USE OF SENTINEL-1, -2 AND -3 SATELLITES FOR WATER CYCLE STUDIES. DEVELOPMENT OF LOW-COST TOOLS FOR CAPACITY BUILDING AND TRANSFER KNOWLEDGE

University of Valencia. Faculty of Physics

Department of Earth Physics and Thermodynamics

Doctorate in Remote Sensing

Doctorate Thesis:

Pierre Ferreira do Prado

Supervisors:

Dr. Ernesto López Baeza

Dra. Iolanda C. S. Duarte

Valencia, January, 2024

VNIVERSITAT DE VALÈNCIA Facultat de Física

Departamento de Física de la Tierra y Termodinámica
Grupo de Teledetección de Medio Ambiente (UV-ERS)

ufscar

Departamento de Biología/Sorocaba
Universidad de San Carlos/Brasil

18 de Enero de 2024

La Tesis Doctoral del doctorando **Pierre Ferreira do Prado** desarrolla un avance significativo en la capacidad de la *Valencia Anchor Station* (VAS) de validación de datos de satélite, en este caso, en relación con el contenido en humedad del suelo. Pierre ha conseguido implementar una metodología basada en modelos de *machine learning* para la predicción de alta resolución (120 m x 120 m) de la humedad del suelo utilizando datos del satélite Copernicus Sentinel-1. La modelización se ha desarrollado en dos etapas, la primera centrada en la predicción de este parámetro alrededor de la propia estación meteorológica de la VAS (se denomina VAS1 en la Tesis) y en la segunda etapa se emplean datos MODIS/TVDI y función de transferencia del proyecto *Ground-Based Observations for Validation (GBOV) of Copernicus Global Land Products* del CE *Joint Research Center* para así extrapolar la predicción de VAS1 a otras zonas de la zona de control de 10 km x10 km de la VAS. Las buenas correlaciones entre valores observados y valores simulados/de predicción abren la vía a una modelización más ambiciosa utilizando modelos generativos y la totalidad de los conocimientos acumulados sobre la VAS.

Pierre ha demostrado en todo momento una capacidad auto-organizativa de trabajo riguroso por objetivos con una base robusta de conocimientos en un contexto de colaboración internacional Brasil-España, lo cual permitiría extender esta metodología a otros países en vías de desarrollo, particularmente Brasil, sin necesidad de invertir costes en equipamiento, donde pueden existir restricciones de tipo económico.

ERNESTO|
LOPEZ|BAEZA
Firmado digitalmente por ERNESTO|LOPEZ| BAEZA
Fecha: 2024.01.19 22:44:55 +01'00'

Fdo.: Ernesto López Baeza

gov.br
Documento assinado digitalmente
IOLANDA CRISTINA SILVEIRA DUARTE
Data: 19/01/2024 17:40:50-0300
Verifique em https://validar.iti.gov.br

Fdo.: Iolanda C. Silveira Duarte

ACKNOWLEDGMENT

For Prof. Dr. Ernesto López Baeza and Prof. Dr. Iolanda C. S. Duarte, respectively from the University of Valencia, Spain, and from the University of São Paulo State/Brazil and Federal University of São Carlos/Brazil. Taken together, these two great persons and these great institutions have provided me with everything I needed to experience an extraordinary and unforgettable life and scientific journey, for which I take this opportunity to express my heartful gratitude.

This work has been carried out mainly within the framework of the Valencian Anchor Station, to which the successive members of the Satellite Climatology Group have devoted so much effort and time. On the one hand, I am grateful to all of them for the fact that I have been able to benefit, directly or indirectly, from their previous work and, on the other hand, I am honoured to be able to contribute modestly, with this Doctoral Thesis, to the growth of knowledge of this scientific research instrument.

GENERAL INDEX

ACRONYMS

ANN	Artificial Neural Network
ASCAT	Advanced SCATterometer
autoML	Automated Machine Learning
ATBD	Algorithm Theoretical Basis Document
C-band	Frequencies of 4-8Hz od electromagnetic spectrum
CAROLS	Cooperative Airborne Radiometer for Ocean and Land Studies
CERES	Clouds and the Earth's Radiant Energy System
EC	European Commission
ED50	European Datum 1950
EO	Earth Observation
EPS	EUMETSAT Polar System
ESA	European Space Agency
ESSP	Earth System Science Pathfinder
FAM	Fuzzy ARTMAP
GCS	Climatology from Satellites Group
GBOV	Ground-based Observations for Validation of Copernicus Global Land Products
GRD	Ground Range Detected
HH	Horizontal – Horizontal polarization
HV	Horizontal – Vertical polarization
IEM	Integral Equation Model
INFOIGME	Information of Spanish Geological Survey (IGME)
ISRO	Indian Space Research Organization
LAI	Leaf Area Index
M1SV	Module-1to predict SSM at the VAS1
M2GE	Method for Geographical Extrapolation
MELBEX	Mediterranean Ecosystem L-Band characterization Experiment
ML	Machine Learning
MODIS	Moderate-Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NISAR	NASA-ISRO Synthetic Aperture Radar
NDVI	Normalized difference vegetation index
OM	Organic Matter
PC	Principal Components

RMSE	Root Mean Squared Error
RTC	Radiometric Terrain Correction
SAR	Synthetic Aperture Radar
SMAP	Soil Moisture Active Passive
SMOS	Soil Moisture and Ocean Salinity
SSM	Surface Soil Moisture
SVRT	SMOS Validation Retrieval Campaign
SVM	Support Vector Machine
TDR	Time-Domain Reflectometry
TVDI	Temperature Vegetation Dryness Index
UTM	Universal Transverse Mercator
VAS	Valencia Anchor Station
VAS1	ThetaProbe at the VAS met station (bare soil)
VAS2	ThetaProbe at MELBEX_II_177_01 (tbc)
VAS3	ThetaProbe at MELBEX_II_177_02 (tbc)
VAS4	ThetaProbe at MELBEX_II_178 (tbc)
VAS5	ThetaProbe at MELBEX_I (matorral)
VAS6	ThetaProbe at La Cubera (vines)
VAS7	ThetaProbe at Nicolas_M1 (matorral)
VAS8	ThetaProbe at Nicolas_T1 (almond and olive tress)
VH	Vertical – Horizontal polarization
VV	Vertical – Vertical polarization

SYMBOLS

Ia	Weighted cross-polarized σ°
R	Pearson correlation coefficient
ε	Electrical permittivity (C/m^2)
σ°	Radar backscatter signal
°	Latitude and Longitude grade
®	Registered mark
x	Box-plot average

FIGURE CAPTIONS

TABLE TITLES

RESUMEN

La urgencia de determinar el contenido en humedad superficial del suelo (SSM) en todo el mundo, especialmente en los suelos agrícolas, está bien establecida. La misión Copernicus Sentinel-1 de la Comisión Europea (CE) juntamente con la Agencia Espacial Europea (ESA) ofrece un radar de apertura sintética (SAR) que, junto con métodos de *machine learning*, contribuye eficazmente a esta demanda tecnológica. Este trabajo tiene como objetivo explorar la capacidad de los datos de humedad del suelo disponibles en la Valencia Anchor Station (VAS) en Valencia, España. También se utilizaron herramientas de MATLAB® para investigar la complejidad de la relación entre la SSM y el coeficiente de retrodispersión lineal (polarizaciones de σ°, VV y VH) de Sentinel-1 con corrección radiométrica del terreno (RTC) codificada en los valores de píxel de las imágenes de esta zona de estudio utilizando métodos de aprendizaje automático supervisado.

El trabajo está estructurado de la siguiente manera: aparte de la Introducción y Objetivos (Sección 1), las características de las zonas estudiadas, los conjuntos de datos terrestres y satelitales, y los métodos para la implementación de las redes neuronales artificiales se describen en la Sección 2 sobre Materiales y Métodos, mientras que la Sección 3 sobre Resultados y Discusión ofrece una descripción y discusión de los resultados obtenidos. En la Sección 4 se presentan las conclusiones del estudio y se ofrecen algunas perspectivas sobre esta rama de investigación.

INTRODUCCIÓN Y OBJETIVOS

La humedad del suelo es un factor crítico para el crecimiento de las plantas, ya que afecta la disponibilidad de agua y nutrientes para las raíces. La medición de la humedad del suelo es importante para la agricultura, la jardinería, la silvicultura y otras aplicaciones. Por lo general, se mide utilizando instrumentos que se pueden organizar en dos conjuntos: 1) métodos directos que miden directamente el contenido de agua del suelo, como la difracción de rayos X, los métodos gravimétricos y de microondas y 2) métodos indirectos que miden algunos parámetros relacionados con el contenido de agua del suelo como, por ejemplo: resistividad, tensiómetro y conductividad eléctrica (Montzka et al, 2020).

Singh et al. (2023) describieron la humedad del suelo como el almacenamiento de agua en los poros del suelo durante un cierto intervalo de tiempo. Esta cantidad de agua, en total, es muy baja en comparación con el agua total disponible en la Tierra. A pesar de ello, la interfase aire-suelo es crucial para los procesos biológicos e hidrológicos. El estudio de esta parcela de agua es clave para la investigación agrícola, los incendios forestales, los estudios de inundaciones, las sequías y los estudios climáticos, entre otros. Actualmente, se dispone de imágenes satelitales como Sentinel-1, Sentinel-2 y Radarsat-2 para cubrir la superficie de la Tierra. Según estos autores, a escala global, ya se ha demostrado (a favor) que los métodos ópticos y de microondas son importantes para la difícil tarea de seguimiento de la humedad superficial del suelo a una escala tan desafiante. Teniendo en cuenta el potencial de la misión NASA-ISRO Synthetic Aperture Radar (NISAR) junto con las mejoras en los métodos de aprendizaje automático, como el AutoML y los algoritmos basados en la física, estos autores afirman que existe una importante vía por explorar en los próximos años en el campo de la teledetección SSM.

También señalan que ESA y NASA han lanzado misiones respectivas para proporcionar humedad del suelo, como Soil Moisture and Ocean Salinity (SMOS), de ESA y Soil Moisture Active and Passive de NASA, esta última en el marco de las misiones Earth System Science Pathfinder (ESSP). Estos productos SSM tienen resoluciones espaciales que varían de 1 ~ 50 km y tiempos de revisita que oscilan entre 1 y 8 días, pero todavía hay algunas lagunas en su cobertura, especialmente relacionadas con la vegetación densa y la topografía compleja. Estos autores también destacaron que la capacidad de penetración en la capa superficial del suelo, la sensibilidad a las propiedades dieléctricas, la robustez a la cobertura de nubes y la disponibilidad diurna y nocturna, y la disponibilidad de material son características clave que impulsan al uso de la teledetección por microondas en esta tarea de estimar la humedad del suelo. Se han propuesto muchos tipos de modelos de retrodispersión, semiempíricos, empíricos y teóricos, que se basan en el gradiente de permitividad (ε), que oscila entre ~80 (agua) y ~2 (suelo seco), y basados en imágenes de microondas de polarización cuádruple (VV, VH, HH y HV). Teniendo en cuenta que este tipo imágenes no están disponibles para algunas regiones, se suelen utilizar métodos de aprendizaje automático que utilizan otros datos y productos como índices de vegetación, topografía y otros datos auxiliares, además de señales de retrodispersión con otras características de polarización.

Ali et al. (2015) evaluaron la recuperación de SSM utilizando métodos de inversión de redes neuronales artificiales (RNA). Estos autores clasificaron el radar de apertura sintética como un tipo de método de teledetección por microondas complementario al uso de radiómetros pasivos y dispersómetros. Según estos autores, las mediciones terrestres de SSM son escasas semanalmente/diariamente, y existe una clara necesidad de esta información a pequeña y gran escala que permita obtener información sensible para la gestión del agua en las explotaciones agrícolas, especialmente en caso de sequías e inundaciones, permitiendo una toma de decisiones fiable y rápida.

La recuperación de SSM a partir de sensores remotos de microondas se basa en su relación con las propiedades dieléctricas del suelo, lo que de hecho es un problema desafiante y complejo debido a la combinación de diferentes características del suelo, como la humedad del suelo, su rugosidad y textura y la cubierta vegetal, todo lo cual afecta una respuesta electromagnética única dada capturada por un sensor (Ali et al., 2015). Esta relación entre la humedad del suelo/propiedades dieléctricas del suelo y la emisión de radiación de microondas puede validarse localmente a partir de los datos proporcionados por instrumentos de medición terrestres distribuidos estratégicamente en varias zonas, como la VAS en el este de España.

La misión Copernicus Sentinel-1 consta de dos satélites que comparten el mismo plano orbital, a saber, Sentinel-1A y Sentinel-1B. Ambos sistemas llevan un instrumento de radar de apertura sintética (SAR) de banda C en cuatro modos que difieren en resolución espacial (con un límite inferior de 5 m) y cobertura (hasta 400 km). También ofrecen una entrega rápida del producto, doble polarización y un corto tiempo de revisión. Esta misión se centra, fundamentalmente, en aplicaciones sobre supervisión del medio ambiente y seguridad. La capacidad del SAR para funcionar de forma independiente durante el día y la noche y en condiciones nubladas pone de manifiesto el potencial de utilizar los datos de Sentinel-1 para vigilar con frecuencia zonas amplias.

La investigación de la humedad superficial apoyada por los productos Sentinel-1 ha sido y sigue siendo un esfuerzo científico activo y dinámico. Con frecuencia, el énfasis de estas investigaciones está en cómo considerar de manera óptima los efectos de la vegetación y la rugosidad de la superficie en la señal de retrodispersión $\sigma°$ e, idealmente, en cómo extender este recurso de libre disponibilidad ofrecido por las políticas espaciales de Europa para subsidiar u optimizar las estrategias de riego inteligente en la agricultura en todo el mundo, especialmente en países menos desarrollados como Brasil y otros (Beale et al., 2019).

Entre los algoritmos más utilizados para este fin, según el análisis bibliométrico, se han utilizado *support vector machines*, redes neuronales y *random forests* soportados por medidas in situ de SSM. Sin embargo, el uso de sistemas de lógica difusa, *deep belief networks,* aprendizaje automático automatizado, árboles de decisión y regresión de procesos gaussianos, entre otros, se han mencionado con menor frecuencia (Singh et al., 2023).

Después de 9 años de operación continua, las características de la VAS se describieron por López-Baeza et al. (2009) en el marco de la validación de productos de baja resolución espacial de misiones de observación de la Tierra desde 2001. Ejemplos de estas misiones son CERES (Clouds and the Earth's Radiant Energy System), EPS (EUMETSAT Polar System) y SMOS, entre otras. Todas estas misiones comparten características como la necesidad de estar altamente instrumentadas y bien caracterizadas, grandes áreas con *footprints* cercanos a los 50 km × 50 km. La homogeneidad es una característica clave de esta zona en lo que respecta a las actividades de validación, como las que se llevaron a cabo durante la fase de comisionado de la misión SMOS, en relación con los productos de nivel 2, humedad del suelo y contenido de agua de vegetación. Durante este proyecto, se eligió una zona de control de 10 km × 10 km donde concentrar las operaciones de campo estableciendo sus características en cuanto a clima, tipo de suelo, litología, elevación, pendiente y condiciones de cobertura vegetal de las unidades fisio-hidrológicas a instrumentar y monitorizar con una red de estaciones de medición de SSM interconectadas mediante comunicación inalámbrica (Cano et al., 2007 y 2012). Más recientemente, López-Baeza et al. (2022a y 2022b) han podido describir algunos detalles clave sobre esta actividad a más largo plazo en la VAS.

Esta tesis tiene dos objetivos innovadores, ramificados e igualmente importantes: por un lado, 1) el objetivo estratégico es explorar la interrelación y el efecto neto de ese cuerpo ya disponible de conocimientos de la VAS (con mayor énfasis en VAS1, la piedra angular de la VAS). Además, todo esto se hizo principalmente, pero no exclusivamente, centrándose en la información disponible de una de las primeras sondas de SSM, la VAS1, operada continuamente casi desde su configuración inicial, situada junto al mástil principal que, naturalmente, está equipada con un cuerpo más sustancial de datos y conocimiento experto para permitir que los resultados de los modelos ML sean interpretados de acuerdo con un marco orientador e informado por la física, 2) el objetivo más inmediato es un estudio de viabilidad de una herramienta impulsada por Sentinel-1 para SSM sobre el área de control de 10 km x 10 km de la VAS. Para lograr 1), fue necesario (a) el uso cruzado (mezcla) de información de diferentes conjuntos de datos disponibles sobre la VAS procedentes de estudios anteriores y el análisis de las pruebas numéricas como la comparación entre las mediciones in situ y los datos de predicción. Y, para lograr 2), (b) se desplegaron algunos métodos de aprendizaje automático, con énfasis en las redes neuronales artificiales, para predecir SSM basándose en este conjunto heterogéneo de datos referenciados en (a) y en el propio conjunto de datos de valores de retrodispersión del radar ESA/Sentinel-1 para VAS1.

Por otra parte, para empezar a captar las múltiples potencialidades de este corpus de datos acumulados durante más de 15 años de funcionamiento de las mediciones SSM en la VAS y como posible referencia para los países de renta más baja, todas las proposiciones exploradas (incluidas las extrapolaciones geográficas) se han concebido de modo que puedan eventualmente ser fácilmente adoptadas por los técnicos de dichos países, que dependen más de la informática, a menudo más accesible en sus centros, que de la instalación y el mantenimiento de las redes de sensores SSM.

METODOLOGÍA

Esta tesis se beneficia de la disponibilidad de información in situ de humedad del suelo proporcionada por la VAS y de la cobertura del satélite Copernicus Sentinel-1 (SAR) sobre la comarca de La Plana de Utiel-Requena en la Comunidad Valenciana en España. Los datos se basan en el conjunto de datos AEMET de precipitación, conjuntos de datos basados en Sentinel-1 y conjuntos de datos de MODIS/TVDI y mediciones ThetaProbe de SSM, ambos relacionados con trabajos anteriores realizados en la VAS. Además, en este trabajo de tesis también se utilizaron métodos estadísticos multivariantes y de aprendizaje automático. Un resumen esquemático del flujo de trabajo de esta tesis se muestra en la Figura 1.

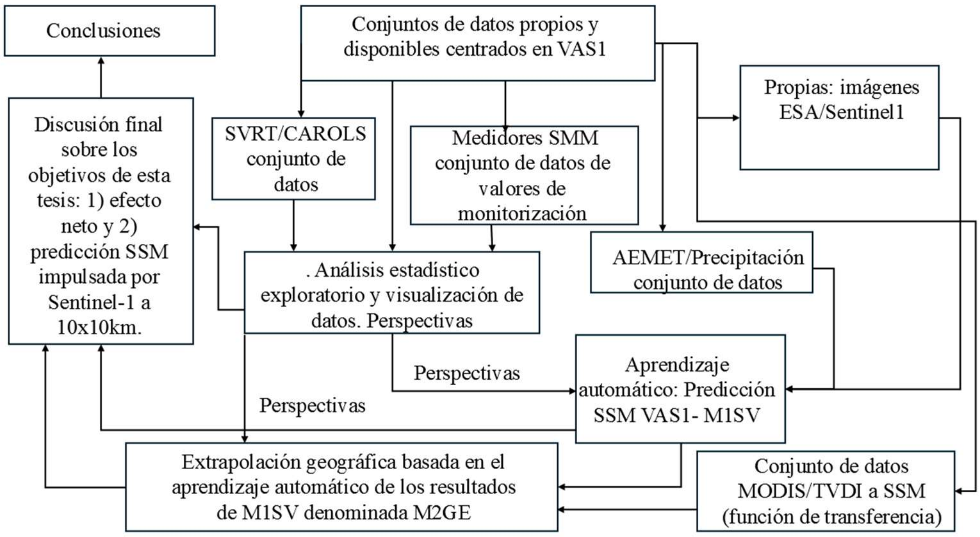

Figura 1 (duplicada en español): Resumen esquemático del flujo de trabajo de esta tesis

La zona experimental es la *Valencia Anchor Station* (VAS) (N 39.570740°, W −1.288176°) que se define como "*una estación meteorológica automática equipada con un gran número de instrumentos para llevar a cabo actividades de validación de datos y productos de teledetección de baja resolución espacial*" (Lopez-Baeza et al., en Bolle et al., 2006). Estos autores también informaron que la VAS se ubica en un ecosistema típicamente mediterráneo que abarca una gran superficie de viñedo y que también contiene otras especies, como matorrales, olivos y almendros, y pinares, distribuidos en un área razonablemente homogénea y plana de 50 × 50 km. El punto de monitorización VAS1 (Figura 2) es uno de los principales puntos de medición de la humedad del suelo de la red VAS actual (Figura 3).

