionado con la capacidad de retención de humedad del suelo y RVG representa el contenido residual de humedad del suelo, es decir el mínimo contenido de agua que puede almacenar el suelo. Además, se puede apreciar una ligera influencia de Ksat, JJA y MAM. En general los pesos de las demás covariables mostraron variación mínima entre ellos y no fue posible un análisis adicional.

Estos resultados demuestran la importancia de la época del año donde se hace la predicción y además la importancia de la resolucion espacial de las covariables, el producto PISCO tiene una resolucion aproximada de 10 Km, similar a la de θ_{SMAP}, por lo tanto, la influencia es más directa.

7.1.8.2. Aproximación mediante árboles de regresión.

Las figuras 22, 23, 24 y 25 demostraron la lógica interna de los modelos de desagregación espacial. Estas relaciones han sido observadas anteriormente en los trabajos de Brocca et al. (2007) y Famiglietti et al. (2008) y otros.

Era de esperar que los índices de humedad topográficos tengan influencia relevante en la distribución de la humedad como sugieren estudios como los de (Beck et al., 2021; Mohanty et al., 2000; Vergopolan et al., 2022), pero a resoluciones tan bajas parece que su efecto es disipado por las covariables a resoluciones más gruesas.

En época seca y época húmeda las covariables interactúan de forma diferente y producen diferentes modelos, esa es la importancia de ajustar un modelo de desagregación de forma diaria. En las figuras 22 y 23 se puede apreciar dos árboles de regresión ajustados

en dos fechas diferentes del periodo de monitoreo (16 de agosto del 2021 y 9 de febrero del 2022). Claramente se puede observar una señal en las covariables, que los modelos aprovechan para generar predicciones. Las figuras 24 y 25 siendo árboles de regresión entre el producto SMAPL3E y las covariables a excepción de PISCO permiten concluir que tanto el DEM como Ksat influyeron en el modelamiento, pero sus efectos fueron disminuidos por los efectos de PISCO, un fenómeno que se mencionó anteriormente.

7.1.9. Generación de mapas de humedad del suelo a alta resolución.

La evaluación visual de los modelos de desagregación a ambas escalas espaciales (9 km y 100 m) y la evaluación estadística sugirieron la capacidad de los modelos de desagregar el producto SMAP-L-E con adecuada precisión.

Por lo tanto, los modelos fueron implementados para la predicción de la humedad del suelo a 100 m de resolucion espacial en la microcuenca K´ayra.

Es fundamental postular que el supuesto de que los modelos *random forest* construidos a bajas resoluciones espaciales también son válidos para predecir la humedad del suelo a más altas resoluciones espaciales usando las covariables a la resolucion deseada debe ser cierto. En otras palabras, se asume que los modelos de desagregación son invariantes respecto a la escala espacial, una propiedad que no siempre es posible comprobar y que varía en cada caso particular como lo sugieren estudios previos (Fang et al., 2018).

En algunos mapas de la figura 26 pueden observarse ciertos artefactos (regiones que cambian abruptamente de contenido de humedad del suelo), tras una cuidadosa evaluación, se puede postular que el motivo es que los modelos para esas fechas en particular fueron entrenados con pocas observaciones y esto generó predicciones en un rango de covariables

con poca variabilidad produciendo el resultado final, esto se puede apreciar en el mapa generado con el modelo entrenado el día 12 de agosto del 2021.

7.2. Determinación de la influencia de la topografía, las propiedades del suelo y la precipitación en la dinámica espacial del producto SMAP-L3-E desagregado mediante *random forest* en el área de estudio.

7.2.1. Análisis espacial del producto SMAP-L3-E desagregado mediante *random forest.*

Los resultados del análisis y la figura 28 muestran la alta variabilidad espacial de la humedad en épocas lluviosas, esto ha sido descrito anteriormente en algunos estudios como los de Famiglietti et al. (2008), Mohanty et al., (2000) y Western & Blöschl (1999). Los resultados de este estudio tienen coherencia con la dinámica de la humedad descrita por los autores mencionados anteriormente, además el área de estudio es significativamente grande y es de esperar que exista variabilidad de la humedad, principalmente por las gradientes de precipitación. Considerando que cada bloque fue muestreado dentro de solo un pixel del producto SMAP-L3-E original a 9 km de resolución espacial y el muestreo no se hizo con reemplazo se obtuvo una representación adecuada de la distribución de humedades para los pixeles originales del SMAP-L3-E de 9 km.

Según el análisis de las figuras 29, 30 y 31, parece que el producto SMAP-L3-E desagregado a 100 m es incapaz de describir la variabilidad espacial natural de la humedad del suelo dentro de superficies de 1 km^2 (desviaciones estándar de 0.05% en promedio para ambas épocas hidrológicas), además de ser insensible a la influencia de la condición general de saturación en la desviación estándar espacial de la humedad (ambas fechas

muestran la misma desviación estándar a pesar de tener condiciones de saturación muy diferentes) , esta variabilidad es una propiedad intrínseca de la humedad del suelo y ha sido descrita en numerosos estudios anteriores (Vergopolan et al., 2021). En general se puede postular que es por ese motivo que el análisis de PCA realizado posteriormente no fue capaz de producir resultados coherentes en el caso de la desviación estándar espacial del contenido de agua en el suelo desagregado a 100 m a partir del producto SMAP-L3-E.

