D.4.4: Implementation and development of systems; SSS, data processing and geo-tagged photos framework alpha versions (final version)

December/2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870378.

Author(s)/Organisation(s)	Georgios Galanis, Nikolaos Tsakiridis, Nikolaos Tziolas, Konstantinos Karyotis (i-BEC)
Contributor(s)	Orestis Sampson, Nikos Iliakis, Valantis Tsiakos (ICCS)
Work Package	WP4
Delivery Date (DoA)	31/12/2021
Actual Delivery Date	31/12/2021
Abstract:	The deliverable constitutes a report describing the technical details from the implementation and development of the final versions of the DIONE in-situ (WP4) components aiming to establish a complete ecosystem of low-cost and easy to use smart tools and techniques for the accumulation of the in-situ information. In particular the report focuses on (i) the microelectromechanical systems (MEMS) based sensor, the measurement protocol, the smartphone application, and the security protocols, (ii) the geo-tagged photos framework iii) the data pre-processing system, including the utilized DBMS, backend processes for data integration and validity, and techniques used for outlier detection iv) tools and methodologies for the production of spatially explicit maps of key indicators for land degradation by combining historical Earth Observation (EO) data with MEMS captures.

Document Revision History			
Date	**Version**	**Author/Contributor/ Reviewer**	**Summary of main changes**
15/11/2021	V0.1	i-BEC	Initial ToC
05/12/2021	V0.6	i-BEC	Submitted for internal review
28/12/2021	V0.9	i-BEC	Small modifications according to reviewers' feedback
31/12/2021	V1.0	ICCS	Approved, final version submitted

Dissemination Level		
PU	Public	x
CO	Confidential, only for members of the consortium (including the EC)	

DIONE Consortium			
Participant Number	**Participant organisation name**	**Short name**	**Country**
1	INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS	ICCS	EL
2	DIABALKANIKO KENTRO PERIBALLONTOS	i-BEC	EL
3	SINERGISE LABORATORIJ ZA GEOGRAFSKE INFORMACIJSKE SISTEME DOO	SINERGISE	SI
4	CORE INNOVATION AND TECHNOLOGY OE	CORE	EL
5	NATIONAL PAYING AGENCY	NMA	LT
6	INOSENS DOO NOVI SAD	INO	RS
7	GILAB DOO BEOGRAD PALILULA	GILAB	RS
8	CYPRUS AGRICULTURAL PAYMENTS ORGANISATION	CAPO	EL

Funding Scheme: Innovation Action (IA) ● Theme: DT-SPACE-01-EO-2018-2020

Start date of project: 01 January 2020 ● Duration: 30 months

Table of contents

List of tables

List of figures

List of Abbreviations and Acronyms	
Abs	Absorbance
CLHS	Conditioned Latin Hypercube Selection
CNN	Convolutional Neural Network
DBMS	DataBase Management System
EO	Earth Observation
ESDAC	European Soil Data Centre
GUI	Graphical User Interface
ISS	Internal Soil Standards
LUCAS	Land Use and Cover Area frame Survey
MAE	Mean Absolute Error
MEMS	Microelectromechanical systems
ODC	Open Data Cube
PLSR	Partial Least-Squares Regression
Ref	Reflectance
RF	Random Forests (RF)
RMSE	Root Mean Squared Error
SG	Savitzky-Golay
SNV	Standard Normal Variates
SSO	Single Sign On
SSL	Soil Spectral Library
SSS	Soil Scanning System
SVM	Support Vector Machines

TLS	Transport Layer Security
USDA	United States Department of Agriculture
VNIR	Visible and Near-Infrared

1 Introduction

1.1 Context and background

The present report contains an extensive description of the work carried out in the context of DIONE project's WP4 for the implementation and development of the DIONE in situ components for complementing EO-data. This entails the development of a set of low-cost and easy to use smart tools that produce a bridging between in situ data accumulation and heterogeneous remote sensing data.

The first part of the toolbox is the Soil Scanning System (SSS) comprising of a MEMS based portable Visible and Near-Infrared (VNIR) soil scanner and a mobile application developed for the operation of the soil scanner. This component aims to the collection of in situ measurements and provide quick estimations of key soil health indices. Along with the reflectance capture, a set of ancillary data and the required meta-data are recorded, aiming to provide ground truth soil observations of key parameters that will later converted to key soil properties via machine learning techniques.

The second component pertains to the use of geo-tagged photos which are captured using smartphones and allow for an improved method for the provision of additional evidence regarding the Common Agricultural Policy (CAP) compliance monitoring while simultaneously ensuring that the process is untampered. The added benefit is the quality and trust of the transmitted data as well as the application characteristics with respect to location accuracy and data collection process through the use of augmented reality features.