La distribución de los sensores SSM en la VAS ha evolucionado de acuerdo con los proyectos en curso dentro de un rango de tiempo determinado. La ubicación hasta 2020 se puede ver en la Tabla 1. Además, los detalles sobre el uso del suelo en esa configuración de estaciones de la VAS se muestran en la Tabla 2. Actualmente, la ubicación de los sensores en la VAS se muestra en la Tabla 3.

El conjunto de datos SSM de la VAS (capa superficial superior de 5 cm) y la precipitación de la estación de La Cubera de AEMet, se obtuvieron del conjunto de datos del proyecto *Ground-based Observations for Validation (GBOV) of Copernicus Global Land*

Products (GBOV, 2021). La Figura 4 muestra un ejemplo de una serie temporal de valores in situ de SSM.

Las imágenes de órbita ascendente y descendente del satélite Copernicus Sentinel-1 GRD A/B utilizadas se relacionaron con un polígono de interés de 120 m × 120 m centrado en VAS 1 y se obtuvieron con el navegador Sentinel-Hub EO (Sentinel-Hub, 2019). Estas señales retrodispersadas de radar de corrección del terreno (RTC) σ°-lineales se utilizaron para calcular un σ° ponderado de polarización cruzada denominado *ia* como se expresa en la ecuación 1.

$$\boldsymbol{ia = 0.67VV + 0.33VH\,, (eq - 1)}$$

El intervalo de tiempo de las imágenes de Sentinel-1 utilizadas osciló entre el 29/03/2015 y el 04/09/2019.

Para obtener una intuición física en profundidad sobre el diseño de IA desarrollado, se exploró estadísticamente una parcela de datos acumulados durante la operación de la VAS a largo plazo. Se utilizó una herramienta de visualización de diagrama de caja que contiene el símbolo x, que indica el valor promedio, mientras que los puntos de cierre por línea indican la mediana. En la estructura en forma de caja, el límite inferior representa el 1er cuartil (Q1) y el límite superior representa el 3er cuartil (Q3). Los puntos aislados muestran valores atípicos y las líneas horizontales aisladas delimitan los valores inferiores y superiores del dataset. Se verá su aplicación más adelante, en el apartado de resultados

El análisis de componentes principales (PC), que es un análisis estadístico multivariante y que se utiliza para reducir las características esenciales de un conjunto de datos caso por caso, se utilizó en dos conjuntos de datos. Este método permite construir gráficos cartesianos para visualizar las interconexiones de cada componente principal y las variables que permiten su construcción. Esto se debe a que los coeficientes subyacentes a la construcción de los componentes principales pueden interpretarse como una especie de proyección (desde un punto de vista vectorial) de las variables a lo largo de un determinado componente principal.

El instrumento CAROLS (Combined Airborne Radio-instruments for Ocean and Land Studies) a bordo de un avión ATR 42 de Météo France se utilizó en una serie de campañas aerotransportadas para la preparación de la validación de SMOS. Los diferentes sitios de muestreo de la campaña SVRT/CAROLS 2008 (Carbó et al., 2022) sobre la VAS se muestran en la Figura 5. El objetivo de estas campañas era establecer el modelo de emisión SMOS L-MEB (L-band Microwave Emission model of the Biosphere) para validar los productos SMOS y datos sobre tierra. Además de los vuelos, en la campaña SVRT/CAROLS 2008, los diferentes equipos recogieron muestras de suelo in situ georreferenciadas de áreas de control sobre el VAS con el fin de obtener materia orgánica del suelo, humedad del suelo y textura del suelo y, posteriormente, clasificar estas regiones en unidades fisio-hidrológicas homogéneas para la validación de SMOS (Carbó et al., 2022). La estructura de datos del SVRT/CAROLS 2008 utilizada es la siguiente: UTM_X (*easting*), UTM_Y (*northing*), HV_ vol es decir, humedad volumétrica de la muestra tomada (utilizada en esta tese como indicación de SSM), MO es decir, cantidad de materia orgánica en %, arena%, limo%, arcilla%.

Inicialmente, se generó e inspeccionó visualmente un gráfico de dispersión de la humedad relativa del suelo (es decir, colores relacionados con gradientes que van desde un valor máximo a mínimo) en función de las coordenadas UTM ED50.

Como se ha indicado, se utilizó un análisis de componentes principales (PC) (número de individuos 4438) en la búsqueda de una comprensión más profunda de los valores de SSM medidos in situ durante el período 2013-2020 (VAS1 a VAS8, excepto VAS6). También se llevó a cabo otro análisis PC (N=849 individuos) para la campaña SVRT/CAROLS 2008 (Figura 6), con el objetivo de obtener una visión más profunda de la varianza conjunta relacionada con las siguientes características físicas: longitud, latitud, SSM, materia orgánica del suelo, contenido de arena, limo y arcilla del suelo. Los datos de muestreo utilizados para la Campaña SVRT/CAROL fueron de 02/05/2008, 22/04/208, 24/04/2008 y 28/04/2008.

El primer PC del repositorio de datos permitía la siguiente visualización. Dicho análisis se realizó utilizando un ordenador portátil equipado con un microprocesador Intel Core I7 que ejecuta el sistema operativo Windows 10. El software estadístico (disponible de forma gratuita) utilizado fue BIOESTAT V5.3 (Mamiraua, 2023), que fue desarrollado y distribuido por el Instituto Mamirua para los estudios de la Amazonía.

A continuación, se emplearon métodos de *machine learning*, fundamentalmente de redes neuronales, para predecir SSM en la VAS1 (M1SV). Las 26 variables de entrada de los modelos (143 muestras) fueron, por un lado, de 1 a 24 (ia1, ..., ia24), que son los recuentos de las intensidades de píxeles lineales de σ° del SAR (de 0 a 255) registradas en el histograma de 24 bins de las imágenes del PI y, por otro lado, las otras 2 variables de entrada del modelo, denominadas IB1 e IB2, respectivamente, que representan la suma de los días posteriores a la fecha de la 1ª muestra utilizada (29/03/2015) y la precipitación diaria (mm) promediada a partir de los datos muestreados por hora.

Se exploró la aplicación Neural Fitting (perceptrón de retroalimentación multicapa) como un recurso para la modelización matemática con parametrización predeterminada. Las muestras de entrenamiento, validación y prueba (75%, 12,5% y 12,5%, respectivamente) fueron seleccionadas aleatoriamente en función de la disponibilidad de datos. Además, para este estudio, también se propusieron los 19 modelos ofrecidos por la aplicación de aprendizaje de regresión de MATLAB® 2020b estos modelos se clasifican en las siguientes clases principales: regresión lineal, árboles de regresión, máquinas de vectores de soporte, proceso gaussiano y conjuntos de árboles. Estos algoritmos alternativos de aprendizaje automático para redes neuronales artificiales también se exploraron en esta tesis. Estos experimentos se implementan en la app de aprendizaje de regresión de MATLAB®, que calcula el mejor modelo de regresión (procedimiento de aprendizaje supervisado) para un conjunto de datos determinado que se utilizará en la predicción de estos datos. Los modelos de regresión pueden entenderse como modelos de aprendizaje automático que aprenden la relación entre una variable dependiente (Y) y una o más variables independientes (X). Entre ellos:

1) **Regresores bayesianos**: estos modelos utilizan la teoría de la probabilidad para estimar los parámetros de la relación entre *y* y *x*. Son más robustos y flexibles que los modelos tradicionales como la regresión lineal, especialmente para datos no lineales.

2) **Los regresores en árboles de decisión** utilizan árboles de decisión, que son estructuras jerárquicas que dividen los datos en subconjuntos más pequeños basados en reglas específicas. Son fáciles de interpretar y pueden manejar diversos problemas de regresión.

3) Los **regresores en bosque aleatorios** se basan en árboles de decisión mediante el uso de un conjunto de muchos árboles de decisión, cada uno entrenado en un subconjunto aleatorio de los datos. Este enfoque de conjunto mejora la solidez en comparación con los árboles de decisión únicos.

4) Los modelos de **máquinas de vectores de soporte** utilizan la teoría de vectores de soporte para identificar un pequeño subconjunto de puntos de datos que representan todo el conjunto de datos. Ofrecen una alta precisión predictiva y robustez en comparación con los modelos tradicionales y de árbol de decisión. Los métodos de las máquinas vectoriales de soporte se basan, matemáticamente, en la separación de datos mediante el uso de hiperplanos (MATLAB, 2023 y referencias allí incluidas). Aunque estos métodos de aprendizaje supervisado se basan en conceptos matemáticos diferentes, comparten el mismo objetivo de ofrecer aproximaciones funcionales (conjuntos de datos). Por lo general, para cada función (conjunto de datos), hay un método más adecuado que minimiza el error para un conjunto determinado de restricciones. Existe una vasta bibliografía disponible sobre estos métodos, como se ha informado en otros lugares (Olivas et al., 2009).

El mejor resultado se obtuvo utilizando 50 neuronas en la capa oculta de ANN que contienen una función de activación no lineal (*tansig*), como se representa esquemáticamente en la Figure 7.

También se utilizó el algoritmo de Levenberg-Marquardt que es un método iterativo para encontrar el mínimo global de una función no lineal. En cuanto al descendente de escala, este algoritmo utiliza una escala para controlar la tasa de aprendizaje. Comienza con una escala alta y la reduce si el algoritmo no progresa. También es un algoritmo eficiente y confiable para encontrar el mínimo global de funciones no lineales. La regularización bayesiana es una técnica para evitar el sobreajuste en los modelos de aprendizaje automático. El algoritmo de reducción es un método iterativo para encontrar el mínimo global de una función no lineal.

En el mejor modelo se utilizó el algoritmo de aprendizaje de Levenberg-Marquardt, aunque también se exploraron la regularización bayesiana y los gradientes de escala como posibles opciones.

Se utilizó el coeficiente de correlación R como criterio para clasificar y elegir los dos modelos principales entre los 22 modelos metodológicos evaluados. El hardware utilizado fue un ordenador portátil equipado con un microprocesador Intel Core I7 en el sistema operativo Windows 10.

Para la extrapolación geográfica de la SSM a partir de VAS1, se desarrolló el módulo 2, M2GE, utilizando la aplicación Neural Fitting de MATLAB® para desplegar una segunda red neuronal superficial (perceptrón de retroalimentación multicapa) cuya entrada se acopló a la salida del M1SV, con el objetivo de proponer un método ampliamente accesible (para técnicos de todo el mundo) para la extrapolación geográfica (M2GE) de los resultados del M1SV.

Las entradas M2GE fueron las coordenadas de latitud y longitud del objetivo, una referencia de tiempo y el MSE generado por M1SV. Los datos de entrenamiento para M2GE fueron SSMs obtenidos utilizando el producto MODIS/TVDI (Temperature Vegetation Dryness Index) y una función de transferencia disponible en el ATBD para productos de humedad del suelo (LP-6, RM-10 y RM-11) de la iniciativa GBOV (López-Baeza et al., 2021) (ecuación 2):

$$\boldsymbol{SSM = -0.37TVDI + 0.30} \text{ con } R^2{=}0{,}83 \text{ (eq-2)}$$

El TVDI se utiliza para evaluar el valor de SSM (López-Baeza et al., 2021) en la VAS en los emplazamientos VAS1, VAS2, VAS3, VAS4, VAS5, VAS6 y VAS9.

Los esquemas de entrenamiento, validación y prueba de M2GE se realizaron utilizando una relación 60%, 20%, 20%, que se seleccionó aleatoriamente entre el conjunto de datos MODIS/TVDI disponible.

El hardware utilizado también fue un portátil equipado con un microprocesador Intel Core I7® en el sistema operativo Windows® 10. Esquemáticamente, esta configuración experimental se muestra en la Figura 8 y la Figura 9 detalla el módulo M2GE.

Para evaluar el rendimiento de dicho acoplamiento de M1SV y M2GE, respaldado por la evidencia inferida de PC, este enfoque de modelización en dos etapas se probó con mediciones in situ de VAS8 (que no se utilizó para el desarrollo de M2GE) y VAS5, que presentaban la correlación más baja con el conjunto de datos VAS1. Este desafío se exploró utilizando un diagrama de dispersión entre la salida del módulo y las medidas in situ, como objetivo. También se trazaron muestras secuenciales para permitir cierta exploración de la dinámica relacionada de los datos. Además, las predicciones de M1SV, impulsadas únicamente por los datos de Sentinel-1, se verificaron con los productos LP-6 del proyecto GBOV que se basa en los datos de MODIS/TVDI.

CONCLUSIONES

Esta tesis profundizó en dos vías innovadoras: 1) examinar el conjunto de datos VAS de efecto neto y 2) inferir la viabilidad de los datos SAR de Sentinel-1 para impulsar la recuperación de la humedad del suelo y predecir la humedad superficial del suelo (SMM) con menos de 1 km de resolución espacial (~120 m x 120 m) dentro de la estación de anclaje de Valencia. Dicha evaluación se basa en un método de aprendizaje automático en dos etapas que produce una R = 0,86 al predecir la SMM en VAS1 utilizando una red neuronal y datos de Sentinel-1. Para la extrapolación geográfica, el resultado obtenido fue R= 0,89, además del buen rendimiento en los retos de predicción para VAS8 (R = 0,92) y VAS5 (R = 0,88), en particular una zona de matorral cuyas características tienden a ser mucho más complejas que las de los viñedos. Considerados conjuntamente, estos resultados numéricos sugieren claramente que este enfoque Sentinel-1 es viable para la zona de control VAS 10 km x 10k m. Los buenos valores R obtenidos en el ajuste de los valores medidos in situ y el hecho de que existan muchas fuentes diferentes de datos del VAS indican que el corpus de datos del VAS puede dotarse, más allá de la autoconsistencia, del efecto neto de datos, lo que constituye una prometedora vía futura de investigación.

La característica de creación de capacidades de esta tesis se basa principalmente en la utilización de redes neuronales simples y poco profundas, metodologías MODIS/TVDI establecidas y extendidas, software disponible gratuitamente y una necesidad reducida de manejar el conjunto de datos de radar de banda C, así como de depender en gran medida del software/computación, lo que hace que este enfoque metodológico sea adecuado para su uso en todo el mundo, especialmente en países de bajos ingresos con disponibilidad limitada de instrumentos y recursos humanos para establecer grandes redes de instrumentos que operen para un largo alcance de estaciones similares.

Futuras mejoras y posibles ampliaciones: aprovechando el vasto corpus de datos del la VAS, podría desarrollarse un modelo multimedia generativo de aprendizaje automático y de entrada de lenguaje natural de última generación basado en estos datos. Nuevo análisis de series temporales: La integración de series temporales de humedad del suelo a largo plazo de alrededor de la zona de control podría mejorar significativamente la solidez y la posibilidad de generalización de las predicciones de SSM dentro de la región interior. Abordar las no linealidades asociadas con la rugosidad del suelo y la dinámica de la vegetación sigue siendo un reto clave para las recuperaciones basadas en RNA. La exploración de fuentes de datos alternativas, como los datos SAR de banda C procedentes de vehículos aéreos no tripulados (UAV), podría ofrecer un mayor realismo y apoyar mejores estrategias de gestión del uso del suelo. Gemelos digitales para la formación en materia de humedad del suelo y teledetección: Un modelo de este tipo podría actuar como un "gemelo digital" dela VAS, aceptando datos multimedia como entrada. Esto podría crear una configuración de "Observatorio Ciudadano", promoviendo la formación y la educación científica en humedad del suelo/detección remota, similar al concepto de "captación escolar" en hidrología. Intercambio mundial de conocimientos: Investigadores y profesionales de todo el mundo podrían beneficiarse de este "robot" comparando sus lugares de estudio locales con la VAS e interactuando con ella utilizando lenguaje natural y datos multimedia, obteniendo profundas percepciones sobre sus propios retos de SSM y teledetección.

En general, esta tesis avanza en el campo de la recuperación de la humedad del suelo a partir de Sentinel-1 para la Valencia Anchor Station, revelando la promesa de las RNA y la exploración del efecto neto. El concepto propuesto de "Observatorio Ciudadano" abre nuevas oportunidades para el intercambio global de conocimientos y las iniciativas educativas en los estudios de teledetección de la humedad del suelo.

ABSTRACT

The urgency of determining surface soil moisture content (SSM) worldwide, especially in agricultural soils, is well established. The Copernicus Sentinel-1 mission of the European Commission (EC) in conjunction with the European Space Agency (ESA) offers a synthetic aperture radar (SAR) that, together with *machine learning* methods, effectively contributes to this technological demand. This work aims to explore the capability of the soil moisture data available at the Valencia Anchor Station (VAS) in Valencia, Spain.

This thesis has two innovative, branching and equally important objectives: on the one hand, 1) the strategic objective is to explore the interrelationship and net effect of that already available body of VAS knowledge (with major emphasis on VAS1, the cornerstone of VAS). In addition, all this was done mainly, but not exclusively, by focusing on the information available from one of the first SSM probes, VAS1, operated continuously almost from its initial configuration, located next to the main mast which, naturally, is equipped with a more substantial body of knowledge, is equipped with a more substantial body of data and expertise to allow ML model results to be interpreted according to a guiding, physics informed framework, 2) the more immediate objective is a feasibility study of a Sentinel-1 driven tool for SSM over the 10 km x 10 km control area of the VAS. To achieve (1), it was necessary to (a) cross-use (blend) information from different available VAS datasets from previous studies and analyze numerical evidence such as the comparison between in situ measurements and prediction data. The Copernicus Sentinel-1 GRD A/B satellite ascending and descending orbit images used were related to a 120 m × 120 m polygon of interest centered on VAS 1 and obtained with the Sentinel-Hub EO navigator (Sentinel-Hub, 2019). This backscattered radar backscattered terrain correction (RTC) σ°-linear signals were used to calculate a cross-polarisation weighted C-band radar backscatter (σ°).

The Neural Fitting Matlab´s application (multi-layer feedback perceptron) was explored as a resource for mathematical modelling. The training, validation and test samples (75%, 12.5% and 12.5%, respectively) were randomly selected based on data availability. In addition, for this study, the 19 models offered by the MATLAB regression learning application® 2020b were also proposed. The best result was obtained using 50 neurons in the ANN hidden layer containing a non-linear activation function (*tansig*).

For the geographical extrapolation of SSM from the VAS1, a module 2, M2GE, was developed using the MATLAB® Neural Fitting application to deploy a second surface neural network (multilayer feedback perceptron) whose input was coupled to the M1SV output, with the aim of proposing a widely accessible method (for technicians worldwide) for the geographical extrapolation (M2GE) of the M1SV results. The result obtained is a two stage machine learning method that produces an R = 0.86 with respect to the SMM prediction in VAS1 using a neural network and Sentinel-1 data. Concerning the 2nd module, for geographical extrapolation of predictions for VAS (M2GE), the result obtained was R = 0.89, in addition to a fair performance in the prediction challenges for VAS8 (R = 0.86) and VAS5 (R = 0.88), notably a natural vegetation area whose features tends to be much more complex than those of vineyards. Taken together, these numerical results in comparison with ground-truth values strongly suggest that this Sentinel-1 approach is feasible for the 10 km x 10 km VAS control area and the called data net-effect very probably occurred.

CHAPTER 1

INTRODUCTION AND OBJETIVES

1. INTRODUCTION AND OBJECTIVES

1.1. INTRODUCTION

The urgency of determining the surface soil moisture (SSM) content worldwide, especially in agricultural soils, is well established. The Copernicus Joint European Commission (EC) and European Space Agency (ESA) Sentinel-1 mission offers a synthetic aperture radar (SAR), the European Radar Observatory, that, in conjunction with machine-learning-based methods, effectively contributes to this technological demand. Soil moisture is a critical factor for plant growth, as it affects the availability of water and nutrients to roots. Soil moisture measurement is important for agriculture, gardening, forestry, and other applications. Usually, it is measured using instruments that can be organized into 2 sets: 1) direct methods which directly measure water content of the soil such as X-ray diffraction, gravimetric and microwave methods and 2) indirect methods that measure some parameters related to soil water content as for example: resistivity, tensiometer and electrical conductivity (Montzka et al, 2020).