7.2.2. Evaluación de los factores relacionados con la distribución espacial del producto SMAP-L3-E desagregado mediante random forest.

La interpretación del *biplot* de las figuras 30 a 31, permitió hacer visible la relación entre la humedad del suelo desagregada y los factores ambientales que determinan su variabilidad espacial. Es apreciable que la humedad del suelo sigue los gradientes espaciales de precipitación del producto PISCO, los puntos con mayor μDWS_H son aquellos donde la μMAM y μDEF son mayores, lo que significa que la precipitación histórica media descrita por PISCO en los meses de diciembre a mayo explican la distribución de zonas con alto contenido de humedad del suelo.

El segundo componente principal es dominado negativamente por σDEM y μKsat y positivamente por μfd8, y μmd, lo que significa que a mayor variabilidad de la elevación y a mayor media de la conductividad hidráulica saturada del suelo la humedad promedio será menor y a su vez que a mayor variabilidad espacial de los índices de humedad topográfico la media de la humedad del suelo es mayor, sin embargo el PC2 solo es capaz de expresar el 13% de la variabilidad total de las variables, por lo que su capacidad explicativa de la variabilidad de la humedad del suelo es menor que la del PC1.

La heterogeneidad del suelo y sus propiedades como la textura, contenido de materia orgánica, densidad aparente y conductividad hidráulica saturada son factores que influencian la capacidad de almacenamiento de agua de los suelos, así como la velocidad de flujo y redistribución de la humedad, lo que a su vez se puede observar dispara la heterogeneidad espacial de la media de la humedad desagregada, estas observaciones coinciden con (Brocca et al., 2007; Crow et al., 2012; Famiglietti et al., 2008). Esta heterogeneidad del suelo tiene un papel mucho más complejo en zonas de aguas subterráneas poco profundas como bofedales, que requieren un análisis adicional apoyado por modelos hidrológicos y data de monitoreo de niveles freáticos. Por ejemplo Warner et al., (2021) obtuvieron excelentes resultados en la desagregación del SMAP mediante el modelo KNN en la red de monitoreo CONUS (Estados Unidos), excepto en áreas dominadas por bofedales (*wetlands*) donde el modelo subestimo sistemáticamente la humedad del suelo.

A su vez, las características topográficas e hidrológicas como la elevación de la superficie, el índice de humedad topográfico modulan la variabilidad de la humedad del suelo hacia zonas de convergencia mediante flujo superficial/escorrentía o flujo lateral subsuperficial. Los resultados muestran una alta variabilidad de la humedad del suelo relacionada a los índices de humedad topográfica μfd8, y μmd y su variabilidad espacial σfd8, y σmd demostrando el rol de la topografía en la distribución de la humedad (Beven & Freer, 2001; Li et al., 2020; Quinn et al., 1995; Raduła et al., 2018) particularmente en épocas húmedas del año (Western & Blöschl, 1999).

De hecho, los resultados sugieren que localidades altas (mayor μDEM) y de condiciones hidro-topográficas de alta divergencia (bajos μfd8, y μmd) y poco variables (bajos σfd8, y σmd), además con suelos altos en contenido de materia orgánica (mayor μOC y μOCS) están relacionadas con condiciones de humedad del suelo mayor en la zona de estudio.

Finalmente, la variabilidad espacial de la humedad del suelo se considera mayor en épocas húmedas (Famiglietti et al., 2008). Con el objetivo de analizar la variabilidad de la humedad tanto en época húmeda como seca se realizaron dos análisis adicionales del coeficiente de variabilidad de la humedad del suelo tanto en época seca como húmeda, sin embargo, estos análisis como el de la desviación estándar espacial de la humedad del suelo fueron altamente difíciles de interpretar y sugieren que los campos de humedad generados no dan pista de la heterogeneidad intrínseca de la humedad del suelo en condiciones naturales.

7.3. Análisis de la relación entre el producto SMAP-L3-E desagregado mediante *random forest* con la humedad del suelo medida *in situ* en el área bajo estudio.

7.3.1. Monitoreo de la humedad del suelo.

La figura 34 muestra las series de tiempo obtenidas mediante monitoreo y mediante la desagregación del producto SMAP-L3-E reconstruido.

A partir de octubre el producto desagregado subestima la humedad del suelo de forma significativa generando diferencias de entre 0.15 a 0.25 cm^3 cm^{-3}, específicamente en noviembre, en este periodo las dos series de tiempo mostraron un comportamiento inverso. A partir de diciembre las dos series de tiempo tienden a aproximarse más y las diferencias

disminuyen nuevamente a valores cercanos a 0.05 cm^3 cm^{-3}. Después de abril el producto desagregado tiende a sobre estimar la humedad del suelo en el pixel donde se encontró la estación de monitoreo con el resultado de que las diferencias entre las dos series de tiempo son negativas en este periodo -0.05 cm^3 cm^{-3}.

7.3.2. Validación del producto desagregado del SMAP-3L-E.

7.3.2.1. Coeficiente de correlación cuantílico multiescala (MQCC).

El análisis MQCC permitió analizar la relación entre las dos series de tiempo, este análisis permite realizar las siguientes observaciones que se aprecian la figura 35: A niveles de agregación temporal a escala semanal la relación entre DWS y OBS es fuerte. En agregaciones de tiempo más grandes (14 días $\leq$ *grainings* $\leq$ 20 días) el coeficiente de correlación disminuye fluctuando entre 0.60 a 0.69. Cabe mencionar que, según el gráfico QQCC el coeficiente de correlación entre DWS y OBS se maximiza en agregaciones temporales de entre 6 a 14 días, siendo su valor máximo en *grainings* de 10 días.