The ensemble of the abovementioned tools forms the DIONE's in situ component and the data collected by which, will populate a DataBase Management System (DBMS) that serves as a data storing and processing system that validates curated data integrity. The DBMS ensures that all data exchange complies with predefined security protocols, ensuring data validity and that only authorized parties may be engaged or produce traffic. Outlier analysis of data is performed to ensure that no erroneous or highly variable records will populate the DBMS while valid records with novel characteristics are recognized via novelty detection mechanism. The system is designed to ensure easy data retrieval and provide user friendly APIs.

The last part of the tools and methodologies described to this report is the production of spatially explicit maps of key indicators for land degradation by combining historical Earth Observation (EO) data with SSS captures. To this end, a collection of openly accessible EO datasets were set together and according to a bottom-up spiking approach that is extensively described in this report were translated to maps of key soil properties.

The main objective of this WP is the development of a set of digital and easy to use tools exploiting automated technologies to ensure more frequent, accurate and inexpensive CAP area-based compliance checks, aiming to the monitoring of CAP compliance and of the soil health through a set of observations that will be turned into spatially explicit maps, providing meaningful information related to soil.

1.2 System's architecture

DIONE toolbox consists of three main components:

i. DIONE Earth Observation component, that aims to harness DIAS/Sentinel Hub to support data management and processing of EO data, area monitoring markers, drones, data fusion, and super resolution techniques towards the provision of enhance resolution maps of permanent pastures, crop-types, non-productive EFA types and farmers' activities (i.e. grassland mowing/ploughing, etc);
ii. DIONE In-situ component, whose components are analysed in the context of this report, containing the SSS, the geo-tagged photos framework, the DBMS and the novel machine learning algorithms to transform the in situ data to spatial explicit maps of key soil properties and indices quantifying the current level of land degradation
iii. DIONE Green Accountability component that includes a compliance monitoring tool, which decides on beneficiaries' compliance and is integrated with the existing tools of paying agencies and an AI-enabled environmental performance tool, accompanied with a visualization engine.

DIONE's components have been developed under common or compatible frameworks, ensuring that high level of interoperability has been achieved, and their outputs are further utilized by other DIONE toolbox components. Key role to this plays the DIONE toolbox API, which supporting data retrieval and storage and integration with external infrastructure, enabling also the gradual access to final data provision.

2 DIONE in situ component

2.1 In situ Soil Scanning System

2.1.1 System description and subcomponents

The implementation of a low-cost system based on portable spectrometers enabling the end-user to acquire a rapid assessment of soil quality was set as a key objective of DIONE's WP4. To this end, a MEMS based soil scanning systems employing microelectromechanical systems (MEMs) technology that is able to acquire and translate Near Infrared soil reflectance to critical agro-environmental – related parameters was developed. The advent of MEMs spectrometers operating inside the VNIR range) enables the rapid and non-destructive measurement of a soil's reflectance spectrum, providing valuable information related to soil texture, SOC concentration, pH and more with the inference of measured top-soil reflectance through supervised ML techniques. The developed system can be connected wirelessly to a mobile device (e.g. smartphone, tablet) and operated by non-experts to accumulate valuable information for soil status in real-time via in situ usage. The VNIR spectra and the associated metadata are securely transmitted using secure transmission and validation protocols, and are stored to DIONE's central database. All data-flow is automatically monitored and erroneous or novel instances are flagged through classification algorithms. Additionally, a mobile application has been implemented serving as a data-mediator between the soil spectrometer and the central database and also enabling the user to operate and amass top-soil reflectance.

2.1.2 Technical specifications of the Soil Spectrometer

At the core of the SSS lies the soil spectrometer which records the in situ soil reflectance. It thus must fulfil some key requirements such as portability and long battery life, operability from standard smartphones and of course the potential to identify key soil properties through the analysis of the covered spectral range by maintaining the cost demand low. Evidently, a low-cost device cannot have the same spectral range and resolution as a laboratory spectrometer; in other words, there is a given trade-off between the technical capabilities of the device and its cost.

The device selected is Spectral Engines Nirone S2.2 sensor which uses a patented Fabry–Pérot interferometer and covers the spectral range from 1750 to 2150 nm (Figure 1), with a spectral resolution of about 18nm. The built-in illumination source is composed from two tungsten vacuum lamps, with estimated life expectancy of more than 40.000 hours. The device is compact with limited dimensions (25x25x17.5mm^3), weighing 150g and is equipped with a 5V re-chargeable battery supply, meeting portability characteristics and able to operate in situ and withstand ambient temperatures between 10° and 50°C. The device further features a Bluetooth connection that can be used to communicate with smartphone applications. The device (Figure 2) was selected among a wide set of available devices and after a thorough market research and was assessed for its potential to calibrate robust models for the estimation of key soil properties over the spectral signatures acquired by it. The reader is referred to DIONE D.4.1: Technical specifications of the in situ soil scanning system (SSS), data processing system and farmer's geo-tagged photos framework for more information regarding the methodology followed for the definite selection of the portable soil scanner.