Singh et al. (2023) described soil moisture as the storage of water in soil pores for a certain time interval. This amount of water, in total, is very low compared with the total water available on Earth. Despite this, the air–soil interface is crucial for biological and hydrological processes. Studying this parcel of water is key for agricultural research, forest fires, flood studies, droughts, and climate studies, among others. Currently, satellite images such as Sentinel-1, Sentinel-2 and Radarsat-2 are available for covering the Earth surface.

According to Singh et al. (2023), on a global scale, optical and microwave methods have already been proven to be important for the difficult tasks of monitoring surface soil moisture on such a challenging scale. Considering the potential of the NASA-ISRO Synthetic Aperture Radar (NISAR) mission together with the improvements in machine learning methods such as automated machine learning (AutoML) and physics-informed algorithms, these authors state that there is an important avenue to explore in the next few years in the field of remote sensing SSM.

Singh et al. (2023) noted that ESA and NASA have been launching missions to provide soil moisture, such as the ESA's Soil Moisture and Ocean Salinity (SMOS) and NASA's Soil Moisture Active Passive (SMAP) under the Earth System Science Pathfinder (ESSP) missions. These SSM products have spatial resolutions varying from 1~50 km and revisitation times ranging from 1 to 8 days, but there are still some gaps in their coverage, especially related to dense vegetation and complex topography. The authors also stressed that the penetration capacity to the topsoil layer, sensitivity to dielectric properties, robustness to cloud coverage and day and night availability, and material availability are the key features that drive the use of microwave remote sensing in the task of estimating soil moisture.

Many kinds of backscattering models, semiempirical, empirical and theoretical, that rely on a permittivity (ϵ) gradient ranging from ~80 (water) to ~2 (dry soil) have been proposed based on quad-polarized (VV, VH, HH, and HV) microwave images. Considering that such quad-polarized microwave images for some regions are unavailable, machine learning methods that use data such as vegetation indices, topography and other ancillary data in addition to backscattering signals with other polarization features are commonly used.

Merzouki et al. (2019) stated that risks related to the agricultural sector could be dramatically reduced by Earth observation satellites, which provide accurate and timely soil moisture information that would allow farmers and all government levels to improve decision making. Weiss et al. (2020) noted in a meta-review publication on remote sensing applications in agriculture that surface soil moisture (SSM) is a primary variable that can be retrieved using different data domains, namely, solar radiation in a passive-multi/hyperspectral approach, thermal radiation in a passive approach and microwave radiation in an active and passive approach.

Ali et al. (2015) evaluated SSM retrieval using artificial neural network (ANN) inversion methods. These authors classified synthetic aperture radar as a type of microwave remote sensing method complementary to the use of radiometers and scatterometers. According to those authors, ground-based SSM measurements are scarce on a weekly/daily basis, and there is a clear need for this information on a small and large scale that allows sensitive information to be obtained for in-farm water management, especially in the event of droughts and floods, allowing reliable and fast decision making.

The retrieval of SSMs from microwave remote sensors relies on its relationship with soil dielectric properties, which is indeed a challenging and complex problem because of the combination of different soil characteristics, such as soil moisture, soil roughness, soil texture and vegetation cover, all of which affect a given unique electromagnetic response captured by a sensor (Ali et al., 2015).

This relationship between soil moisture/soil dielectric properties and microwave radiation emission can be locally validated based on data provided by ground-based measuring instruments strategically distributed across several sites, such as the Valencia Anchor Station in Eastern Spain. All these materials are inserted within the framework of the Copernicus Joint European Commission (EC) and the European Space Agency (ESA) Sentinel-1 mission and the so-called European Radar Observatory (Attema et al., 2008; Torres et al., 2012). This Copernicus mission (Sentinel-1) comprises 2 satellites sharing the same orbital plane, namely, Sentinel-1A and Sentinel-1B. Both systems carry an imaging C-band synthetic aperture radar (SAR) instrument in 4 modes that differ in spatial resolution (with a lower bound of 5 m) and coverage (up to 400 km). They also offer fast product delivery, dual polarization, and a short revisiting time. The focus of this mission is monitoring the environment and security thematic areas. The ability of SAR to operate independently during daylight and at night and under cloudy conditions manifests the potential of using Sentinel-1 data for frequently monitoring wide areas.

The investigation of surface moisture supported by Sentinel-1 products has been and continues to be an active and dynamic scientific endeavour. Frequently, the emphasis of these investigations is on how to optimally consider the effects of vegetation and surface roughness on the backscattering σ° signal and, ideally, on how to extend this freely available resource offered by Europe's Space policies for subsiding or optimizing smart irrigation strategies in agriculture worldwide, especially in underdeveloped countries such as Brazil and others, as reviewed elsewhere (Beale et al., 2019).

Rasheed et al. (2022) suggested that there are a variety of in situ methods for measuring SSMs that are limited or difficult to handle concerning large-scale efforts. These authors also noted that these remote sensing methods are normally based on satellite imaging, radar, visible and thermal infrared data, and many of them use a combination of two or more approaches. Moreover, the use of machine learning (data-driven) in the substitution of deterministic models for SSM prediction/forecast has also been frequently reported, especially for the generation of high spatial resolutions that are potentially very useful for water management in arid regions. However, choosing between methods for SSM studies and optimizing resource usage is a complex task that involves field conditions, among other factors.

Singh et al. (2023) noted that, among the methods for evaluating SSM, remote sensing is currently the most common method for publishing records, and the random forest algorithm is the most popular machine learning method for simulating SSM. Concerning the in-situ approach, electrical capacitance is ranked as the most frequently used method after systematically reviewing this theme. Machine learning methods usually require in situ measurements for training and test steps before their application. Additionally, optical and thermal images are frequently used as synergetic sources of data for predicting SSM. These authors also reported that scatterometer-based methods currently predominate, but the Sentinel-1 mission, and the expected NISAR, suggest that SAR-based methods will gain much more relevance concerning SSM studies and SSM products. After the Sentinel-1 mission started operation, the number of SAR-based methods available in the literature notoriously increased, especially because of the need to improve the considerations of surface roughness and the machine learning order to bypass the limitations inherent to the dual-polarized mode that is characteristic of this mission (Singh et al., 2023).

Among the algorithms most used for this purpose, according to bibliometric analysis, support vector machines, neural networks and random forests supported by in situ SSM measurements have been used. In addition, the use of fuzzy logic systems, deep belief networks, automated machine learning, decision trees, and Gaussian process regression, among others, was reported less frequently (Singh et al., 2023). Accordingly, in this work, in such ML (machine learning) models, the target variable and the input features share information. It is a frequent practice to consider the relevant features concerning the related physics of the SSM. However, this process must be performed in a balanced way because excess input information can influence the performance of a machine learning model. Consequently, the selection of input data is considered more critical than the choice of machine learning model. In addition, random forest models have been the most common option, and the China Academy of Sciences has emerged in bibliometric studies as the most prolific organization for publication in this field. The most common application linked with soil moisture sensors is remote sensing, which includes point-to-pixel comparisons. The statistical significance of the difference implies that at least 30 observations of the testing data and in situ soil moisture data are used to train the machine learning method (Singh et al., 2023).

Prado and Duarte (2020ab) survey and analyse the recent progress in using Fuzzy ARTMAP (FAM), a type of neural network model, and its variants, for mapping tasks through remote sensing data classification. This paper provides a perspective on this specific research area. Furthermore, this paper outlines potential research paths for scientists planning to employ FAM architectures in their investigations. Such a roadmap-style overview of FAM, offered by these authors, highlight applications in diverse research fields and point out recommendations for readily available code repositories to facilitate FAM practical implementation. Finally, Prado and Duarte (2020) identify current knowledge gaps regarding FAM's mapping applications, thereby proposing new avenues for future research and development in this domain.

Alexakis et al. (2017) reported that artificial neural network (ANN) methods potentially offer the possibility of capturing these complexities via a data-driven modelling approach that could be used for practical purposes, such as mapping SSMs, despite providing a complete mechanistic/causal understanding of the physical processes related to the enormous variety of farming and natural conditions observed in the Sentinel-1 coverage area.

Gaoet al. (2017) used Sentinel-2 optical data to monitor the SSM of two experimental fields in a 20 km by 20 km area located in Urgell (Catalonia, Spain and Mediterranean climate) and modelled it using a change detection approach, obtaining an R=0.31. Ezzahar et al. (2020), considering the difficulty of collecting soil roughness data, obtained results promising for the inversion of the σ° signal (VV and VH polarizations) of Sentinel-1 for SSM monitoring of agricultural bare soil (Tensift basin in Morocco) via a data-driven approach (support vector machine, SVM) compared to two widely used (classical) methods called the integral equation model (IEM) and semiempirical model (Oh) and supported by a ground-truth dataset. Ezzahar et al. (2020) used linear regression to predict the use of Sentinel-1 data to retrieve SSM from agricultural soil in the Tensift Basin (Morocco); they obtained R=0.84 using VV polarization and R=0.61 for VH polarization, which is considered more sensitive to volume scattering effects. In a further step, by the same authors, experiments with SSM retrieval relying on machine learning (support vector machine) were considered a reliable option. The RMSE obtained by the SVM approach was 2.8% against 2.7% obtained by the IEM.

Consolidated supervised ML relies on important data availability, and this is a strategic advantage for this work thanks to the ground-based availability of SSM information provided by the Valencia Anchor Station (VAS) and by the Sentinel-1 (SAR) coverage over the natural region of the *Utiel-Requena Plateau* at the *Valencia Autonomous Community* in Spain.

After 9 years of continuous operation, the VAS characteristics were described by Lopez-Baeza et al., 2009). The VAS has been used for the validation of low spatial resolution products from Earth observation missions since 2001. Examples of these missions are CERES (Clouds and the Earth's Radiant Energy System), EPS (EUMETSAT Polar System), and SMOS, among others. All these missions share features such as the necessity of being highly instrumented and well characterized large areas in addition to footprints near 50 km × 50 km. Fair homogeneity is a key feature of this area regarding validation activities, such as those carried out during the mission commissioning phase of SMOS, especially for the validation of level 2 products, soil moisture and vegetation water content. During this project, a control area of 10 km × 10 km was chosen due to its features regarding climate, soil type, lithology, elevation, slope and vegetation cover conditions of physio-hydrological units to be instrumented and monitored with a network of SM measuring stations interconnected by wireless communication (Cano et al., 2007 and 2012).

More recently, a *"time-line and road map"* communication was updated by Lopez-Baeza et al. (2022a and 2022b), who described some key details concerning this long-term activity at the VAS.

Together with concept of data net-effect or data network-effect with resemblances to the named Metcalf law (Gregory et al 2021) adopted in this thesis, that is as follows: the joint use of different sources and types of data (distinct domains) could potentially allow the emergence of a nonlinear effect (positive feedback loop) concerning their, jointly considered, explanatory power in the framework of machine learning and prediction tasks. Additionally, these references, offer insight into all the efforts done and probably still untouched potential of VAS-centred measurements and studies that can be drivers of machine learning methods to optimize the extraction of benefits of this vast available corpus of data.

1.2. OBJECTIVES

This thesis has two innovative, branching and equally important purposes: on the one hand, 1) the strategic purpose is to explore the interrelationship and net effect of that already available body of VAS knowledge (with greater emphasis on VAS1, the cornerstone of VAS). Moreover, all this was done mainly, but not exclusively, by focusing on the information available from one of the first SSM probes, VAS1, continuously operated almost from its initial configuration, located next to the main mast, which, naturally, is equipped with a more substantial body of data and expert knowledge to allow the results of the ML models to be interpreted according to a physics-informed and guiding framework, 2) the most immediate purpose is a viability study of a Sentinel-1-driven tool for SSM over the 10 km x 10 km control area of the VAS. In order to achieve 1), it was necessary to (a) cross use (mixing) of information of different datasets available about VAS from previous studies and analysis of the numerical evidence like comparison of in-situ measurements and predicted data. And, in order to achieve 2), (b) some machine learning methods were deployed, with emphasis in artificial neural networks, to predict SSM based on this heterogeneous set of data referenced in (a) and on the own dataset of ESA/Sentinel-1 radar backscatter values for VAS1.

On the second hand, to begin to capture the multiple potentialities of this body of data accumulated during more than 15 years of operation of SSM measurements on the VAS and as a possible reference for low-income countries, all the propositions explored (including geographical extrapolations) have been designed so that they can eventually be easily adopted by technicians in such countries who rely more on IT, often more accessible in their centres, than on the installation and maintenance of SSM sensor networks.

CHAPTER 2

MATERIAL AND METHODS

2. MATERIAL AND METHODS

This thesis benefits from the ground-based availability of soil moisture information provided by the Valencia Anchor Station and from the Copernicus Sentinel-1 (SAR) coverage over the natural region of the *Utiel-Requena Plateau* at the *Valencia Autonomous Community* in Spain. The data relies on the AEMET dataset of precipitation, Sentinel-1-based datasets and datasets of MODIS/TVDI and ThetaProbe measurements of SSM, both of which are related to previous work conducted at the VAS. Moreover, multivariate statistical and machine learning methods were also used in this thesis work. A schematic overview of the workflow of this thesis is as follows (Figure 1).

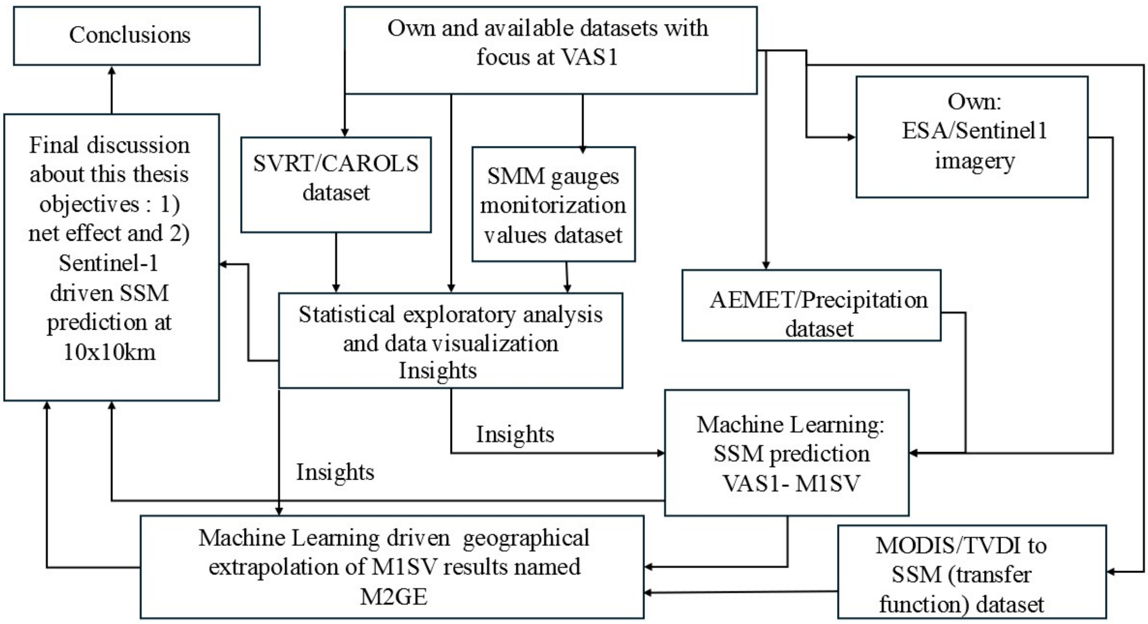

Figure 1: Schematic overview of the workflow of this thesis.

2.1. EXPERIMENTAL SITE

The *Valencia Anchor Station* (VAS) (N 39.570740°, W −1.288176°) is defined as "*an automatic meteorological station equipped with a large number of instruments to carry out validation activities of low spatial-resolution remote sensing data and products*" (Lopez-Baeza et al., in Bolle et al., 2006). These authors also reported that the VAS is located on a typical

Mediterranean ecosystem covering a large vineyard area also containing other species, such as shrubs, olive and almond trees, and pine forests, distributed on a reasonably homogeneous and flat area of 50 × 50 km. This monitoring infrastructure is located at the Utiel-Requena Plateau, 80 km in the west direction from Valencia city (eastern Spain). The monitoring point VAS1 (Figure 2) is one of the principal soil moisture measuring points from the current VAS network (figure 3).

Figure 2: Detail of VAS1 located at the center of the 1 km × 1 km control area; this was one of the first SSM monitoring points that anchored further developments of the VAS initiative.

The distribution of SM sensors at the VAS has evolved according to ongoing projects within a given time range. The location until 2020 is shown in Table 1. Moreover, details about land use in such a configuration of the VAS stations are displayed in Table 2.

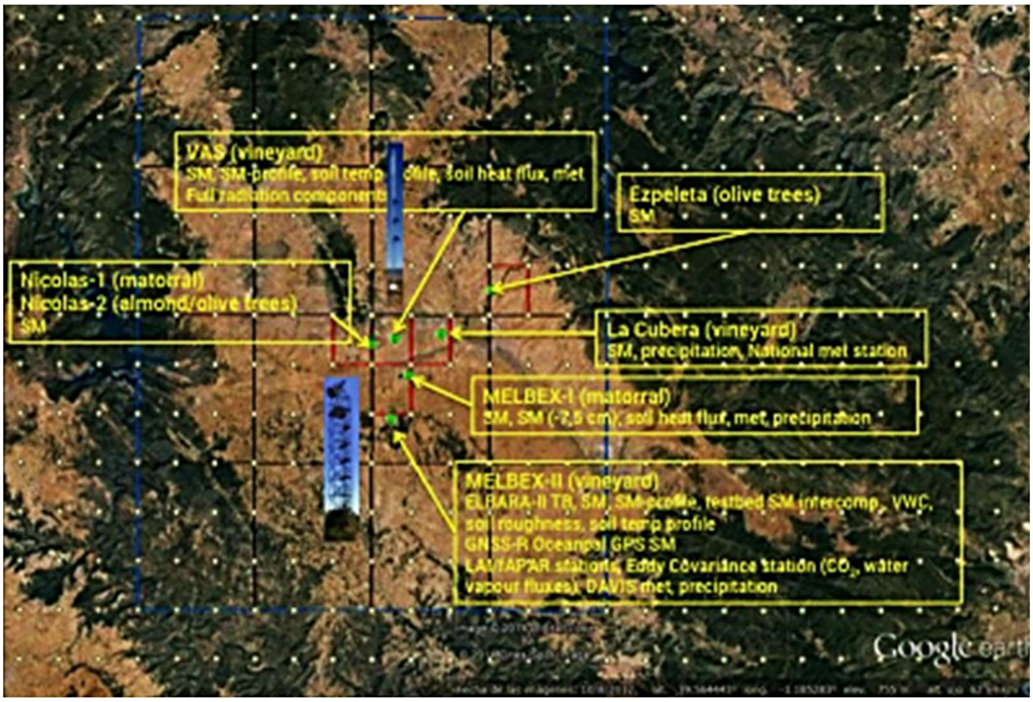

Figure 3: Schematics of the localization of a set of in situ sensors endowed with land use information from 2012 through 2021. (Source: GCS)

Table 1: Geographical location of sensors, including SSMs at the VAS until 2020.

ID	Latitude (°)	Longitude (°)
VAS1	39.57072	-1.28822
VAS2- MELBEX_II_177_01	39.52200	-1.29200
VAS3- MELBEX_II_177_02	39.52200	-1.29200
VAS4-MELBEX_II_178	39.52200	-1.29200
VAS 5-MELBEX_I	39.54807	-1.27730
VAS6- La Cubera	39.57333	-1.25137
VAS7-Nicolas_M1	39.56731	-1.30738
VAS8-Nicolas_T1	39.56731	-1.30738
VAS9-Ezpeleta	39.60035	-1.21133

In the table, MELBEX is a short abbreviation for *Mediterranean Ecosystem L-Band characterization Experiment*; for further details, see Wigneron et al., 2012 and Fernandez-Moran et al. (2014) and references therein. Currently, the location of the sensors at this site

(VAS) is as displayed in Table 3). Noteworthy that any data from this recent configuration was used in this thesis.

Table 2: Details about installation and land use related to different VAS monitoring stations.