Adicionalmente se esperaba que los coeficientes de correlación varíen significativamente a diferentes niveles cuantílicos (por ejemplo, que la correlación sea mayor en el cuantil $\tau = 0.1$ que en el cuantil $\tau = 0.9$, el gráfico de dispersión y el grafico quantil-quantil empírico sugieren que la correlación es mayor en condiciones de baja humedad (cuantiles pequeños) y menor condiciones mayor humedad (cuantiles mayores), sin embargo, el comportamiento de la relación entre DWS y OBS en diferentes cuantiles es constante e independiente del valor de τ.

Por ejemplo, tanto en valor de $\tau = 0.1$ (considerando valores de humedad bajos de las series de tiempo de OBS y DWS) como en $\tau = 0.9$ (considerando valores de humedad más

altos de las series de tiempo de OBS y DWS) el coeficiente de regresión fluctúa entre 0.65 a 0.95 modulado solamente por la escala de agregación temporal.

Estos resultados muestran consistencia con investigaciones previas, por ejemplo (Hu et al., 2020) al desagregar SMAP obtuvieron correlaciones para 30 estaciones en Mongolia en un rango entre 0.246 a 0.705. Adicionalmente (Abbaszadeh et al., 2019) al desagregar SMAP obtuvieron coeficientes de correlación entre 0.65 a 0.70 para las 300 estaciones de monitoreo de humedad del suelo de la red CONUS a través de Estados Unidos. Tambien (Wakigari & Leconte, 2022) obtuvieron coeficientes de correlación entre 0.68 a 0.83 en su zona de estudio ubicada en la región nororiental de Estados Unidos. (Huang et al., 2020b) por su parte validó una estrategia de desagregación del SMAP basada en *random forest* quantilico (QRF) en diferentes redes de monitoreo a través del mundo, obteniendo coeficientes de correlación entre 0.754 a 0.632, además los autores analizaron la relación de la correlación con el uso de suelo (*land cover),* siendo el contenido de humedad mejor explicado por los modelos en pasturas (r = 0.696) que en áreas cultivadas (r = 0.624) que en bosques (r = 0.611) y que en suelos sin cobertura (r = 0.433). (Singh et al., 2019b) encontraron coeficientes de correlación entre 0.416 y 0.943.

7.3.2.2. Gráficos de dispersión y Gráfico Q-Q.

Respecto al gráfico de dispersión de la figura 36 se puede observar la acumulación de puntos en la zona de baja humedad (0.1 a 0.2 cm^3cm^{-3}), indicando el promedio general de la humedad para la zona, la línea de regresión en rojo diverge de la línea de relación perfecta 1:1 a medida que la humedad aumenta, se observa que el producto subestima la humedad observada a valores altos de contenido de agua en el suelo. La distribución de puntos es

muy parecida a la obtenida por Singh et al. (2019) con una concentración de puntos en las regiones de humedad baja a intermedia (0.10 a 0.20 cm^3 cm^{-3}).

Considerando el diagrama cuantil-cuantil de la figura 37 se aprecia que entre 0.1 a 0.2 cm^3 cm^{-3} las dos series de tiempo mantienen una relación lineal, a valores de humedad mayores a 0.30 cm^3 cm^{-3} los valores OBS medidos en campo son significativamente mayores a DWS, estas diferencias entre DWS y OBS son máximas cerca a los 0.4 cm^3 cm^{-3} de contenido volumétrico de humedad del suelo (por ejemplo DWS estima la humedad en 0.25 a 0.30 cm^3 cm^{-3} cuando en realidad las mediciones en campo indican que la humedad es aproximadamente 0.4 cm^3 cm^{-3}, a valores mayores de 0.4 cm^3 cm^{-3} y cercanos a 0.5 cm^3 cm^{-3} DWS no subestima tanto la humedad como en valores cercanos a 0.4 cm^3 cm^{-3} pero igual se mantiene una subestimación de aproximadamente 0.1 cm^3 cm^{-3} o 10% de humedad del suelo.

VIII. CONCLUSIONES Y SUGERENCIAS.

La presente tesis propuso evaluar una técnica de aprendizaje automático denominada *random forest* para mejorar la resolución espacio-temporal de las estimaciones remotas de la humedad del suelo del producto SMAP-L3-E del satélite SMAP desde el 2015 hasta el 2022 y evaluar tales predicciones durante un año hidrológico en un área de estudio perteneciente a la cuenca alto Urubamba**.**

8.1. Evaluación de la capacidad de modelos basados en *random forest* para desagregar espacio-temporalmente el producto SMAP-L3-E en el área de estudio.

Después de entrenar los modelos de desagregación espacio-temporales usando *random forest* como función de desagregación, quedó demostrado que la desagregación temporal (reconstrucción de las series de tiempo) asemeja adecuadamente la dinámica temporal del producto SMAP-L3-E. Respecto a la desagregación espacial, la validación visual mostró que la desagregación es coherente con la distribución original de la humedad del suelo, pero además la mejora significativamente. El análisis de la validación estadística en ambos casos mostró que el error de generalización de los modelos de desagregación es adecuado para aplicaciones científicas y prácticas.

Además, mediante el análisis explicativo de los modelos de desagregación espacial (LIME y arboles de regresión) es adecuado concluir que los modelos encontraron correlaciones coherentes entre las covariables y el producto SMAP-L3-E, mediante este análisis fue posible explicar por qué los modelos fueron capaces de desagregar el producto SMAP-L3-E con adecuada precisión.

En conclusión, se demostró que *random forest* es capaz de desagregar espacio-temporalmente el producto SMAP-L3- en el área de estudio.

8.2. Influencia de la topografía, las propiedades del suelo y la precipitación en la dinámica espacial del producto SMAP-L3-E desagregado mediante *random forest* en el área de estudio.

Mediante la aplicación del análisis de componentes principales de la media y desviación estándar de las covariables a alta resolución espacial en 80 polígonos muestreados sistemáticamente en el área de estudio se demostró que a altas resoluciones espaciales (~ 100 m) y en condiciones de humedad del suelo entre moderadas a altas (figura 32) el producto SMAP-L3-E desagregado depende fundamentalmente de la elevación, del contenido de carbono orgánico del suelo, del contenido de arcilla y la conductividad hidráulica saturada del suelo. En condiciones de menor contenido de agua en el suelo, su distribución se hace más aleatoria y deja de depender directamente de las covariables usadas en este estudio.

Respecto a la precipitación, esta explica la mayor parte de la dinámica espacial del producto SMAP-L3-E a resoluciones bajas (~ 9 Km).

8.3. Análisis de la relación entre el producto SMAP-L3-E desagregado mediante *random forest* con la humedad del suelo medida *in situ* en el área bajo estudio.

Mediante análisis de gráficas de dispersión y diagramas cuantil-cuantil quedó demostrado que la serie de tiempo del producto SMAP-L3-desagregado a 100 m, y la serie de tiempo de la humedad del suelo observada en campo mediante monitoreo con sensores

dieléctricos durante 400 días, muestran una relación coherente y altamente significativa entre sí.

Mas específicamente podemos concluir que el producto SMAP-L3-E desagregado mediante el modelo *ramdom forest* explica adecuadamente las mediciones *in situ* de la humedad del suelo en el área bajo monitoreo en condiciones de bajo contenido de agua en el suelo, sin embargo, la relación diverge de ese comportamiento en condiciones de contenidos de humedad entre 0.4 a 0.5 cm^3 cm^{-3}, por lo que el esquema de desagregación propuesto en este estudio no dio resultados adecuados en esas condiciones específicas.

Además, en agregaciones temporales gruesas (aproximadamente promedios semanales) los coeficientes de correlación cuantílicos entre las dos series de tiempo son en promedio 0.98, independientemente de la estación del año, por lo que el producto SMAP-L3-E desagregado mediante *random forest* explica de forma casi perfecta las mediciones *in situ* en agregaciones de tiempo semanales.

8.3.1. Sugerencias.

- Ampliar el periodo de monitoreo de la humedad del suelo y considerar otros usos de suelo.
- Realizar muestreo espacial de con alta densidad por pixel del producto SMAPL3E y escalar dicha información a agregaciones espaciales más coherentes con el soporte de los pixeles del producto SMAPL3E para robustecer la validación.
- Optimizar los hiper-parámetros del *random forest* mediante validación cruzada anidada.

- Aplicar otras técnicas de aprendizaje automático como aprendizaje profundo o XGBoost y comparar los resultados con los obtenidos en este estudio, además comparar con un modelo *baseline* como regresión lineal múltiple.
- Ampliar el dominio de predicción para estimar la humedad del suelo a profundidades mayores a 10 cm (zona radicular) mediante la aplicación de modelos de flujo insaturado.
- Fusionar los productos de varias plataformas de teledetección de la humedad del suelo (p. ej. SMOS) para robustecer la metodología propuesta en este estudio.
- Usar otras covariables en la construcción de los modelos, entre ellas covariables que describan la dinámica y estructura de la vegetación (p. ej. NDVI).
- Realizar un análisis de la relación entre la humedad del suelo y la precipitación de la estación K´ayra.
- Utilizar el producto SMAP-L3-E desagregado en esta investigación como covariable en estudios de rendimiento de cultivos a nivel de subcuenca y probar la significancia del producto desagregado en la predicción de rendimientos en condiciones de cultivo por secano.
- Asimilar las series de tiempo del producto SMAP-L3-E desagregado en esta investigación como condiciones iniciales o variable de estado en modelos hidrológicos distribuidos y evaluar si la data de humedad del suelo mejora los resultados de simulaciones hidrológicas como forma de validación adicional.

IX. BIBLIOGRAFÍA

Abbaszadeh, P., Moradkhani, H., & Zhan, X. (2019). Downscaling SMAP Radiometer Soil Moisture Over the CONUS Using an Ensemble Learning Method. Water Resources Research, 55(1), 324-344. https://doi.org/10.1029/2018WR023354

Adab, H., Morbidelli, R., Saltalippi, C., Moradian, M., & Ghalhari, G. A. F. (2020). Machine Learning to Estimate Surface Soil Moisture from Remote Sensing Data. Water, 12(11), 3223. https://doi.org/10.3390/w12113223

Atkinson, P. M. (2013). Downscaling in remote sensing. International Journal of Applied Earth Observation and Geoinformation, 22, 106-114. https://doi.org/10.1016/j.jag.2012.04.012

Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H., & Tuller, M. (2019). Ground, Proximal, and Satellite Remote Sensing of Soil Moisture. Reviews of Geophysics, 57(2), 530-616. https://doi.org/10.1029/2018RG000618

Bai, J., Cui, Q., Zhang, W., & Meng, L. (2019). An Approach for Downscaling SMAP Soil Moisture by Combining Sentinel-1 SAR and MODIS Data. Remote Sensing, 11(23), 2736. https://doi.org/10.3390/rs11232736

Beck, H. E., Pan, M., Miralles, D. G., Reichle, R. H., Dorigo, W. A., Hahn, S., Sheffield, J., Karthikeyan, L., Balsamo, G., Parinussa, R. M., van Dijk, A. I. J. M., Du, J., Kimball, J. S., Vergopolan, N., & Wood, E. F. (2021). Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors. Hydrology and Earth System Sciences, 25(1), 17-40. https://doi.org/10.5194/hess-25-17-2021

Beven, K., & Freer, J. (2001). A dynamic TOPMODEL. Hydrological Processes, 15(10), 1993-2011. https://doi.org/10.1002/hyp.252

Biecek, P., & Burzykowski, T. (2021). Explanatory model analysis: Explore, explain, and examine predictive models (1.a ed.). CRC Press.

Bivand, R., Keitt, T., & Rowlingson, B. (2021). rgdal: Bindings for the «Geospatial» Data Abstraction Library 1.5-23. http://rgdal.r-forge.r-project.org, https://gdal.org,

Breiman, L. (1999). Random Forests. University of California Berkeley, Dept. of Statistics.

Breiman, L., & Cutler, A. (2003). Random Forests for Scientific Discovery.

Breiman, L., Friedman, H. J., Olshen, R., & Stone, C. (Eds.). (1998). Classification and regression trees (Repr). Chapman & Hall [u.a.].

Brocca, L., Ciabatta, L., Massari, C., Camici, S., & Tarpanelli, A. (2017). Soil Moisture for Hydrological Applications: Open Questions and New Opportunities. Water, 9(2), 140. https://doi.org/10.3390/w9020140

Brocca, L., Morbidelli, R., Melone, F., & Moramarco, T. (2007). Soil moisture spatial variability in experimental areas of central Italy. Journal of Hydrology, 333(2-4), 356-373. https://doi.org/10.1016/j.jhydrol.2006.09.004

Brus, D. J. (2019). Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma, 338, 464-480. https://doi.org/10.1016/j.geoderma.2018.07.036

Chan, S., Bindlish, R., O'Neill, P. E., Jackson, T., Njoku, E. G., Dunbar, S., Chaubell, J., Piepmeier, J. R., Entekhabi, D., Colliander, A., Chen, F., Cosh, M., Caldwell, T., Walker,

J., & Berg, A. (2018). Development and assessment of the SMAP enhanced passive soil moisture product. Remote Sensing of the Environment.

Chaubell, J. (2016). Algorithm Theoretical Basis Document SMAP L1B Enhancement Radiometer Brightness Temperature Data Product. 24.

Chen, S., She, D., Zhang, L., Guo, M., & Liu, X. (2019). Spatial Downscaling Methods of Soil Moisture Based on Multisource Remote Sensing Data and Its Application. Water, 11(7), 1401. https://doi.org/10.3390/w11071401

Choi, J.-E., & Shin, D. W. (2022). Quantile correlation coefficient: A new tail dependence measure. Statistical Papers, 63(4), 1075-1104. https://doi.org/10.1007/s00362-021-01268-7

Chue Hong, N. (2019). How to cite software: Current best practice. https://doi.org/10.5281/ZENODO.2842910

Colliander, A., Jackson, T. J., Bindlish, R., Chan, S., Das, N., Kim, S. B., Cosh, M. H., Dunbar, R. S., Dang, L., Pashaian, L., Asanuma, J., Aida, K., Berg, A., Rowlandson, T., Bosch, D., Caldwell, T., Caylor, K., Goodrich, D., al Jassar, H., … Yueh, S. (2017). Validation of SMAP surface soil moisture products with core validation sites. Remote Sensing of Environment, 191, 215-231. https://doi.org/10.1016/j.rse.2017.01.021

Cooper, D. (2016). Soil water measurement: A practical handbook. John Wiley & Sons.

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D., & Walker, J. P. (2012). Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products: UPSCALING SOIL MOISTURE. Reviews of Geophysics, 50(2), Article 2.

Das, N. N. (2019). SMAP-Sentinel L2 Radar/Radiometer Soil Moisture (Active/Passive) Data Products: L2_SM_SP. 62.

de Gruijter, J., Brus, D., Bierkens, M., & Knotters, M. (2006). Sampling for Natural Resource Monitoring (1.a ed.). Springer.

de Sousa, L. M., Poggio, L., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Riberio, E., & Rossiter, D. (2020). SoilGrids 2.0: Producing quality-assessed soil information for theglobe [Preprint]. Soils and the natural environment. https://doi.org/10.5194/soil-2020-65

Dorigo, W., & de Jeu, R. (2016). Satellite soil moisture for advancing our understanding of earth system processes and climate change. International Journal of Applied Earth Observation and Geoinformation, 48, 1-4. https://doi.org/10.1016/j.jag.2016.02.007

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., … Van Zyl, J. (2010). The Soil Moisture Active Passive (SMAP) Mission. Proceedings of the IEEE, 98(5), 704-716. https://doi.org/10.1109/JPROC.2010.2043918