VAS1	ThetaProbe at the VAS met station (bare soil on vine area)
VAS2	ThetaProbe at MELBEX_II_177_01 (vine)
VAS3	ThetaProbe at MELBEX_II_177_02 (between vine rows)
VAS4	Stevens Probe at MELBEX_II_178 (between vine rows)
VAS5	ThetaProbe at MELBEX_I (matorral)
VAS6	ThetaProbe at La Cubera (vine)
VAS7	ThetaProbe at Nicolas_M1 (matorral)
VAS8	ThetaProbe at Nicolas_T1 (almond and olive trees)

.

Table 3: The location of the sensors at the VAS.

ID	Latitude (°)	Longitude (°)
SM Station_1 VAS	39.57072	-1.28822
SM Station_2 VINE1	39.57120	-1.28842
SM Station_3 VINE2	39.57055	-1.28802
SM Station_4 FRUTAL:	39.56917	-1.29133
SM Station_5 VIÑAS:	39.57036	-1.28930
SM Station_6 FRUTAL (slightly out of the 1x1 km2 area)	39.54824	-1.27701

2.2. GROUND-BASED IN SITU MEASUREMENTS

The dataset from the Valencia Anchor Station, namely, the SSM (top superficial layer 5 cm) and precipitation from the rain gauge station at La Cubera by AEMet, was retrieved from the Ground-based Observations for Validation (GBOV) of Copernicus Global Land Products project dataset (GBOV, 2021). An example of an in-situ time series of SSM values is displayed in Figure 4.

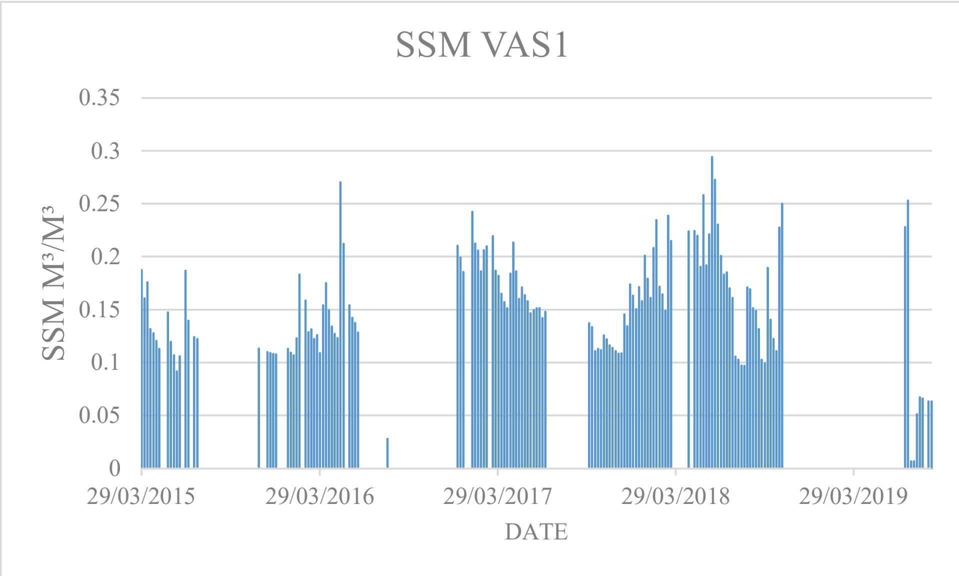

Figure 4: Ground-truth soil moisture at VAS1 for the dates when there is availability of Sentinel-1 ground range detected (GRD) imagery and daily precipitation records, summing up 143 samples in total (concerning different days).

2.3. SENTINEL-1 SATELLITE AND PRODUCTS

The Sentinel-1 GRD A/B ascending and descending orbit imagery used was related to a polygon of interest of 120 m × 120 m centred at VAS 1 and obtained with the Sentinel-Hub EO browser (Sentinel-Hub, 2019). Such radiometric terrain correction (RTC) radar backscattered signals σ°–linear were used to calculate a weighted cross-polarized σ° named ***ia*** as follows (equation 1):

$$\boldsymbol{ia = 0.67VV + 0.33VH} \qquad \boldsymbol{(eq-1)}$$

The time range of the Sentinel-1 imagery used ranged from 29/03/2015 to 04/09/2019.

2.4. GAINING PHYSICS INSIGHTS THROUGH STATISTICAL EXPLORATORY ANALYSIS OF AVAILABLE DATA

To obtain in-depth physical intuition about the AI design evoked, a parcel of accumulated data during the long-term VAS operation was statistically explored. A boxplot visualization tool has been used that contains the symbol x, which indicates the average value, whereas the close-by-line points indicate the median. In the box-shaped structure, the lower limit represents the 1st quartile (Q1), and the upper limit represents the 3rd quartile (Q3). The isolated points show outliers, and the isolated horizontal lines delimit lower and higher values of the dataset. Its application will be discussed later in the results section.

The principal component (PC) analysis, which is a multivariate statistical analysis used to reduce the essential characteristics of a dataset on a case-by-case basis, was used on two datasets. This method allows the construction of Cartesian plots to visualize the interconnections of each principal component and the variables that allow its construction. This is because the coefficients underlying the construction of the PC can be interpreted as a kind of projection (from a vectorial viewpoint) of the variables along a given PC.

The CAROLS (Combined Airborne Radio-instruments for Ocean and Land Studies) instrument onboard an ATR 42 aircraft from Météo France was used on a series of airborne campaigns for the preparation of SMOS validation. The different SVRT/CAROLS 2008 campaign (Carbó et al., 2022) sampling sites over the VAS are shown in Figure 5. The objective of these campaigns was to establish the SMOS emission model L-MEB (L-band Microwave Emission model of the Biosphere) to validate the SMOS products and land data. In addition to the flights, in the SVRT/CAROLS 2008 campaign, the different teams collected georeferenced in situ soil samples from control areas over the VAS in order to obtain soil organic matter, soil moisture and soil texture and, subsequently, classify these regions in homogeneous physio-hydrological units for the validation of SMOS (Carbó et al., 2022). The structure of data of SVRT/CAROLS 2008 used is as follows: UTM_X (easting), UTM_Y (northing), HV_ vol i.e.

volumetric humidity of the sample taken (used as an indication of SSM), MO i.e. amount of organic matter in % , sand%, silt %, clay%.

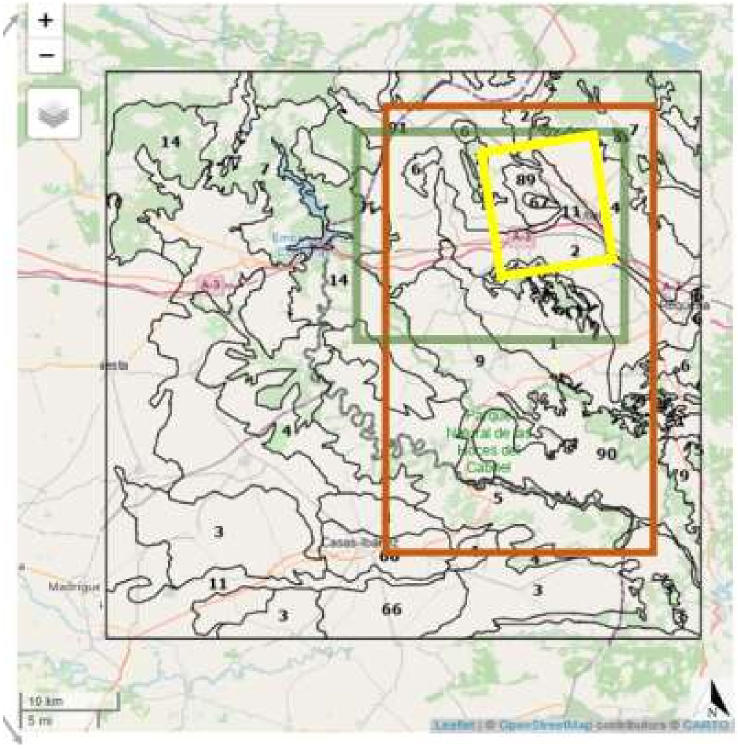

Figure 5: Map of SVRT/CAROLS campaign (2008) location (minor square traced in yellow). Source: GCS and (Carbó et al., 2022).

Initially, a scatter graph of relative (i.e., colours related to gradients ranging from the max-min SSM) soil moisture as a function of UTM ED50 coordinates was generated and visually inspected.

A Principal Components analysis (number of individuals 4438) was used in the search for a more in-depth understanding of the SSM values measured in situ during the 2013–2020 period (VAS1 to VAS8, except VAS6). Moreover, another PC analysis (N=849 individuals) was also carried out for the SVRT/CAROLS 2008 campaign (Figure 6), aiming at gaining more depth insights into the joint variance related to the following physical features: latitude, longitude, SSM, soil organic matter, sand, silt and clay content of the soil. The sampling data used for SVRT/CAROL´s campaing were from dates 02/05/2008, 22/04/2008, 24/04/2008 and 28/04/2008.

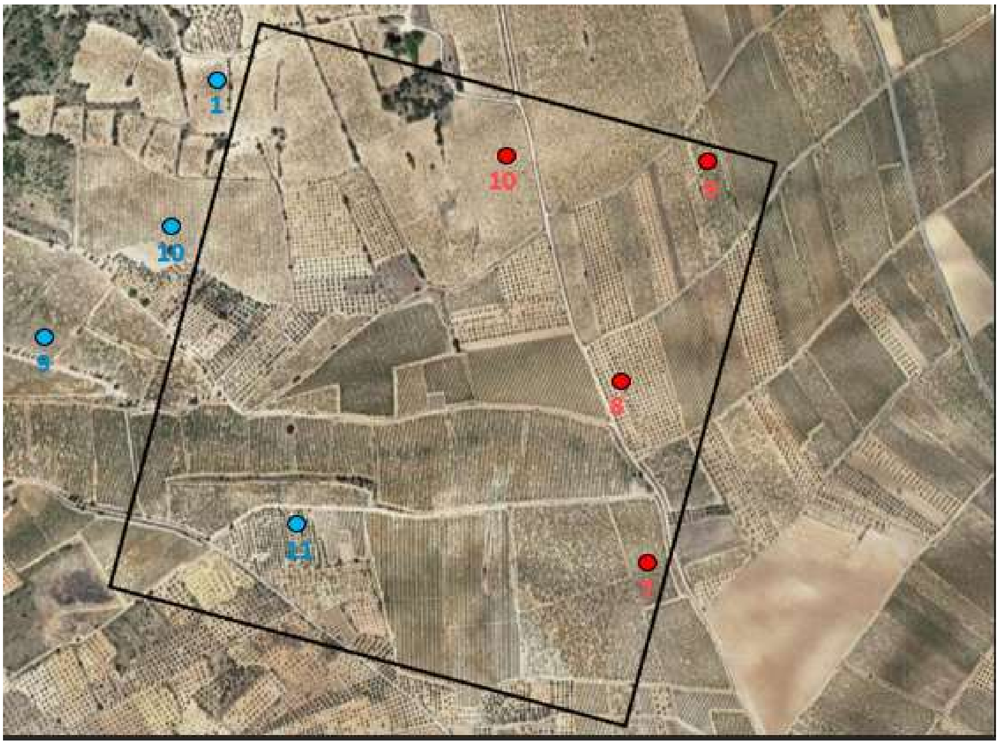

Figure 6: Schematics illustrated by multiple team collection soil samples at VAS and, specifically, in this case, the sampling points of 2 distinct teams around the VAS1 site in the framework of the SVRT/CAROLS campaign for analysis of soil moisture, organic matter and texture features. Source: GCS.

The first PC of the data repository allowed the following visualization. Such an analysis was performed using a notebook equipped with an Intel core I7 microprocessor running the Windows 10 operational system. The statistical software (available for free) used was BIOESTAT V5.3 (Mamiraua, 2023), which was developed and distributed by the Mamirua Institute for Amazon´s studies.

2.5. MACHINE LEARNING ANN METHODS FOR MODULE 1 TO PREDICT SSM AT VAS1 (M1SV)

The 26 input variables of the models (143 samples) were, on the one hand, 1 to 24 (ia_1, ..., ia_{24}), which are the counts of SAR´s $\sigma°$ – linear pixel intensities (from 0 to 255) recorded at 24 bins´ histogram of the PI´s imagery and, on the other hand, the other 2 model input variables, named ib_1 and ib_2, respectively, representing the sum of the days after the date of the 1st sample used (29/03/2015) and the daily precipitation (mm) averaged from hourly sampled data.

The Neural Fitting App (multilayer feed-forward Perceptron) was explored as a resource for math modelling with default parametrization. The training, validation and test samples (75%, 12.5% and 12.5%, respectively) were randomly selected based on the data availability. Additionally, all 19 models offered by MATLAB®′s 2020b regression learner app were proposed for this study; these models are classified into the following main classes: linear regression, regression trees, support vector machines, Gaussian process, and ensembles of trees.

These alternative machine learning algorithms for artificial neural networks were also explored in this thesis. These experiments are implemented in the regression learning app in MATLAB®, which calculates the best regression model (supervised learning procedure) for a given dataset to use in the prediction of these data. Regression models can be understood as machine learning models that learn the relationship between a dependent variable (y) and one or more independent variables (x). Among them:

1) **Bayesian regressors**: these models use probability theory to estimate the parameters of the relationship between y and x. They are more robust and flexible than traditional models like linear regression, especially for non-linear data.

2) **Decision tree regressors** that use decision trees, which are hierarchical structures that divide the data into smaller subsets based on specific rules. They are easy to interpret and can handle diverse regression problems.

3) **Random forest regressors** build on decision trees by using an ensemble of many decision trees, each one trained on a random subset of the data. This ensemble approach improves robustness compared to single decision trees.

4) **Support vector machine models** that use support vector theory to identify a small subset of data points that represent the entire dataset. They offer high predictive accuracy and robustness compared to both traditional and decision tree models. Support vector machines methods rely,

mathematically, on separating data by using hyperplanes (MATLAB® 2023 and references there in). Although these methods for supervised learning are based on different mathematical concepts, they share the same objective of offering functional (dataset) approximations. Usually, for each function (dataset), there is a more suitable method that minimizes the error for a given set of constraints. There is a vast body of literature available about these methods, as reported elsewhere (Olivas et al., 2009).

The best result was obtained by using 50 neurons in the ANN hidden layer containing a nonlinear activation function (*tansig*), as schematically represented in Figure 7.

The Levenberg-Marquardt algorithm was also used, that is an iterative method for finding the global minimum of a non-linear function. Concerning the scale-descendent, this algorithm uses a scale to control the learning rate. It starts with a high scale and reduces it if the algorithm is not progressing. It is also an efficient and reliable algorithm for finding the global minimum of non-linear functions. Bayesian regularization is a technique for avoiding overfitting in machine learning models. The scale-down algorithm is an iterative method for finding the global minimum of a non-linear function.

The Levenberg–Marquardt learning algorithm was used in the best model, although Bayesian regularization and scale gradients were also explored as possible options.

The correlation coefficient *R* was used as the criterion for ranking and choosing the top two models among the 22 methodological different models evaluated. The hardware used was a notebook equipped with an Intel core I7 microprocessor on the Windows 10 operational system.

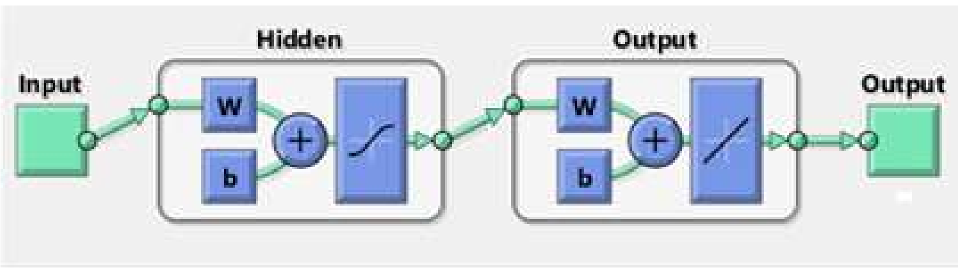

Figure 7: Schematic representation of the M1SV/ANN: 26 input variables. The best results were achieved using 50 artificial neurons in the hidden layer and 1 neuron as the output of the ANN.

2.6. MACHINE LEARNING ANN METHODS FOR MODULE 2- FOR GEOGRAPHICAL EXTRAPOLATION OF SSM AT VAS1 (M2GE)

The MATLAB® Neural Fitting App was used to deploy a 2nd shallow neural network (multilayer feed-forward perceptron) whose input was coupled to the output of the M1SV, aiming at the proposition of a widely accessible (for technicians worldwide) method for geographical extrapolation (M2GE) of the results of the M1SV.

The M2GE inputs were the target latitude and target longitude coordinates, a time reference and the SSM generated by M1SV. The training data for M2GE were SSMs obtained using the MODIS/TVDI (Temperature Vegetation Dryness Index) product and a transfer function available from the Algorithm Theoretical Basis Document for soil moisture products (LP-6, RM-10 and RM-11) from the GBOV initiative (Lopez-Baeza et al., 2021) (equation 2):

$$\boldsymbol{SSM = -0.37TVDI + 0.30 \quad (eq-2)}, \text{ with } R^2\text{=0,83}$$

The TVDI is used to assess the SSM value (Lopez-Baeza et al., 2021) at the VAS for the sites VAS1, VAS2, VAS3, VAS4, VAS5, VAS6 and VAS9. The training, validation and test schemes of M2GE were performed using a 60%:20%:20% ratio, which was randomly selected from among the available MODIS/TVDI dataset. The hardware used was also a notebook equipped with an Intel core I7® microprocessor on the Windows® 10 operational system. Schematically, this experimental setup was as shown in Figure 8, and the isolation procedure for detailed analysis of M2GEs is shown in Figure 9.

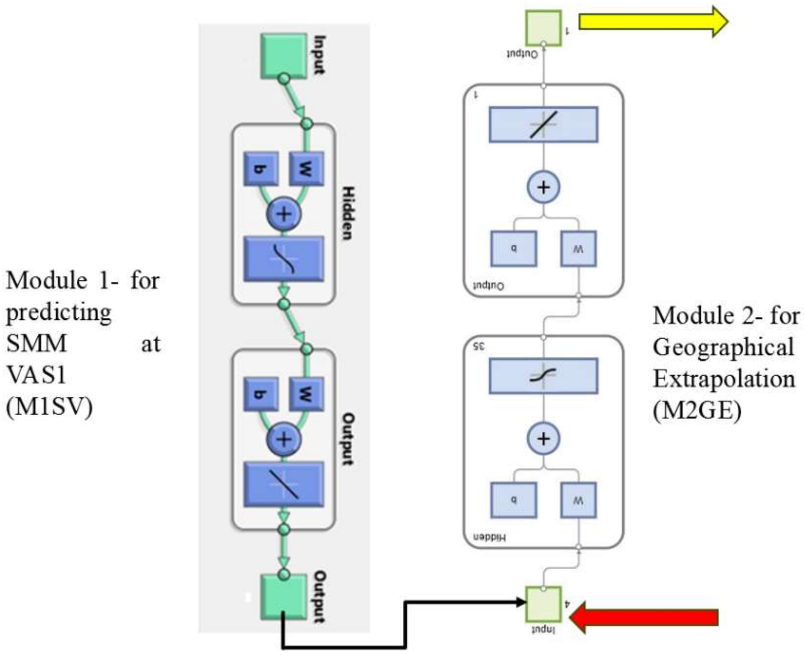

Figure 8: Schematic of the proposed coupling between both ML models. The red arrow indicates the *Lat*, *Long* and time inputs (the same as M1SV), for which the SSM value in VAS1 is extrapolated and the SSM value in that input in red is represented by the yellow arrow.

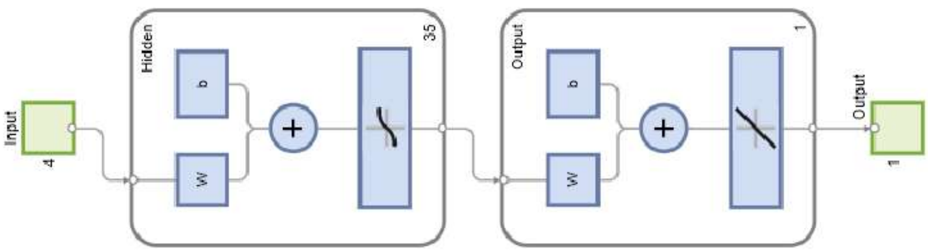

Figure 9: Detail schematics of the M2GE.