Famiglietti, J. S., Ryu, D., Berg, A. A., Rodell, M., & Jackson, T. J. (2008). Field observations of soil moisture variability across scales: SOIL MOISTURE VARIABILITY ACROSS SCALES. Water Resources Research, 44(1), Article 1. https://doi.org/10.1029/2006WR005804

Fang, B., Lakshmi, V., Bindlish, R., & Jackson, T. J. (2018). Downscaling of SMAP Soil Moisture Using Land Surface Temperature and Vegetation Data. Vadose Zone Journal, 17(1), 170198. https://doi.org/10.2136/vzj2017.11.0198

Fang, K., Pan, M., & Shen, C. (2019). The Value of SMAP for Long-Term Soil Moisture Estimation With the Help of Deep Learning. IEEE Transactions on Geoscience and Remote Sensing, 57(4), 2221-2233. https://doi.org/10.1109/TGRS.2018.2872131

Gaskin, G. J., & Miller, J. D. (1996). Measurement of Soil Water Content Using a Simplified Impedance Measuring Technique. Journal of Agricultural Engineering Research, 63(2), 153-159. https://doi.org/10.1006/jaer.1996.0017

Grossman, R. B., & Reinsch, T. G. (2002). 2.1 Bulk Density and Linear Extensibility. En Methods of Soil Analysis: Part IV .Physical Methods (Soil Science Society of America, p. 28).

Gruber, A., De Lannoy, G., Albergel, C., Al-Yaari, A., Brocca, L., Calvet, J.-C., Colliander, A., Cosh, M., Crow, W., Dorigo, W., Draper, C., Hirschi, M., Kerr, Y., Konings, A., Lahoz, W., McColl, K., Montzka, C., Muñoz-Sabater, J., Peng, J., … Wagner, W. (2020). Validation practices for satellite soil moisture retrievals: What are (the) errors? Remote Sensing of Environment, 244, 111806. https://doi.org/10.1016/j.rse.2020.111806

Gruber, S., & Peckham, S. (2009). Land-Surface Parameters and Objects in Hydrology. En Geomorphometry: Concepts, software, applications (1st ed). Elsevier.

Guevara, M., & Vargas, R. (2019). Downscaling satellite soil moisture using geomorphometry and machine learning. PLOS ONE, 14(9), e0219639. https://doi.org/10.1371/journal.pone.0219639

Gupta, S., Lehmann, P., Bonetti, S., Papritz, A., & Or, D. (2021). Global Prediction of Soil Saturated Hydraulic Conductivity Using Random Forest in a Covariate-Based GeoTransfer Function (CoGTF) Framework. Journal of Advances in Modeling Earth Systems, 13(4). https://doi.org/10.1029/2020MS002242

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hengl, T., & MacMillan, R. A. (2019). Predictive Soil Mapping with R. OpenGeoHub foundation.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE, 12(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, 6, e5518. https://doi.org/10.7717/peerj.5518

Hengl, T., Reuter, H. I., & Institute for Environment and Sustainability (European Commission. Joint Research Centre) (Eds.). (2009). Geomorphometry: Concepts, software, applications (1st ed). Elsevier.

Hernandez-Sanchez, J. C., Monsivais-Huertero, A., Judge, J., & Carlos Jimenez-Escalona, J. (2020). Comparison of SMAP Retrieval Soil Moisture Level 2 Product with in Situ Measurements Over Corn Fields in Central Mexico. IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 4727-4730. https://doi.org/10.1109/IGARSS39084.2020.9324106

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in

digital soil mapping. Geoderma, 265, 62-77. https://doi.org/10.1016/j.geoderma.2015.11.014

Heuvelink, G. B. M., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G. F., & Sanderman, J. (2021). Machine learning in space and time for modelling soil organic carbon change. European Journal of Soil Science, 72(4), 1607-1623. https://doi.org/10.1111/ejss.12998

Hillel, D. (2004). Introduction to environmental soil physics. Elsevier Academic Press. http://www.123library.org/book_details/?id=44290

Hijmans, R. J., & van Etten, J. (2012). raster: Geographic data analysis and modeling R package version 2.0-41. http://CRAN. 764 R-project.org/package=raster.

Husson, F., Lê, S., & Josse, J. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), Article 1. https://doi.org/10.18637/jss.v025.i01

Husson, F., Lê, S., & Pagès, J. (2017). Exploratory Multivariate Analysis by Example Using R. CRC Press, 263.

Hvitfeldt, E., Pedersen, T. L., & Benesty, M. (2022). Lime: Local Interpretable Model-Agnostic Explanations. R package version 0.4.1. http://arxiv.org/abs/1602.04938

Im, J., Park, S., Rhee, J., Baik, J., & Choi, M. (2016). Downscaling of AMSR-E soil moisture with MODIS products using machine learning approaches. Environmental Earth Sciences, 75(15), 1120. https://doi.org/10.1007/s12665-016-5917-6

Imfeld, N., Sedlmeier, K., Gubler, S., Correa Marrou, K., Davila, C. P., Huerta, A., Lavado-Casimiro, W., Rohrer, M., Scherrer, S. C., & Schwierz, C. (2021). A combined view on

precipitation and temperature climatology and trends in the southern Andes of Peru. International Journal of Climatology, 41(1), 679-698. https://doi.org/10.1002/joc.6645

Jackson, T. J. (1993). III. Measuring surface soil moisture using passive microwave remote sensing. Hydrological Processes, 7(2), 139-152. https://doi.org/10.1002/hyp.3360070205

Jackson, T. J., Cosh, M. H., Bindlish, R., Starks, P. J., Bosch, D. D., Seyfried, M., Goodrich, D. C., Moran, M. S., & Du, J. (2010). Validation of Advanced Microwave Scanning Radiometer Soil Moisture Products. IEEE Transactions on Geoscience and Remote Sensing, 48(12), 4256-4272. https://doi.org/10.1109/TGRS.2010.2051035

Jackson, T. J., & Schmugge, T. J. (1991). Vegetation effects on the microwave emission of soils. Remote Sensing of Environment, 36(3), 203-212. https://doi.org/10.1016/0034-4257(91)90057-D

Jacobs, J. (2004). SMEX02: Field scale variability, time stability and similarity of soil moisture. Remote Sensing of Environment, 92(4), 436-446. https://doi.org/10.1016/j.rse.2004.02.017

James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). An introduction to statistical learning: With applications in R. Springer.

Kerr, Y. H. (2007). Soil moisture from space: Where are we? Hydrogeology Journal, 15(1), 117-120. https://doi.org/10.1007/s10040-006-0095-3

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling, 81, 401-418. https://doi.org/10.1016/j.apm.2019.12.016

Koenker, R., Chernozhukov, V., He, X., & Peng, L. (Eds.). (2017). Handbook of Quantile Regression (1.a ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315120256

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kutner, M. H. (2004). Applied linear statistical models. McGraw-Hill Irwin.

Li, X., McCarty, G. W., Du, L., & Lee, S. (2020). Use of Topographic Models for Mapping Soil Properties and Processes. Soil Systems, 4(2), 32. https://doi.org/10.3390/soilsystems4020032

Mao, H., Kathuria, D., Duffield, N., & Mohanty, B. P. (2019). Gap Filling of High-Resolution Soil Moisture for SMAP/Sentinel-1: A Two-Layer Machine Learning-Based Framework. Water Resources Research, 55(8), 6986-7009. https://doi.org/10.1029/2019WR024902

Mohanty, B. P., Cosh, M. H., Lakshmi, V., & Montzka, C. (2017). Soil Moisture Remote Sensing: State-of-the-Science. Vadose Zone Journal, 16(1), vzj2016.10.0105. https://doi.org/10.2136/vzj2016.10.0105

Mohanty, B. P., & Skaggs, T. H. (2001). Spatio-temporal evolution and time-stable characteristics of soil moisture within remote sensing footprints with varying soil, slope, and vegetation. Advances in Water Resources, 24(9-10), 1051-1067. https://doi.org/10.1016/S0309-1708(01)00034-3

Montzka, C. M., Cosh, B., Bayat, A., Bitar, A., Berg, R., Bindlish, H. R., Bogena, J. D., Bolten, Cabot, F., Caldwell, T., Chan, S., Colliander, A., Crow, W., Das, N., De Lannoy, Dorigo,

Evett, Gruber, A., Jagdhuber, … O´Neill, P. (2020). Soil Moisture Product Validation Good Practices Protocol. https://doi.org/10.5067/DOC/CEOSWGCV/LPV/SM.001

NASA. (2014). SMAP Handbook Soil Moisture Active Passive. Jet Propulsion Laboratory California Institute of Technology,.

Neteler, M., & Mitasova, H. (2008). Open source GIS: A GRASS GIS approach (3. ed). Springer.

Njoku, E. G., & Entekhabi, D. (1996). Passive microwave remote sensing of soil moisture. Journal of Hydrology, 184, 101-129.

O'Neill, P., Bindlish, R., Chan, S., Chaubell, J., Njoku, E., & Jackson, T. (2020b). Algorithm Theoretical Basis Document Level 2 & 3 Soil Moisture (Passive) Data Products. 100.

Pebesma, E., & Bivand, R. S. (2005). Classes and Methods for Spatial Data: The sp Package. 21.

Peng, J., Loew, A., Merlin, O., & Verhoest, N. E. C. (2017). A review of spatial downscaling of satellite remotely sensed soil moisture: Downscale Satellite-Based Soil Moisture. Reviews of Geophysics, 55(2), 341-366. https://doi.org/10.1002/2016RG000543

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery, 9(3). https://doi.org/10.1002/widm.1301

Qin, C., Pei, T., Li, B., Yang, L., & Zhou, C. (2007). An adaptive approach to selecting a flowpartition exponent for a multiple flow direction algorithm. International Journal of Geographical Information Science, 443-458. https://doi.org/10.1080/13658810601073240

Quinn, P. F., Beven, K. J., & Lamb, R. (1995). The in(a/tan/β) index: How to calculate it and how to use it within the topmodel framework. Hydrological Processes, 9(2), 161-182. https://doi.org/10.1002/hyp.3360090204

R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/

Raduła, M. W., Szymura, T. H., & Szymura, M. (2018). Topographic wetness index explains soil moisture better than bioindication with Ellenberg's indicator values. Ecological Indicators. Ecological Indicators.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). «Why Should I Trust You?»: Explaining the Predictions of Any Classifier (arXiv:1602.04938). arXiv. http://arxiv.org/abs/1602.04938

Schmugge, T. J., Kustas, W. P., Ritchie, J. C., Jackson, T. J., & Rango, A. (2002). Remote sensing in hydrology. Advances in Water Resources, 19.