To evaluate the performance of such a coupling of M1SV and M2GE, supported by evidence inferred from PC, these two-stage modelling approach were tested against in situ measurements of VAS8 (which was not used for M2GE deployment), and VAS5, which presented the lowest correlation with the VAS1 dataset. This challenge was explored using a scatter plot of the target compound. Moreover, sequential samples were plotted in order to allow some exploration of the related dynamics of the data. Besides, the predictions of M1SV, only driven by Sentinel-1 data, were verified against LP6 products from GBOV initiative that rely on MODIS/TVDI data.

CHAPTER 4

RESULTS AND DISCUSSION

3. RESULTS AND DISCUSSION

3.1. EXPLORATORY ANALYSIS OF DATA AVAILABLE IN THE VALENCIA ANCHOR STATION

Some soil properties inferred during the SVRT/CAROLS campaign (texture, soil moisture, organic matter) (Carbó et al., 2022) were graphically represented (as a boxplot) to allow for some environmental analysis praxis (Figures 10, 11 and 12)

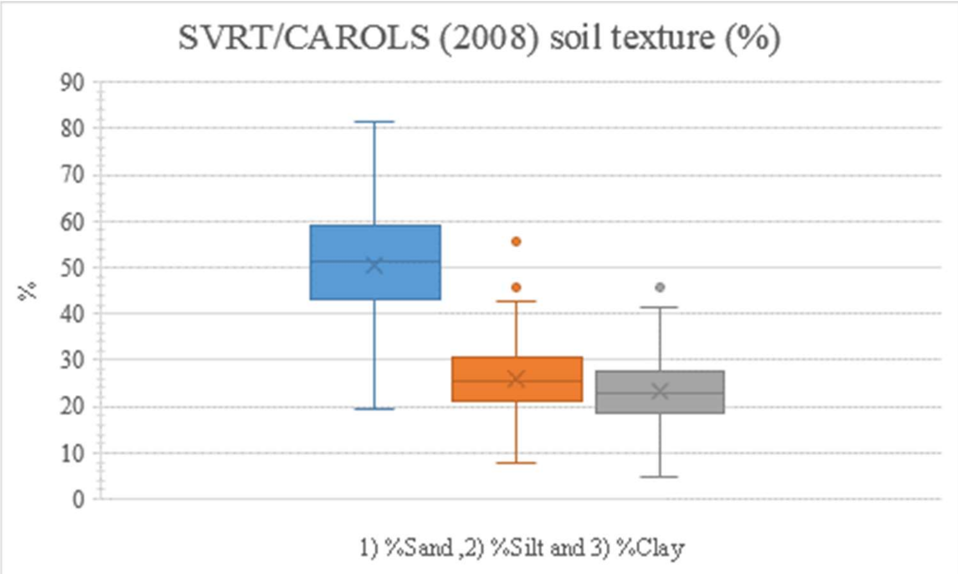

Figure 10: Box-plot graph related to the soil texture inferred during the SVRT/CAROLS campaign. In each box, the x symbol indicates the average value, and the horizontal line indicates the median. In the box-shaped structure, the lower limit represents the 1st quartile (Q1), and the upper limit represents the 3rd quartile (Q3). The isolated points show outliers, and the segment short horizontal lines delimit the lower and higher values of the dataset.

Notably, the average contents of clay and silt are fairly equal (~25%), whereas the average amount of inferred sand is 51%. The sand soil content is consolidated as a factor that strongly favours soil drainage and discharge in water bodies. Additionally, the same strategy was used to display soil moisture and organic matter content (Figure 11)

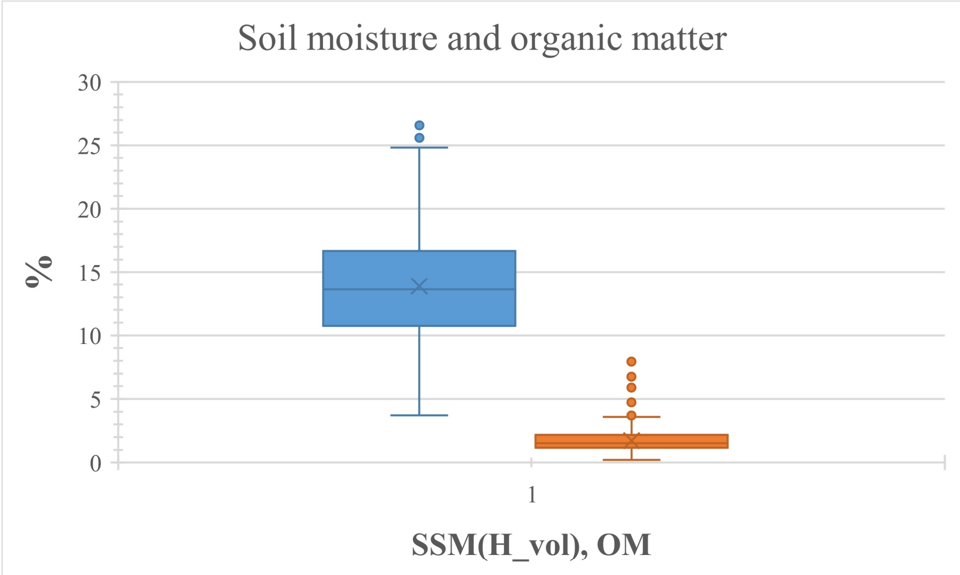

Figure 11: Box-plot graph displaying SSM (left, %) and organic matter (right, in %).

Although the SVRT/CAROLS campaign was occasional in time, the information it provides on organic matter could also be somewhat valid in time, whereas the SSM has a comparatively more punctual characteristic.

Figure 12 shows a 3-D plot scattering of the SVRT/CAROLS inferred Hvol % (as an indicator of SSM) against the geographical coordinate data performed for spatial visualization of soil moisture and its distribution patterns.

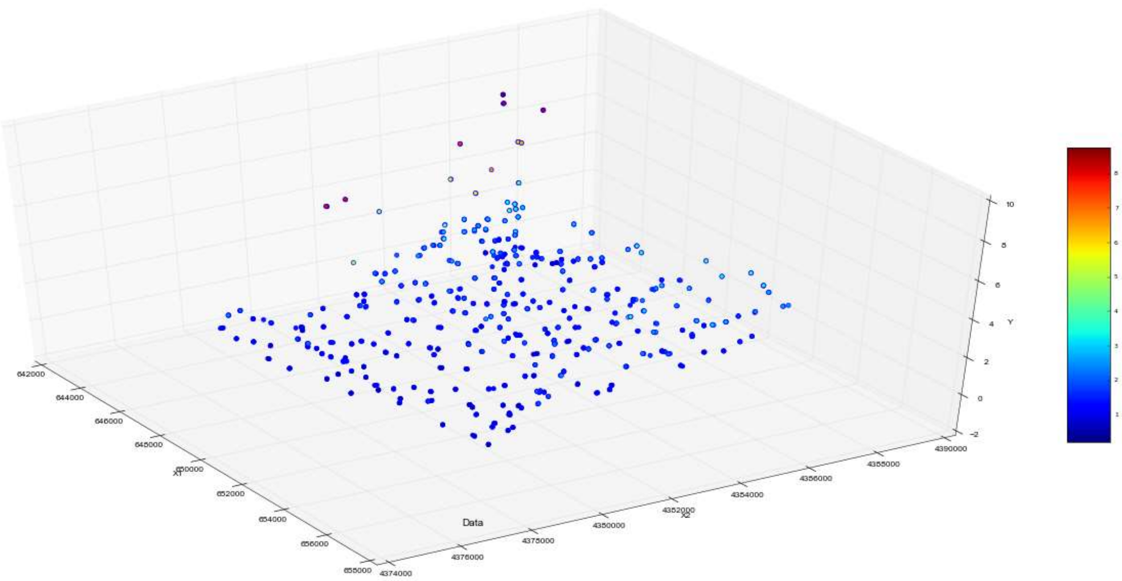

Figure 12: 3-D scatter plot of soil moisture and SSM, which was plotted and visually inspected as a function of UTM ED50 coordinates. Horizontally, the geographical latitude (X-axis) varied between (using UTM ED50) 4374000 (left) and 4390000 (right) , and the longitude (Y-axis) varied between 642000 and 655000 (using UTMED50) at its intersection with the horizontal axis. On the vertical Z-axis, a relative/normalized colour representation of the SSM gradient ranging from dark blue (lowest SSM relative value, 0.5 related to 2.7%) to red (maximum inferred relative value, 8.5 related to 29%) is shown.

Visual inspection of this graph seems to reveal a certain pattern in the distribution of SSM over latitude (UTM-Y), i.e., a gradient of increasing SSM towards north, although the effect of longitude (UTM-X) is more unclear. Table 4 shows the results of the correlation matrix obtained from the principal component (PC) analysis of the SVRT/CAROLS (2008) campaign dataset.

Table 4: Correlation matrix between the CAROLS parameters used.

Correlation Matrix	X_UTM	Y_UTM	SSM	OM	Sand	Silt	Clay
X_UTM	1	---	---	---	---	---	---
Y_UTM	-0.113	1	---	---	---	---	---
SSM	0.0495	0.0278	1	---	---	---	---
OM	-0.2117	0.4402	0.0726	1	---	---	---
Sand	-0.2482	0.2232	-0.1145	0.1103	1	---	---
Silt	0.3335	-0.0956	0.1046	0.0415	-0.8084	1	---
Clay	0.0531	-0.2636	0.0769	-0.2241	-0.7813	0.2641	1

The variance of these datasets attributed to each of the PCs was plotted as shown in Figure 13 and 14.

Almost all the variance in the SVRT/CAROLS dataset is encoded only by 6 PCs, but it can be inferred that the patterns related to this dataset are quite complex because PC1 and PC2 explain ~ 55% of the variance, which contrasts with frequently observed cases of studies in which the variance is captured almost entirely by PC1 and PC2. This is probably due to some transient variables, such as climatological variables, not captured by this study of very short temporal duration and/or a lack of consideration of physical features, as others are more robust (explanative) variables. To detail these PCs, Table 5 shows all numerical values of the PCs.

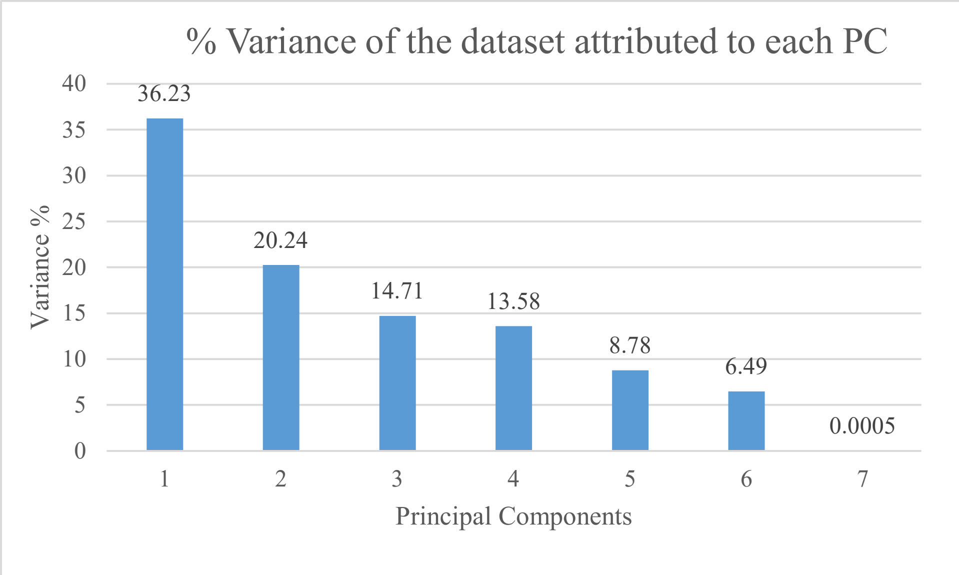

Figure 13: This bar chart is intended to illustrate how the variance of the data is captured for each of the 7 PCs specific to this SVRT/CAROLS (2008) dataset.

Table 5. Complete set of numerical values of the 7 PC coefficients.

Coefs. Autovetores	Coef. X-UTM	Coef. Y-UTM	Coef. SSM	Coef. OM	Coef. Sand	Coef. Silt	Coef. Clay
PC 1 =	0.2668	-0.2754	0.0938	-0.2105	-0.5943	0.4702	0.475
PC 2 =	-0.0431	0.5549	0.2875	0.6583	-0.2096	0.3592	-0.0368
PC 3 =	0.8179	0.071	0.0305	-0.1142	0.1591	0.2162	-0.4899
PC 4 =	-0.0093	-0.1525	0.9508	-0.1429	0.1204	-0.1939	0.0083
PC 5 =	0.2597	0.694	-0.0139	-0.3433	-0.0278	-0.3718	0.4402
PC 6 =	0.4364	-0.326	-0.0574	0.6091	0.0637	-0.44	0.3623
PC 7 =	-0.0001	-0.0003	0	-0.0002	0.7472	0.4836	0.4559

To facilitate the visualization and search for clustering features, the scattering diagramme of PC1 and PC2 was plotted for the CAROLS dataset (Figure 14).

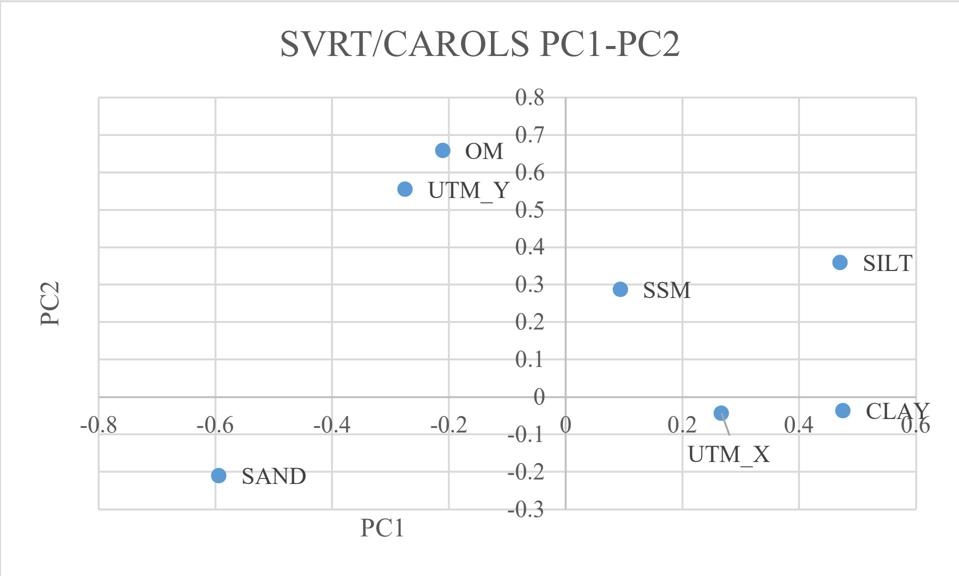

Figure 14-Scattering plot of the projection of each variable on PC1 (horizontal-axis) and PC2 (Y-axis). A total of 1571 individuals were used for this analysis.

In such a framework, i.e., considering only the two more important PCs (PC1 and PC2), SSM had a more important projection (in vector reasoning) along OM, probably due to the so-called OM buffer effect (Werner et al., 2020), closely related to land use characteristics for SSM studies, especially for the typical VAS land use (vineyards, trees and scrubs) matrix. The SSM mixed relationship with latitude (comparatively stronger) and longitude (Figure 15) possibly has some considerable correlation with a diagonal gradient of altitude as indicated by the location of the Magro River and its tributaries (the main drainage of the related hydrological basins that names it also), which is approximately parallel to a given meridian and located near the eastern extreme concerning the 10 km × 10 km VAS control area. Moreover, the geographical location of one of its main tributaries, the Madre River, which drains the region of Caudete de las Fuentes, seems to corroborate this SSM diagonal gradient (hydraulic) driven by the drainage discharge and that seems to be a pattern already observed in many other hydrological basins in Europe, as described elsewhere (Ghajarnia et al., 2020).

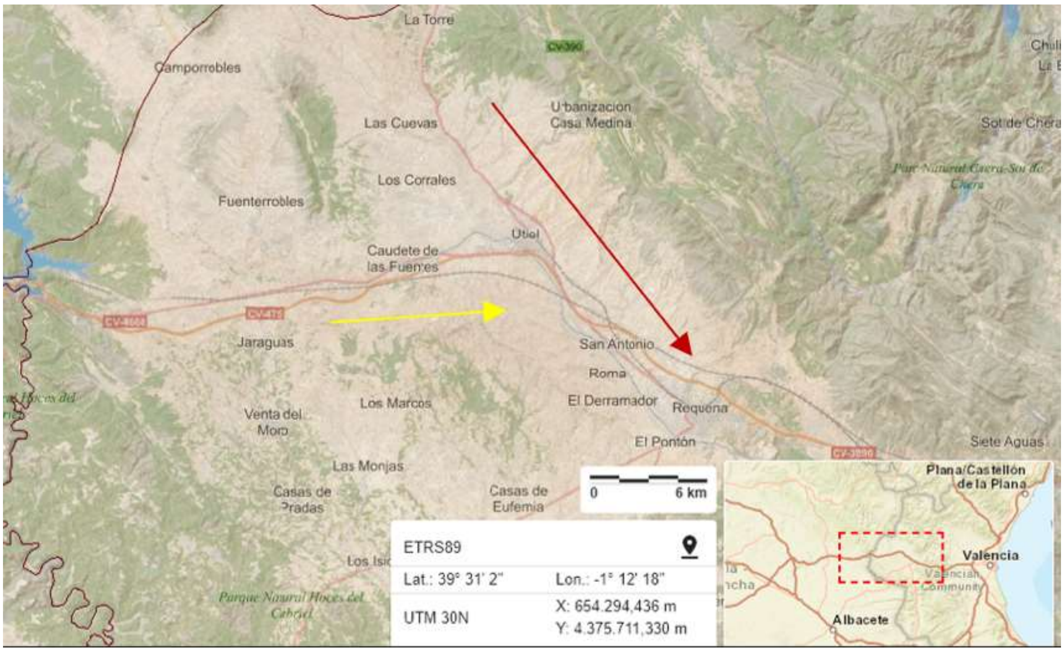

Figure 15. Magro River basin. The red arrow shows the Rio Magro, whereas the yellow arrow illustrates the Rio Madre disposition/orientation features (own authorship using INFOIGME).

This same area was mapped by high-resolution aerial imaging by instruments on board of ATR-42 (Météo France) the same that carried CAROLS instrument (Figure 16):

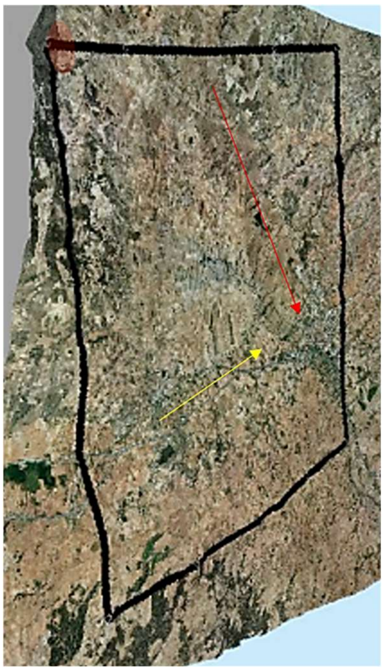

Figure 16: Magro River basin. The red arrow shows the Rio Magro, whereas the yellow arrow illustrates the Rio Madre disposition/orientation features superimposed on a high-resolution aerial image capture.

In Figure 17, SSM clearly shows a comparatively lower projection along PC1 aligned with the clay variable, which, consequently, agrees with established knowledge that relates SSM accumulation and soil clay content. The same phenomenon extends to the intermediary texture of the soil, silt, which also has the most important projection (in vector-like reasoning) in Fig. 18 along clay and PC1. The connection between longitude (UTM_X) and clay content at this point is unclear, although it is well known that the clay fraction of soil can slowly move along hydraulic gradients, and these characteristics could be the initial starting points for exploring this inference in greater depth. As expected, the clayey features in that figure are in diametral contrast with those of the sand fraction, as intuitively expected.

An exploratory analysis of the temporal series of SSM data from VAS1 (V1) to VAS8 (V8) was subsequently performed. The average SSM values of these time series are plotted in Figure 17.

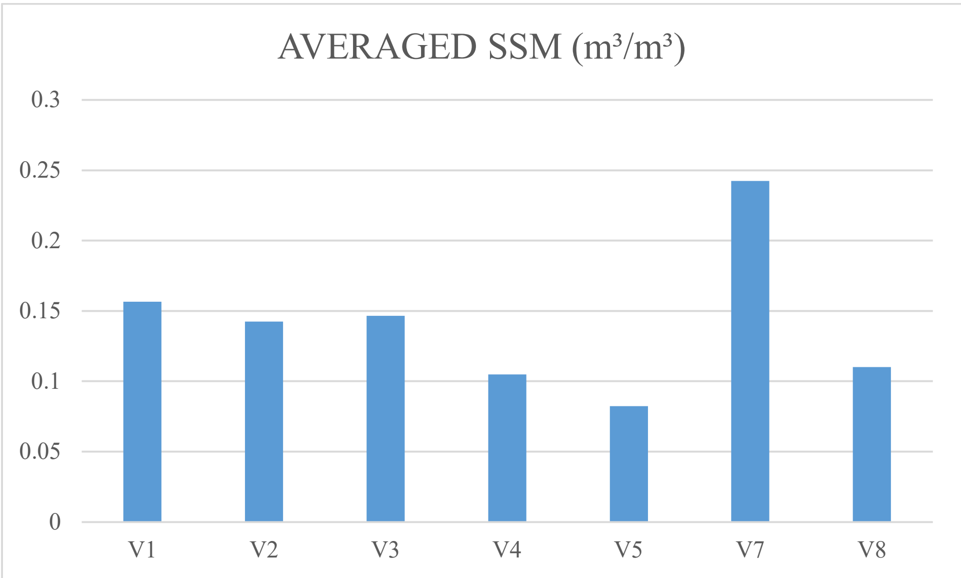

Figure 17: The average SSM values of these time series were plotted.

The higher values inferred concerning V7 (VAS7) may be due to its location/surroundings as the natural type of Mediterranean vegetation called matorral.

The correlation matrix results obtained from the PC analysis are given in Table 6.

Table 6: Correlation matrix between the VAS1, VAS2, VAS3, VAS4, VAS5, VAS7, and VAS8 SSM time series.

Correlation Matrix	VAS1	VAS2	VAS3	VAS4	VAS5	VAS 7	VAS8
VAS1	1	---	---	---	---	---	---
VAS2	0.886	1	---	---	---	---	---
VAS3	0.7035	0.531	1	---	---	---	---
VAS4	0.4992	0.4048	0.7786	1	---	---	---
VAS5	0.4971	0.4747	0.7681	0.8449	1	---	---
VAS7	0.5401	0.4018	0.8507	0.8289	0.7624	1	---
VAS8	0.7526	0.6184	0.8867	0.8107	0.7911	0.8327	1

To detail those PCs, the full table of numerical values of the PC components is presented Table 7.

Table 7: Complete set of numerical values of the 7 PC coefficients.

Coefs. Autovectors	Coef. VAS1	Coef. VAS2	Coef. VAS3	Coef. VAS4	Coef. VAS5	Coef. VAS7	Coef. VAS8
PC 1 =	0.3508	0.3063	0.4078	0.3832	0.3797	0.3874	0.4192
PC 2 =	-0.5461	-0.6484	0.0928	0.337	0.2652	0.2976	0.0172
PC 3 =	-0.1917	0.3371	-0.4246	0.2897	0.6434	-0.3763	-0.1728
PC 4 =	-0.0428	-0.1688	0.497	-0.5812	0.478	-0.3762	0.1217
PC 5 =	-0.2073	0.307	-0.0924	-0.4746	0.2279	0.6719	-0.3543
PC 6 =	-0.1196	-0.0435	-0.5248	-0.2945	0.0418	0.1004	0.7809
PC 7 =	-0.6948	0.4975	0.3374	0.0768	-0.2989	-0.1291	0.2096

The variance of this SSM dataset attributed to each of the PCs was graphically represented in Figure 18.

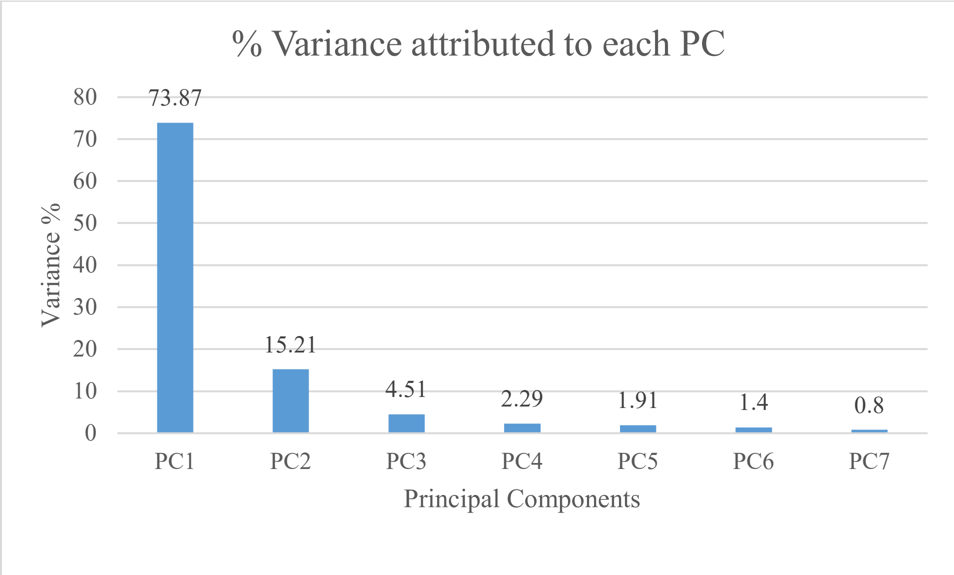

Figure 18: This bar graph aims to illustrate the amount of variance in the data captured by each of the 7 PCs (in direct proportion with the variable quantity).

To provide a friendly visualization and to search for clustering features, the PC1 and PC2 dispersions were plotted for the PC of SVRT/CAROLS dataset (Figure 19).

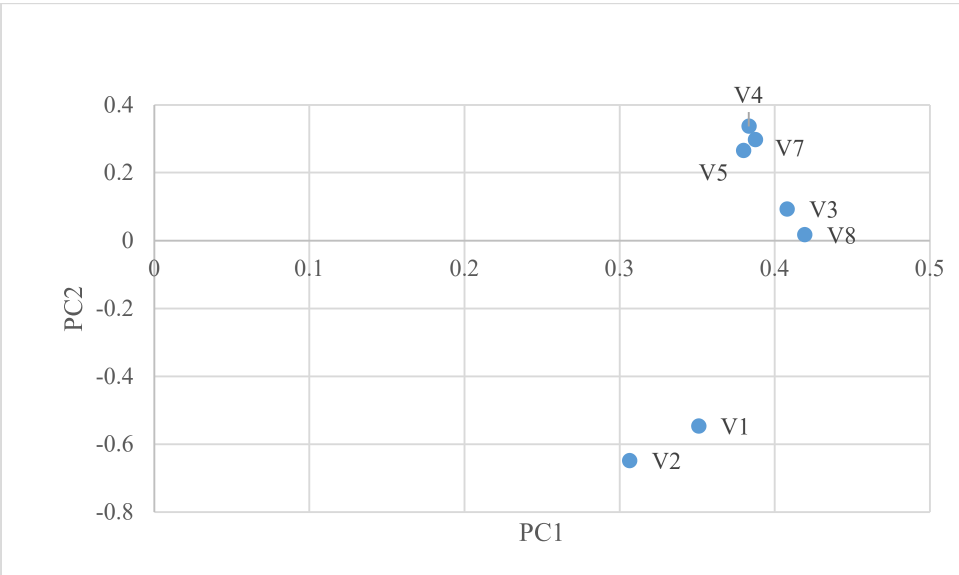

Figure 19: Scattering plot of the projection of each Var in PC1 (horizontal) and PC2 (vertical). The sample was composed of 1571 individuals.

The clustering observed in Figure 19 seems to be coherent with the analogue clustering related to SVRT/CAROLS (2008) dataset because it is explained by land use (and consequently OM), i.e., VAS2 (V2) and VAS1 (V1) are both vineyards, VAS3 (V3) and VAS8 (V8) are both related to scarce vegetation, i.e., at VAS3 were installed between vineyard plantation lines

(Figure 20) and at VAS8 (V8) there were almond trees. The other clusters, VAS5 (V5) and VAS7 (V7), are directly linked to the matorral vegetation, whereas VAS4 (a different type of SSM sensor, Stevens probe, in contrast with the VAS3 driven by a Delta-T ThetaProbe) was installed under the MELBEX-II tower in a clearer area between vineyards rows and, certainly, close to a matorral area.

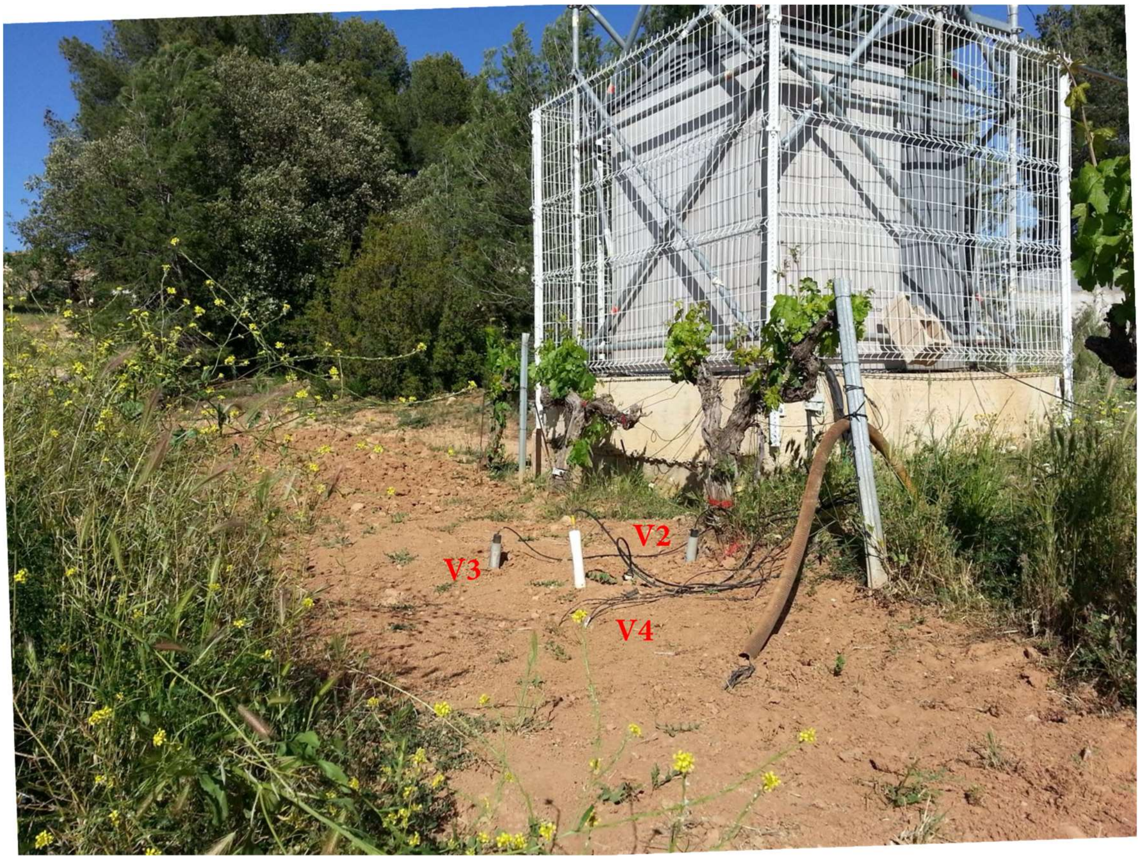

Figure 20: Details of in-field VAS2 (Delta-T ThetaProbe close to the vine stump), VAS3 (Delta-T ThetaProbe between rows) and VAS4 (Stevens probe between rows close to the matorral) sensors at MELBEX-II experimental setup

3.2. SOIL MOISTURE PREDICTION IN THE VAS1 AREA TAKEN AS A REFERENCE

Focusing on the M1SV, the R value inspection suggested that, among several options, an artificial neural network using the Levenberg–Marquardt learning algorithm based on SSM recovery from Sentinel-1 SAR data for this site is preferable. Among other options, the so-called ensemble tree regression model also presented relevant results. These tests allowed us to better understand the complexity of the relationship between SAR backscatter and SSM at this VAS1 site.

The best option offered by the Neural Fitting Tool and Regression Learner app was the ANN-Levenberg–Marquardt algorithm (see the MATLAB® scatterplot with R= 0.86 and RMSE = 0.065 m^3/m^3). The best and best fitting model obtained with the regression learner app was an ensemble tree regression model with R=0.68 and RMSE= 0.038 m^3/m^3 (see the MATLAB® scattering plot in Figure 21).

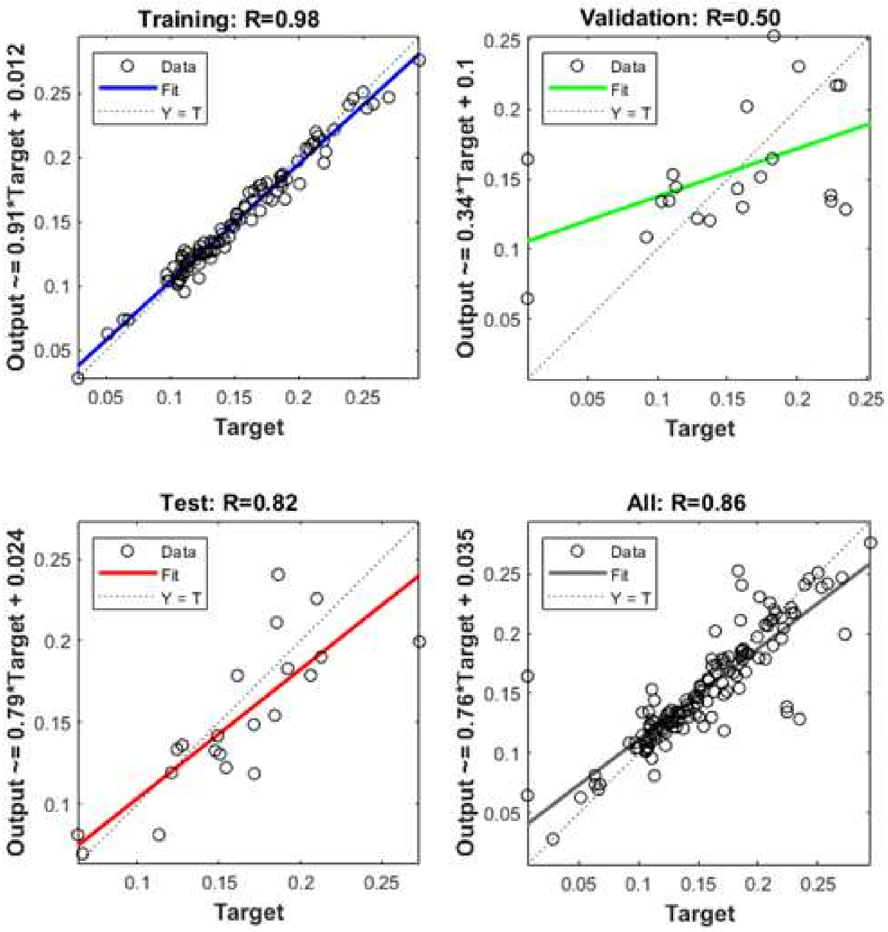

Figure 21: The best option offered with respect to the predicted x true response offered by MATLAB®´s neural fitting tool ANN-Levenberg–Marquardt approach with R= 0.86 and RMSE = 0.065 m^3/m^3 considering the 143 available samples. The target refers to the in-situ measurements, and the output refers to the respective values predicted by this machine learning model.

The regression equations obtained for each stage were summarized, and the R values were calculated (Table 8).

Table 8: Regression models related to M1SV.

Output (O) × Target (T) relation	Stage	R
O ~ = 0.91T+0.012	Training	0.98
O ~ = 0.34T+0.1	Validation	0.50
O ~ = 0.79T+0.024	Test	0.82
O ~ = 0.76T+0.035	All/general	0.86

A dispersion plot was generated concerning the best option offered by the regression learner app, the Ensemble tree (Figure 22).

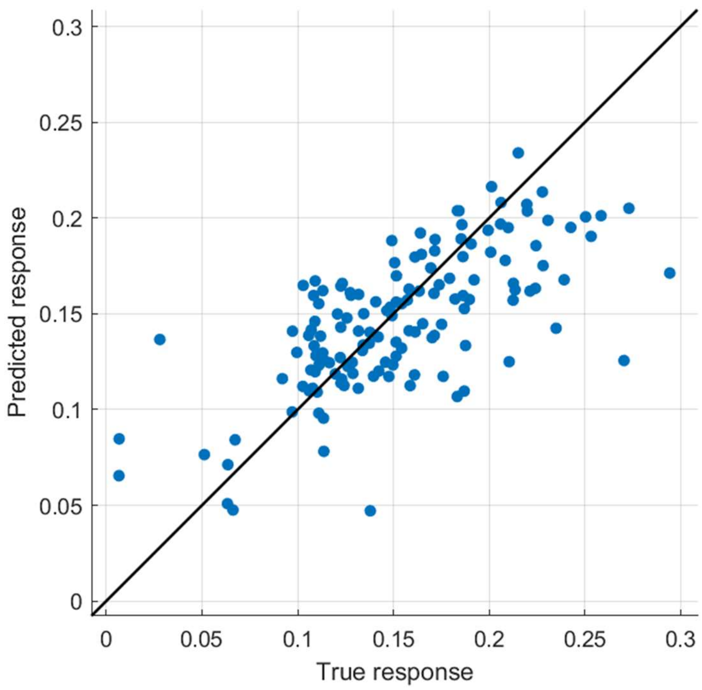

Figure 22: Best option offered with respect to the predicted vs true response by the Regression Learner App as output: the ensemble tree regression model provided R = 0.68 and RMSE = 0.038 m³/m³. This model was obtained by a 5-fold cross-validation process and was parameterized using a minimum leaf size = 4, number of learners = 37 and learning rate = 0.08.

Figure 23 shows now the M1SV plot of the target (in-situ) and the predicted SSM for the same observation dates. This figure shows that a value of R=0.86 indicates the feasibility of the neural network based M1SV as a tool for predicting the dynamics and magnitude of SSM in VAS1.

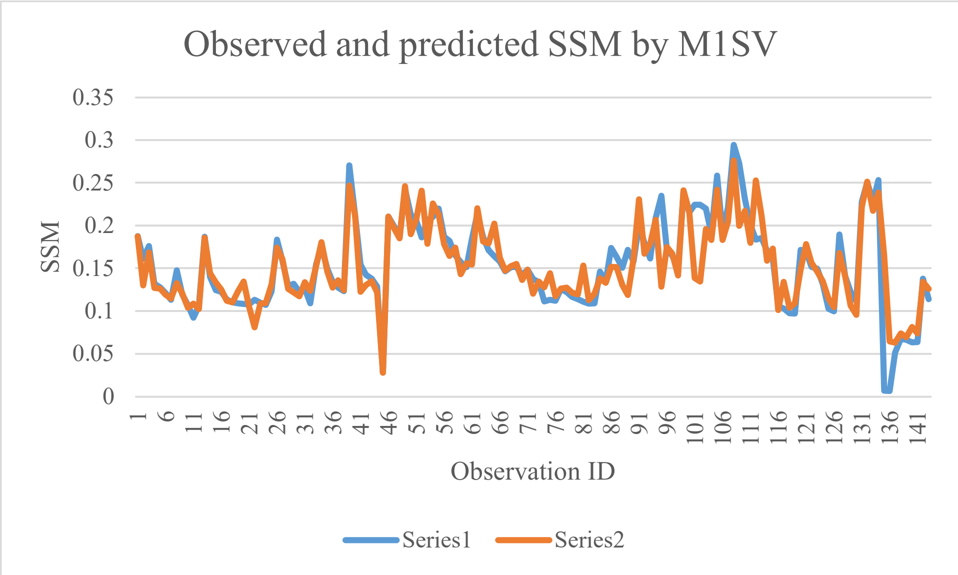

Figure 23: Plot of the target (in situ) and predicted SSMs for the same dates. Series 1 – in situ observation VAS1, Series 2 predicted by M1SV.