Schratz, P., Becker, M., Lang, M., & Brenning, A. (2021). Mlr3spatiotempcv: Spatiotemporal resampling methods for machine learning in R. ArXiv:2110.12674 [Cs, Stat]. http://arxiv.org/abs/2110.12674

Sociedad Americana de Ciencia de suelos (Ed.). (2008). Glossary of soil science terms. Soil Science Society of America.

Srivastava, P. K., Han, D., Ramirez, M. R., & Islam, T. (2013). Machine Learning Techniques for Downscaling SMOS Satellite Soil Moisture Using MODIS Land Surface Temperature for Hydrological Application. Water Resources Management, 27(8), 3127-3144. https://doi.org/10.1007/s11269-013-0337-9

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tindal, J. A., & Kunkel, J. R. (1999). Unsaturated Zone Hydrology for Scientists and Engineers (1.a ed.). Prentice Hall.

Topp, G., & Ferré, P. A. T. (2018). 3.1 Water Content. En J. H. Dane & G. Topp (Eds.), SSSA Book Series (pp. 417-545). Soil Science Society of America. https://doi.org/10.2136/sssabookser5.4.c19

Tu, L. (2019). Downscaling SMAP Soil Moisture Data Using MODIS Data [Master of Science]. Louisiana State University.

Van der Meer, F. (2012). Remote-sensing image analysis and geostatistics. International Journal of Remote Sensing, 33(18), 5644-5676. https://doi.org/10.1080/01431161.2012.666363

van Genuchten, M. Th. (1980). A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils. Soil Science Society of America Journal, 44(5), 892-898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

Vergopolan, N., Sheffield, J., Chaney, N. W., Pan, M., Beck, H. E., Ferguson, C. R., Torres-Rojas, L., Eigenbrod, F., Crow, W., & Wood, E. F. (2022). High-Resolution Soil Moisture Data Reveal Complex Multi-Scale Spatial Variability Across the United States. Geophysical Research Letters, 49(15). https://doi.org/10.1029/2022GL098586

Wackernagel, H. (2010). Multivariate Geostatistics. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-05294-5

Wakigari, S. A., & Leconte, R. (2022). Enhancing Spatial Resolution of SMAP Soil Moisture Products through Spatial Downscaling over a Large Watershed: A Case Study for the Susquehanna River Basin in the Northeastern United States. Remote Sensing, 14(3), 776. https://doi.org/10.3390/rs14030776

Wang, C., Zuo, Q., & Zhang, R. (2008). Estimating the necessary sampling size of surface soil moisture at different scales using a random combination method. Journal of Hydrology, 352(3-4), 309-321. https://doi.org/10.1016/j.jhydrol.2008.01.011

Warner, D. L., Guevara, M., Callahan, J., & Vargas, R. (2021). Downscaling satellite soil moisture for landscape applications: A case study in Delaware, USA. Journal of Hydrology: Regional Studies, 38, 100946. https://doi.org/10.1016/j.ejrh.2021.100946

Weil, R. R., & Brady, N. C. (2017). The nature and properties of soils (Fifteenth edition, global edition). Pearson Prentice Hall.

Western, A. W., & Blöschl, G. (1999). On the spatial scaling of soil moisture. Journal of Hydrology, 217(3-4), 203-224. https://doi.org/10.1016/S0022-1694(98)00232-7

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data. Journal of Statistical Software, 77(1). https://doi.org/10.18637/jss.v077.i01

Xu, C., Ke, J., Zhao, X., & Zhao, X. (2020). Multiscale Quantile Correlation Coefficient: Measuring Tail Dependence of Financial Time Series. Sustainability, 12(12), 4908. https://doi.org/10.3390/su12124908

Xu, Y. (2019). Mapping Soil Moisture from Remotely Sensed and In-situ Data with Statistical Methods (p. 100) [Doctoral Dissertation]. Louisiana State University and Agricultural and Mechanical College.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. Water Resources Research, 55(6), 5053-5073. https://doi.org/10.1029/2019WR024873

Zappa, L., Forkel, M., Xaver, A., & Dorigo, W. (2019). Deriving Field Scale Soil Moisture from Satellite Observations and Ground Measurements in a Hilly Agricultural Region. Remote Sensing, 11(22), 2596. https://doi.org/10.3390/rs11222596

Zhang, D., & Zhou, G. (2016). Estimation of Soil Moisture from Optical and Thermal Remote Sensing: A Review. Sensors, 16(8), 1308. https://doi.org/10.3390/s16081308

Zhang, H., Yang, R., Guo, S., & Li, Q. (2020). Modeling fertilization impacts on nitrate leaching and groundwater contamination with HYDRUS-1D and MT3DMS. Paddy and Water Environment, 18(3), 481-498. https://doi.org/10.1007/s10333-020-00796-6

Zhao, W., Sánchez, N., Lu, H., & Li, A. (2018). A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression. Journal of Hydrology, 563, 1009-1024. https://doi.org/10.1016/j.jhydrol.2018.06.081

X. ANEXOS.

Anexo 1: Implementación de los algoritmos en R.

Los códigos realizados en R para esta tesis pueden encontrarse en el siguiente repositorio de GitHub:

https://github.com/kundun14/soil_moisture_SMAP_machine_learning

Anexo 2: Ubicación e instalación de sensores de humedad del suelo.