Santi et al. (2019), and references therein, claim that the artificial neural network is a method related to minimizing the statistical variance, with advantages regarding the trade-off between retrieval accuracy and computational cost. Additionally, satellite microwave data (active remote sensing) are considered good potential options for soil moisture retrieval. These authors considered that the ANN approach was suitable for merging polarimetric SAR parameters among other possibly different data sources. However, these authors also caution that soil moisture retrieval is indeed a complicated task due to the relationships between σ° and observations that can be affected by surface roughness and vegetation effects despite other sources of nonlinearity. The use of different polarizations can be useful for overcoming these difficulties, for example, co-polarized and cross-polarized SAR data, because they offer additional geometrical information about the target observed.

These two high R values obtained in this work are greater than those obtained using additional inputs, such as the Normalized Difference Vegetation Index (NDVI), incidence angle and thermal infrared and correction for vegetation and roughness of the radar signal, as performed by Alexakis et al (2017), focusing on the SSM stations of TUC, Marathi and Alikampos in Greece. These authors considered that an R= 0.63 for these sites agreed well with ground-based measurements. Furthermore, they suggested that the ANN inversion model is suitable for estimating the relationship between SSM and SAR observations and stressed that, due to its nonparametric nature, little is known about the relationship between the variables in the model. In that study, NDVI was considered the most sensitive model input variable explored. In their paper, the authors also noted that the lack of HH polarization implies the use of Sentinel-1 images only for the SSM model of areas with low vegetation growth and density.

It should be noted that these conditions indeed have similarities with the vegetation present at VAS1, i.e., vineyards. The results of soil moisture modelling on the VAS by Coll (2017) and Khodayar et al. (2019), using a SURFEX (ISBA) SVAT (externalized surface (interaction between soil-biosphere-atmosphere) soil-vegetation-atmosphere-transfer (SVAT) model, yielding R^2 values in the range of 0.7 to 0.95 and a RMSD of ~ 0.02 across all stations in the VAS network, provide strong evidence that the capability of this ANN approach can also handle the complexity of the SSM monitoring station measurements at VAS1 well and can potentially also be extended to the entire set of available SSM stations designed to be representative of the entire VAS site.

The distribution of soil ploughing and plant roots in vineyards and the lack of homogeneity in compaction constitute additional barriers for the extrapolation of soil moisture content predicted from soil moisture probe measurements across the entire VAS field (Monerris & Schumugge, 2009). L-band microwave polarization was not considered relevant for retrieving soil moisture content at the VAS site (Yin et al. 2019), which possibly agrees well with our findings despite the use of VH polarization. This is because in spite of the vines having a preferential orientation towards the south in order to take advantage of the greatest amount of radiation in a homogeneous way for the vines, this orientation is marked by much larger dimensions (spacing between vines approximately 2.5 m and spacing between rows approximately 3 m) than the wavelength of the C-band (approximately 5 cm) and the L-band (approximately 20 cm). Based on the assumption that vegetation and ground roughness are constant for the purpose of L-band radiometry, Yin et al. (2019) used ANN methods to retrieve SSMs at VAS 4 (MELBEX-III) from the Oceanpal Global Navigation Satellite System Reflectometry (GNSS-R) system and obtained R^2=0.90 in comparison with the ELBARA-II passive radiometer with R^2=0.65. Santi et al. (2019) studied the retrieval of soil moisture at a site in Manitoba (Canada) where vegetation effects were not relevant. An analysis using linear SAR σ° and compact polarimetry features allowed them to conclude that the use of artificial neural networks for soil moisture retrieval is promising, with an emphasis on the use of compact polarimetry data as an information source; however, this approach reinforces the implicit site dependence of data-driven approaches.

Li et al. (2020) also performed a study with certain resemblances to this paper, focusing on the use of Sentinel-1 SAR data related to the location for soil moisture monitoring as inputs of feedforward multilayer perceptron neural networks. It was implemented using MATLAB®′s toolbox for the inversion of soil moisture, i.e., a data-driven approach, and compared with two distinct approaches: 1) the results outputted by the inversion/modification of the Dubois model and 2) a multilinear regression-based approach.

The conclusions of these authors supported the use of the ANN approach in such a comparative analysis. The authors also suggested the possibility of using geographical coordinates as inputs to ANNs, and the consideration of a vegetation coefficient could be extended to an area without vegetation.

Zribi et al. (2021) proposed a new reflective index, and a comparison with ground truth data provided by ThetaProbes was considered satisfactory because of the use of Sentinel-1 data from three different sites (Occitania-France, Merguellil, Tunisia, Banizombou, Niger), despite some vegetation corrections. These authors stated that the agreement between remotely sensed data and ground truth data is normally high, but a possible source of discrepancy is the difficulty in capturing information on the full effects of precipitation with a Sentinel-1 acquisition due to the 6-to-12-day repetition cycle intrinsic to the constellation setup.

Merzouki et al. (2019) explored and retrieved SSMs based on RADARSAT constellation mission products related to an eastern Canadian site and obtained a considerable good correlation ($R > 0.70$). Using a method based on change detection, an $R = 0.78$ was achieved. Both were considered to have high RMSEs, possibly due to the small number of ground samples and the relatively small size of the radar signal database. Beale et al. (2019) stated that speckle (of SAR images) and model simplifications are related to errors in retrieving SSM from SAR data and that an LAI (leaf area index) > 0.5 can potentially make it challenge if not impossible to retrieve SSM from C-band SAR.

The work of Robertson et al. (2020) addresses this concern, limitations and potential of the pre-processing step to provide a methodological blueprint for image pre-processing steps to provide a basis for further catalysing the use of SAR data for agricultural use worldwide, considering that few countries use SAR data to map agricultural landscapes. Canada is an exception in this regard and has been consistently incorporating SAR data into agriculture-related mapping activities. Sites from Argentina, Canada, and the USA were used in this research. The sites studied were of different sizes: approximately 3 ha in Canada, 28 ha in Argentina, and approximately 80 ha in the USA. All the sites had the same crop types (soybean, corn, and wheat). The Sentinel-1 and RadarSat-2 data were used as sources of radar data. In this work, four different speckle filtering strategies were tested—adaptive, multitemporal, multiresolution, and nonlocal—used with a variable window configuration. Additionally, the effects of order were tested in the pre-processing steps. Robertson et al. (2020) concluded that the multiresolution Touzi filter was the best option for all sites. The order of the operations did not impact the classification accuracies verified but had effects on the processing time. Their conclusions pointed to the merit of SAR data at least as a source for agricultural inventories for compatible sites worldwide. Considering that crops can potentially affect SSM retrieval, these results suggest positive evidence in favor of the use of SAR for SSM monitoring.

Sentinel-1 data have also been consolidated as a source of information for SSM mapping, including the emergent framework of citizens' science efforts. Specifically, citizen observatories are involved in integration and standardization efforts to interconnect small stakeholders and optimize the use of available monitoring infrastructure, allowing for a realistic understanding of the key SSM variable at each site, as reported by Kibirige et al. (2020). Their study site was the Hungarian village of Tard (western county of Borsod-Abaúj-Zemplén).

Sentinel-1 data were used in conjunction with the Landsat-derived NDVI and low-cost ground-truth data. The results obtained concerning the soil moisture distribution in the landscape using low-cost monitoring strategies were considered satisfactory. In this context, Sentinel-1 data provided high spatial resolution data and were useful for ensuring the continuity of low-resolution data, such as those embedded in long time series observations produced by sensors such as Advanced SCATterometer (ASCAT) and SMOS, as noted by Zribi et al. (2020).

The joint analysis of both the PC1 and PC2 sets suggested that among all the variables analysed, organic matter could be used as an indirect indicator of SSM variation for geographic extrapolation of SSM. The strong relationship between soil organic matter and land cover has been settled, and this finding resonates with the TVDI feature for SSM inference via remote sensing.

3.3. GEOGRAPHICAL EXTRAPOLATION OF SOIL MOISTURE TO OTHER SITES IN THE VAS FROM THE VAS1 REFERENCE.

Using this M1SV framework as a basis, the feasibility of the second module, the so-called M2GE, can also be deduced. After the analogue and typical training stage, a 35 hidden neuron shallow network (multilayer feed-forward perceptron) was deployed and trained using the Levenberg–Marquardt learning algorithm setup endowed with a tansig nonlinear activation function at the hidden layer and a linear activation function at the output layer, as represented in Figure 24, which gave results with an RMSE of 0.024 m^3/m^3. The Levenberg–Marquardt learning algorithm was used in the best model, although Bayesian regularization and scale-conjugated gradients were also explored as possible options. Reinforcing, schematically, the M2GE was obtained as follows (Figure 24):

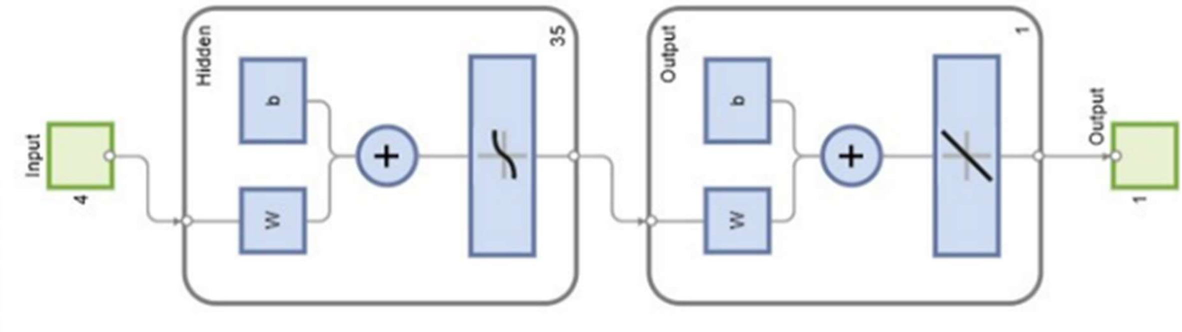

Figure 24: Schematics of the M2GE.

The training, validation and test steps are described more in detail by linear regression (Figure25). In this graph, the targets are the values inferred in situ, and the outputs are the respective values devised by M2GE processing. These graphs represent a step-by-step visualization of the results of each stage of the MATLAB® workflow toward deploying a multilayer perceptron neural network. The split of the available dataset creates subsets of data, and all of them are named the target for the training, validation and test processes. Such an organization allows them to construct different linear regressions, as shown below.

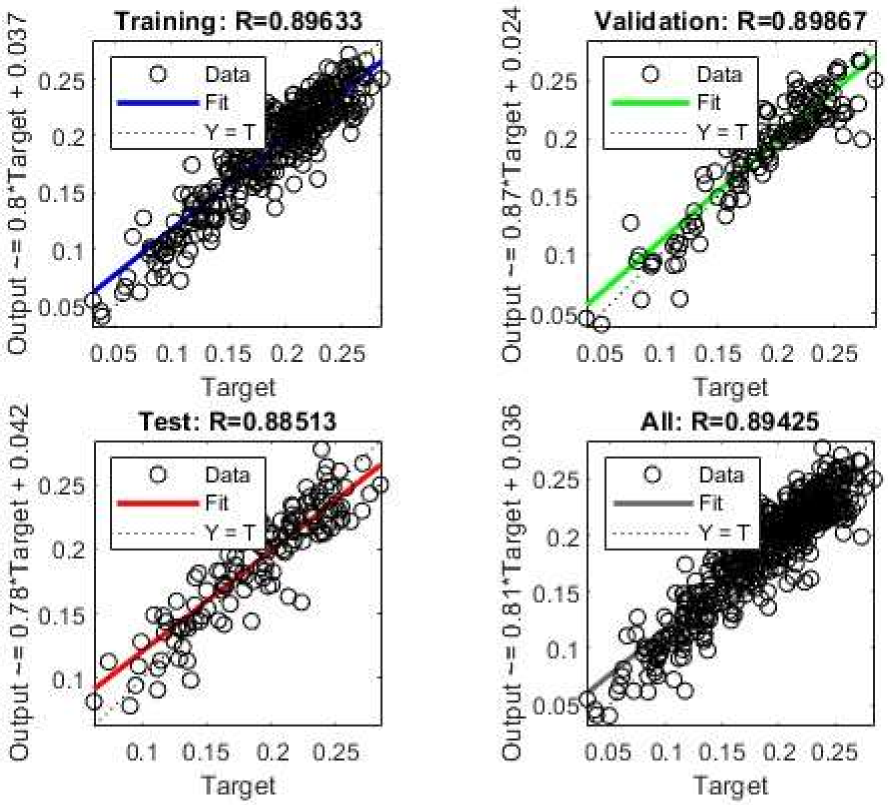

Figure 25: Typical performance analysis of ML methods by means of multiple linear regression related to the deployment process on this kind of shallow neural network.

The regression equations obtained for each stage were summarized, and the results were analysed with respect to R (Table 9).

Table 9: Regression models related to M2GE.

Output (O) vs Target (T) relation	Stage	R
O ~= 0.8T+0.037	Training	0.87
O ~= 0.87T+0.024	Validation	0.90
O ~= 0.78T+0.042	Test	0.88
O ~= 0.81+0.036	All/General	0.89

It is worth noting that the R value obtained for M2GE (R = 0.89) was greater than that for M1SV (R = 0.86), suggesting the viability of this proposed geographical extrapolation of

M1SV for the 10 km × 10 km control area. To evaluate the M2GE in more depth, it was challenging to predict an hourly dataset of VAS8 SSM in situ values that were not used for the M2GE training, validation, or test steps, as described above. Despite this contrast between the challenging hourly dataset and the dataset used for training (daily values), M2GE model performed fairly well (Figures 26 and 27).

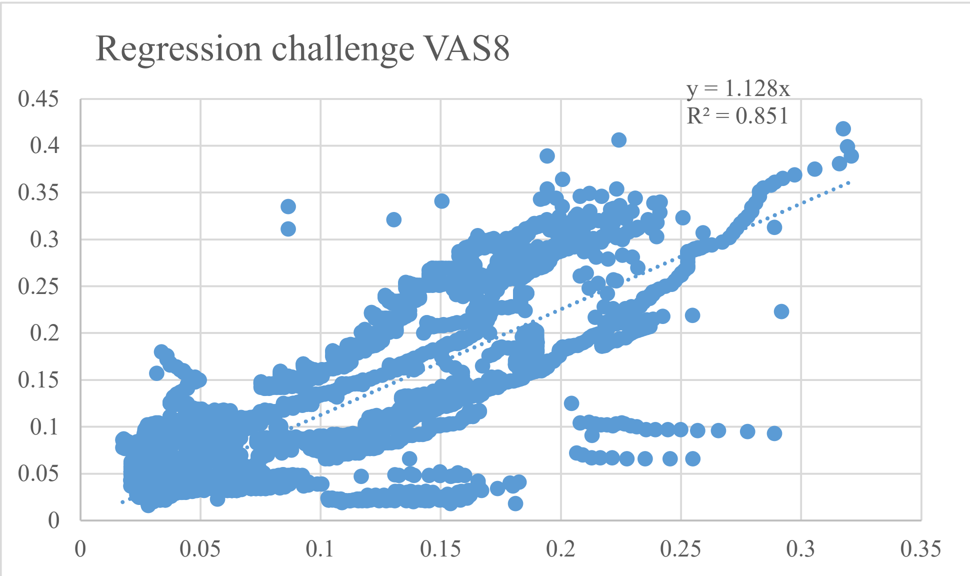

Figure 26: Regression plot obtained for the VAS-8 challenge for M2-GE.

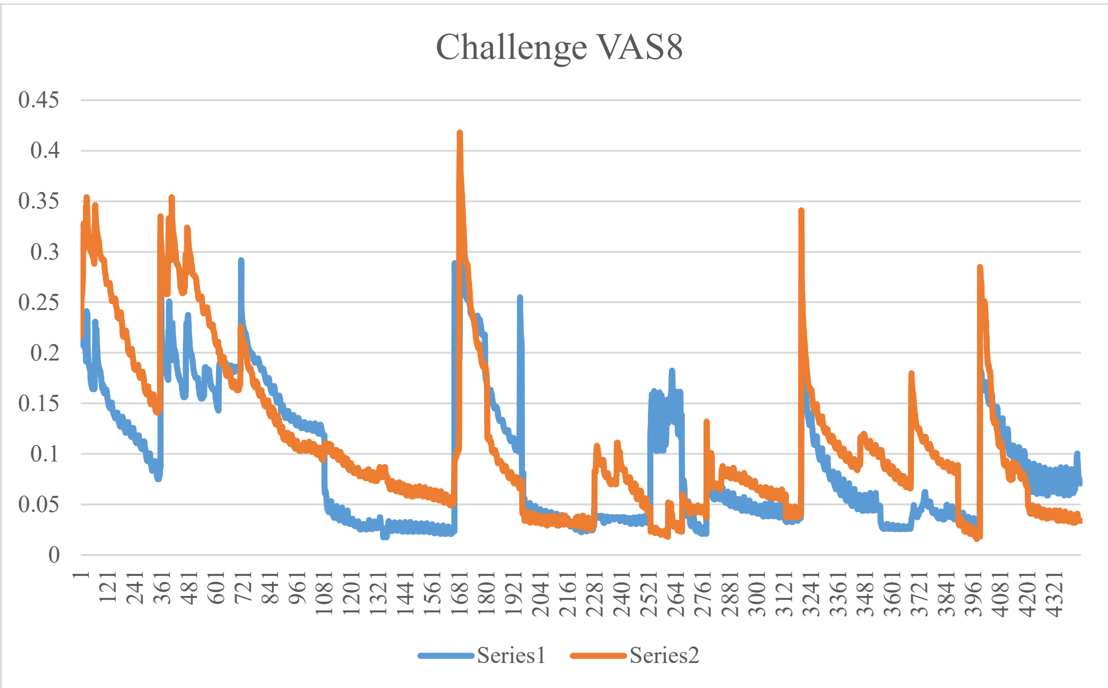

Figure 27: Plot of the target (in situ) SSM and predicted SSM for the same dates. Series 1 – in situ observation VAS1, Series 2 predicted by M1SV.

It is important to emphasize that both M1SV and M2GE have ~120 m × 120 m spatial resolution, which, in this case, was verified with point in situ probe measurements, most likely because of their more dynamic characteristics.

Another challenging test for M2GE was performed using only the VAS5 dataset; despite its presence in the training dataset, M2GE was negatively correlated with the VAS1 score (Figures 28 and 29).

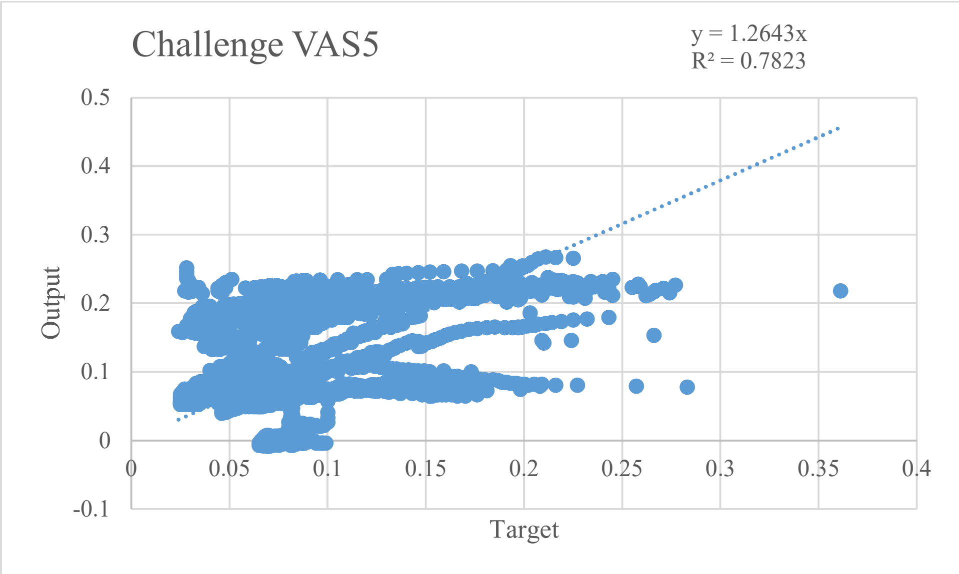

Figure 28: Regression plot obtained for the VAS-5 challenge for M2-GE.

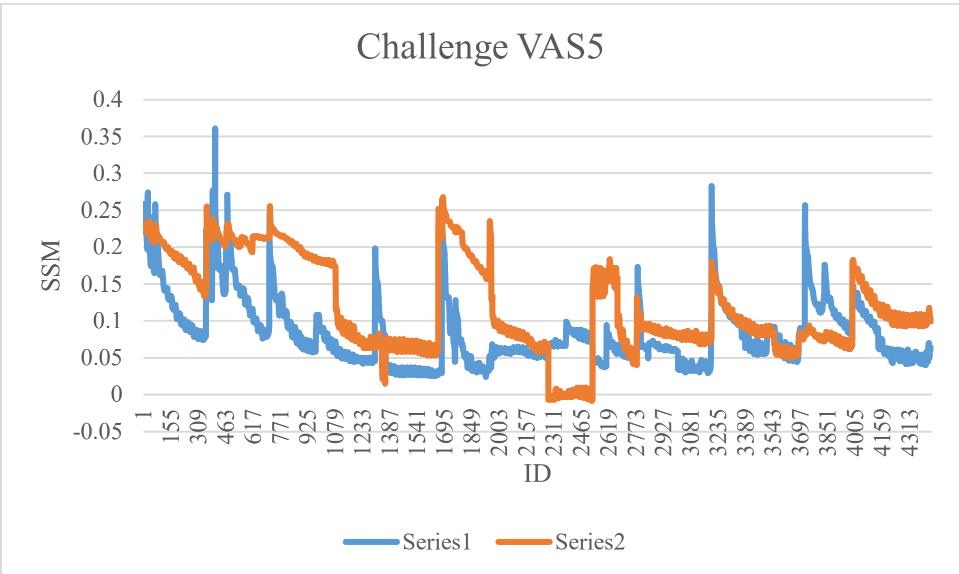

Figure 29: Plot of the target (in situ) SSM and predicted SSM for the same dates. Series 1 – in situ observation VAS5, Series 2- values predicted by M2GE.

The results obtained for VAS5 corroborate the lower (comparatively to other VASs analysed using PC) correlation between VAS1 and VAS5. Regarding amplitude, an average positive bias of 26% was verified, although the modelling of the dynamics of the process was more satisfactory. VAS5 modelling is challenging because of the variety of issues to be attacked, especially some inversions of peaks that are difficult to understand at this point and possibly strongly connected with the *matorral* type of land cover verified at the VAS5 location.

This type of averaged approach, encoded in the contrast of ~120 m ×120 m spatial resolution versus a comparison of point in situ probe measurements, probably contributes to positive biases of 12% and 26% (inferred at the linear regression coefficient) for both challenges VAS8 (V8) and VAS5 (V5), respectively. It remains unclear whether other drivers could be considered at this point.

These prospective and viability results suggest that an operational and reliable system for monitoring SSM (with ~120x 120 m spatial resolution) in the 10 km×10 km control area of the VAS driven by SAR Sentinel-1 products and ML methods could be reliable.

CHAPTER 4

CONCLUSIONS

4. CONCLUSIONS

This thesis delved into two innovative avenues: 1) prospecting the net-effect VAS dataset and 2) inferring the viability of Sentinel-1 SAR data to drive soil moisture retrieval and predict surface soil moisture (SSM) with less than 1 km spatial resolution (~120 m x 120 m) within the Valencia Anchor Station. Such an evaluation is based on a two-stage machine learning method which produces an R = 0.86 when predicting SMM at VAS1 using neural-network and Sentinel-1 data. For geographical extrapolation, the result obtained was R= 0.89 besides the fair performance in the prediction challenges for VAS8 (R = 0.92) and VAS5 (R = 0.88), notably a matorral area which features tends to be much more complex than those of vineyards. Jointly considered these numerical results strongly suggest that this Sentinel-1 approach is viable for the VAS 10 km x 10k m control area. The good R-values obtained in the fit of the measured values in situ and the fact that there are many different sources of VAS data suggests that the VAS data corpus can be endowed, beyond self-consistency, with the net effect, which is a promising future avenue for research.

The capacity building feature of this thesis is the fact that it is mainly based on simple and shallow neural networks, established and widespread MODIS/TVDI methodologies, freely available software and a reduced need to handle the C-band radar data set as well as relying heavily on software/computing, which makes this methodological approach suitable for worldwide use, especially in low-income countries with limited availability of instruments and human resources to establish large networks of instruments operating for a long range of similar VAS.

Future improvements and potential enhancements: by taking advantage of the vast corpus of VAS data, a state-of-the-art generative multimedia machine learning and natural language input model could be developed based on this data. New time Series analysis: Integrating long-term soil moisture time series from around the control area could significantly improve the robustness and generalizability of SSM predictions within the inner region. Addressing the nonlinearities associated with soil roughness and vegetation dynamics remains a key challenge for ANN-based retrievals. Exploring alternative data sources, such as C-band SAR data from Unmanned Aerial Vehicles (UAVs), could offer greater realism and support better land use management strategies. Digital Twins for soil moisture/remote sensing training: Such a model could act as a "digital twin" of the VAS, accepting multimedia data as input. This could create a "Citizen's Observatory" configuration, promoting soil moisture/remote sensing science training and education, similar to the "school catchment" concept in hydrology.

Global Knowledge Sharing: Researchers and practitioners worldwide could benefit from this "robot" by comparing their local study sites to the VAS and interacting with it using natural language and multimedia data, gaining profound insights into their own SSM and remote sensing challenges. Overall, this thesis advances the field of soil moisture retrieval from Sentinel-1 for the VAS site, revealing the promise of ANNs and net-effect exploration. The proposed "Citizen's Observatory" concept further unlocks opportunities for global knowledge sharing and educational initiatives in soil moisture remote sensing studies.

RFERENCES

Alexakis, D. D., Mexis, F. D. K., Vozinaki, A. E. K., Daliakopoulos, I. N., & Tsanis, I. K. (2017). Soil moisture content estimation based on Sentinel-1 and auxiliary earth observation products. A hydrological approach. *Sensors*, *17*(6), 1455.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., & Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing*, *7*(12), 16398-16421.

Attema, E., Davidson, M., Floury, N., Levrini, G., Rosich, B., Rommen, B., & Snoeij, P. (2008, June). Sentinel-1 ESA's new European radar observatory. In *7th European conference on synthetic aperture radar* (pp. 1-4). VDE.

Bauer-Marschallinger, B., Freeman, V., Cao, S., Paulik, C., Schaufler, S., Stachl, T., ... & Wagner, W. (2018). Toward global soil moisture monitoring with Sentinel-1: Harnessing assets and overcoming obstacles. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(1), 520-539.

Beale, J., Snapir, B., Waine, T., Evans, J., & Corstanje, R. (2019). The significance of soil properties to the estimation of soil moisture from C-band synthetic aperture radar. *Hydrology and Earth System Sciences Discussions*, 1-32.

Cano, A., Añón,J., Reig, C., Millán-Scheiding, C., & López-Baeza , E. (2012)), "Automated Soil Moisture Monitoring Wireless Sensor Network for Long-Term Cal/Val Applications," Wireless Sensor Network, Vol. 4 No. 8, 2012, pp. 202-209. doi: 10.4236/wsn.2012.48030.

Cano, A., Lopez-Baeza, E., Anon, J. L., Reig, C., & Millán-Scheiding, C. (2007, October). Wireless sensor network for soil moisture applications. In *2007 International Conference on Sensor Technologies and Applications (SENSORCOMM 2007)* (pp. 508-513). IEEE.

Carbó, E., Juan, P., Añó, C., Chaudhuri, S., Diaz-Avalos, C., & López-Baeza, E. (2021). Modeling Influence of Soil Properties in Different Gradients of Soil Moisture: The Case of the Valencia Anchor Station Validation Site, Spain. *Remote Sensing*, *13*(24), 5155.

Dingle Robertson, L., Davidson, A., McNairn, H., Hosseini, M., Mitchell, S., De Abelleyra, D., ... & Cosh, M. H. (2020). Synthetic Aperture Radar (SAR) image processing for operational space-based agriculture mapping. *International Journal of Remote Sensing*, *41*(18), 7112-7144.

Ezzahar, J., Ouaadi, N., Zribi, M., Elfarkh, J., Aouade, G., Khabba, S., ... & Jarlan, L. (2019). Evaluation of backscattering models and support vector machine for the retrieval of bare soil moisture from Sentinel-1 data. *Remote Sensing*, *12*(1), 72.

Fernandez-Moran, R., Wigneron, J. P., Lopez-Baeza, E., Salgado-Hernanz, P. M., Mialon, A., Miernecki, M., ... & Kerr, Y. H. (2014, July). Evaluating the impact of roughness in soil moisture and optical thickness retrievals over the VAS area. In *2014 IEEE Geoscience and Remote Sensing Symposium* (pp. 1947-1950). IEEE.

Gao, Q., Zribi, M., Escorihuela, M. J., & Baghdadi, N. (2017). Synergetic use of Sentinel-1 and Sentinel-2 data for soil moisture mapping at 100 m resolution. *Sensors*, *17*(9), 1966.

GBOV, 2021 https://gbov.acri.fr/https://doi.org/10.1016/j.rse.2019.111402.

Ghajarnia, N., Kalantari, Z., Orth, R., & Destouni, G. (2020). Close co-variation between soil moisture and runoff emerging from multi-catchment data across Europe. *Scientific reports*, *10*(1), 4817.

Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). The role of artificial intelligence and data network effects for creating user value. Academy of management review, 46(3), 534-551.

Hossain, M. R. H., & Kabir, M. A. (2023). Machine Learning Techniques for Estimating Soil Moisture from Smartphone Captured Images. *Agriculture*, *13*(3), 574.

Khodayar, S., Coll, A., & Lopez-Baeza, E. (2019). An improved perspective in the spatial representation of soil moisture: Potential added value of SMOS disaggregated 1 km resolution "all weather" product. *Hydrology and Earth System Sciences*, *23*(1), 255-275.

Kibirige, D., & Dobos, E. (2020). Soil moisture estimation using citizen observatory data, microwave satellite imagery, and environmental covariates. *Water*, *12*(8), 2160.

Kobiyama, M., Checchia, T., Corseuil, C. W., Lino, J. F., Lopes, N. H., Grison, F., ... & Basso, S. (2007). Forest hydrology project (UFSC-MOBASA) for water resources management in Rio Negrinho City, Santa Catarina, Brazil. *Changes in Water Resources Systems: methodologies to maintain water security and ensure integrated management*, 250-257.

Li, Y., Yan, S., Chen, N., & Gong, J. (2020). Performance evaluation of a neural network model and two empirical models for estimating soil moisture based on Sentinel-1 SAR data. *Progress In Electromagnetics Research C*, *105*, 85-99.

Lopez-Baeza, E., Antolin, M. C., Balling, J., Belda, F., Bouzinac, C., Buil, A., ... & Domenech, C. (2009). The ESA SMOS Mission: Validation Activities at the Valencia Anchor Station. In *Advances in studies on desertification, Murcia 16-18 September, 2009*.

Lopez-Baeza, E., Garcia Rodriguez, D., Albero Peralta, E., Garcia-Celda, A., Asensi Ortega, V., Catalan Alcober, D. J., & Martinez Dura, J. J. (2022, May). The Valencia Anchor Station: 20 Years of Uninterrupted Scientific Activity on Validation of Low and Middle Resolution Earth Observation Remote Sensing Data and Products. In EGU General Assembly Conference Abstracts (pp. EGU22-6685).

Lopez-Baeza, E., Albero Peralta, E., Asensi-Ortega, V., Benlloch Casaban &. Martinez Diaz, B. (2021). Ground-Based Observations for Validation (GBOV) of Copernicus Global Land Products - Algorithm Theoretical Basis Document -Soil Moisture Products (LP-6, RM-10 and RM-11). Ref.: GBOV-ATBD-RM-10-RM-11-LP-6 - Version: 2.2 - Date: 15/09/2021 - Framework Contract ref.: 932059.

Lopez-Baeza, E., Garcia Rodriguez, D., Albero Peralta, E., Garcia-Celda, A., Asensi Ortega, V., Catalan Alcober, D. J., & Martinez Dura, J. J. (2022, May). The Valencia Merzouki, A., McNairn, H., Powers, J., & Friesen, M. (2019). Synthetic aperture radar (SAR) compact polarimetry for soil moisture retrieval. *Remote Sensing*, *11*(19), 2227.

Mamirua, 2023; https://mamiraua.org.br/downloads/programas

Monerris, A., Schmugge, T., & Jedlovec, G. (2009). Soil moisture estimation using L-band radiometry. *Advances in Geoscience and Remote Sensing*, (21), 441-462.

Montzka, C., M. Cosh, B. Bayat, A. Al Bitar, A. Berg, R. Bindlish, H. R. Bogena, J. D. Bolten, F. Cabot, T. Caldwell, S. Chan, A. Colliander, W. Crow, N. Das, G. De Lannoy, W. Dorigo, S. R. Evett, A. Gruber, S. Hahn, T. Jagdhuber, S. Jones, Y. Kerr, S. Kim, C. Koyama, M. Kurum, E. Lopez-Baeza, F. Mattia, K. McColl, S. Mecklenburg, B. Mohanty, P. O´Neill, D. Or, T. Pellarin, G. P. Petropoulos, M. Piles, R. H. Reichle, N. Rodriguez-Fernandez, C. Rüdiger, T. Scanlon, R. C. Schwartz, D. Spengler, P. Srivastava, S. Suman, R. van der Schalie, W. Wagner, U. Wegmüller, J.-P. Wigneron, F. Camacho, and J. Nickeson (2020): Soil Moisture Product Validation Good Practices Protocol Version 1.0. In: C. Montzka, M. Cosh, J. Nickeson, F. Camacho (Eds.): Good Practices for Satellite-Derived Land Product Validation (p. 123), Land Product Validation Subgroup (WGCV/CEOS), doi:10.5067/doc/ceoswgcv/lpv/sm.00

Olivas, E. S., Guerrero, J. D. M., Martinez-Sober, M., Magdalena-Benedito, J. R., & Serrano, L. (Eds.). (2009). Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques. IGI global.

Prado, P. F., Duarte, I. C. S. (2020). An overview and recent advances in fuzzy ARTMAP classifier usage for mapping purposes using remotely sensed data. Environmental Informatics Letters, 3(2), 86-97.

Prado, P. F., Duarte, I. C. S. (2020). Fuzzy ARTMAP driven features selection: towards a system to design information fusion for the geographic discovery. Journal of Environmental Informatics Letters. 10.3808/jeil.202000025.

Rasheed, M. W., Tang, J., Sarwar, A., Shah, S., Saddique, N., Khan, M. U., ... & Sultan, M. (2022). Soil moisture measuring techniques and factors affecting the moisture dynamics: A comprehensive review. *Sustainability*, *14*(18), 11538.

Santi, E., Dabboor, M., Pettinato, S., & Paloscia, S. (2019). Combining machine learning and compact polarimetry for estimating soil moisture from C-band SAR data. *Remote Sensing*, *11*(20), 2451.

Sentinel Hub (2023) https://www.sentinel-hub.com/

Sentinel-1 (2023), https://sentinel.esa.int/web/sentinel/missions/sentinel-1/overview

Singh, A., Gaurav, K., Sonkar, G. K., & Lee, C. C. (2023). Strategies to measure soil moisture using traditional methods, automated sensors, remote sensing, and machine learning techniques: review, bibliometric analysis, applications, research findings, and future directions. *IEEE Access*.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., ... & Rostan, F. (2012). GMES Sentinel-1 mission. *Remote sensing of environment*, *120*, 9-24.

Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote sensing of environment*, *236*, 111402.

Werner, W. J., Sanderman, J., & Melillo, J. M. (2020). Decreased soil organic matter in a long-term soil warming experiment lowers soil water holding capacity and affects soil thermal and hydrological buffering. *Journal of Geophysical Research: Biogeosciences*, *125*(4), e2019JG005158.

Wigneron, J. P., Schwank, M., Baeza, E. L., Kerr, Y., Novello, N., Millan, C., ... & Mecklenburg, S. (2012). First evaluation of the simultaneous SMOS and ELBARA-II observations in the Mediterranean region. *Remote Sensing of Environment*, *124*, 26-37.

Yin, C., Lopez-Baeza, E., Martin-Neira, M., Fernandez-Moran, R., Yang, L., Navarro-Camba, E. A., ... & Yang, D. (2019). Intercomparison of soil moisture retrieved from GNSS-R and from passive L-band radiometry at the Valencia Anchor Station. *Sensors*, *19*(8), 1900.

Zribi, M., Albergel, C., & Baghdadi, N. (2020). Editorial for the special issue “Soil moisture retrieval using radar remote sensing sensors”. *Remote Sensing*, *12*(7), 1100.

Zribi, M., Foucras, M., Baghdadi, N., Demarty, J., & Muddu, S. (2020). A new reflectivity index for the retrieval of surface soil moisture from radar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 818-826.

APPENDIX

Two On-going project of signal processing with connections soil moisture inference done during doctorate period:

1) **Smartphone base soil moisture inference using colours: initial prototyping of a low-cost tool**

As a low-cost process for estimating SSM, several image processing and smartphone-driven solutions have been evaluated that could offer benefits for many users worldwide, especially due to their low cost (see Hossian and Kabir,2023). Moreover, it works synergistically with remote sensing tools to improve precision in defining soil moisture for example.

The MATLAB® GUI/code designed as follows:

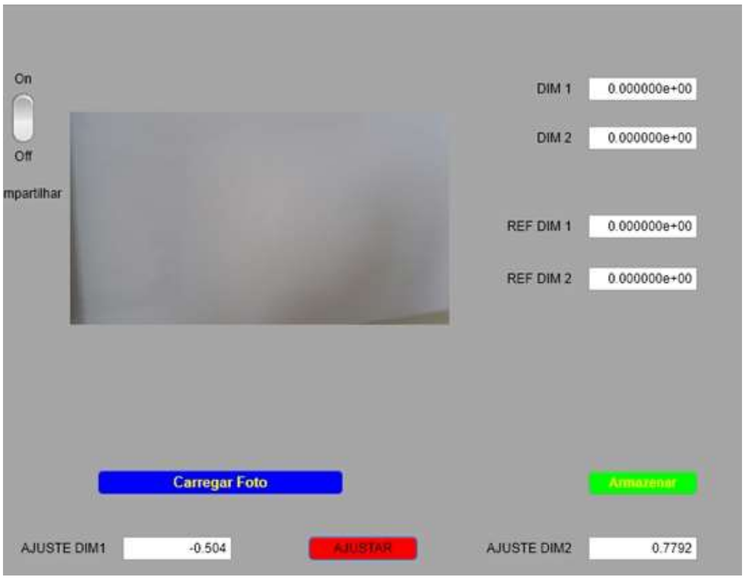

Figure appendix 1 - Usuary interface of a working progress to prototype a complementary Smartphone driven tool to infer about SMM.

2) Perspectives about applying multi-fractal analysis in soil moisture studies:

A work done by this author hints a secondary utilization of such a approach for SMM to give insights about multiple process occurring on soil watering:

ELSEVIER

Sensing and Bio-Sensing Research

Volume 24, June 2019, 100285

Using Matlab's wavelet toolbox to compare electric signals outputted by microbial fuel cells

P.F. Prado[a], M.A.G. Mayen[b], G.F. Silva[b], I.C.S. Duarte[b]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.sbsr.2019.100285

Get rights and content

Under a Creative Commons license

open access

Abstract

Motivation: microbial fuel cells (MFC) represents a wastewater treatment technology with the potential for a relevant electric energy generation. The monitoring of the electric current outputted generates time series of data. It was generated two time series using an experimental setup with a low-cost dual chamber microbial fuel cell treating cassava wastewater and, also, it was used the electric current data from a previously described experiment with this same setup treating cheese whey. All of this, innovatively, according our best knowledge, was studied using power spectral density, multifractal and wavelet coherence analysis. Results are promising and indeed point out this approach as an under-considered tool towards gaining knowledge about biofilms attached on anodes of a type of microbial fuel cell reactor.

Figure appendix 2 – Abstract of recent multi-fractal and wavelet methods for exploring electric signals in biotechnology that possible could also be used for in depth understanding (fine detailing) of soil watering dynamics as further research investigation.