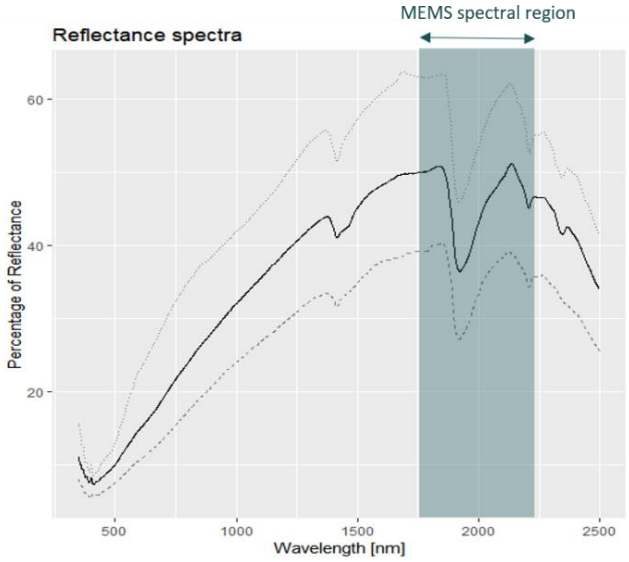


Figure 1: MEMS' Spectral region over the VNIR electromagnetic spectrum

Figure 2: S2.2 spectral sensor from Spectral Engines Company

With the definite selection of the handheld spectrometer, the next step is to ensure that the spectral data generated by the sensor meet DIONE's toolbox purposes. To this end, a set of soil samples were measured to assess the device's potential to produce high accuracy measurements of low standard deviation for many repetitions under the same conditions, i.e. high signal to noise ratio. The version that is commercially available from Spectral Engines (Figure 2) covers a wide variety of applications ranging from food quality to powder analysis or plastics. To acquire high quality soil spectral measurements, a significant modification related to the sensor's spectral scanning area was identified as crucial. The factory-ready version comes with a limited scanning field, unable to cover the minimum soil particle's size limit, which is 2mm (Roderick, 1962). To overcome this obstacle, the increase of distance between the soil sample and the sensor was proposed. As described in detail at D.4.1, the spectrometer was redesigned and customized to meet the above soil spectroscopy principle. To produce the updated version, a synthetic bracket was 3D printed and mounted to the sensor (Karyotis et al., 2021), increasing the distance from the sensor's aperture to the sensed surface by 20mm and eventually increasing the sensing area to 3.3mm diameter, meeting the initial objective. The definitive version of the device is depicted at Figure 3.

Figure 3: Soil Spectrometer with mounted 3d bracket

The device used for DIONE's tasks contains the following units (*Figure 4*):

- The soil spectrometer device
- The white calibration panel, which is a spectrally featureless material with high reflectance used as a reference panel (for further information please refer to paragraph 2.3.2)
- USB 2.0 male to USB-C male cable for charging
- Detailed user's guide in English, Greek and Lithuanian and an English narrated video tutorial for proper usage of the spectrometer

Figure 4: Soil Spectrometer packaging with its accessories

2.1.3 Smartphone application

As described in section 2.1.2, the Soil Spectrometer is totally dependent to a mobile device which is required for its operation. To this end, a mobile application has been developed that allows the interconnection of the Spectrometer with the mobile device and the later transmission of the captured data and metadata to DIONE's database. The developed application is based on a Graphical User Interface (GUI) and intents to cover the following needs:

- Allow access only to authenticated and authorized users through server-side validation
- Establish connection between Soil Spectrometer and the mobile device via Bluetooth
- Record topsoil reflectance and moisture content estimation
- Capture a photo of the soil sample through the smartphone integrated camera
- Assign a sample's name according to user's input
- Save collected data and corresponding metadata at the device's internal storage space and transmit them to the DIONE's DBMS synchronously or asynchronously
- Provide detailed help section

The application was developed with the help of the Capacitor Ionic Framework and is distributed through an apk file or an ipa file. For the installation to Android devices, the user is prompted to provide access to the device's location, storage space, and enable the Bluetooth connection in order to enable the full functionality of the application. The installation apk is delivered to end users along with the soil scanner. It is easily installable to every device with Android version 5.0 or newer by

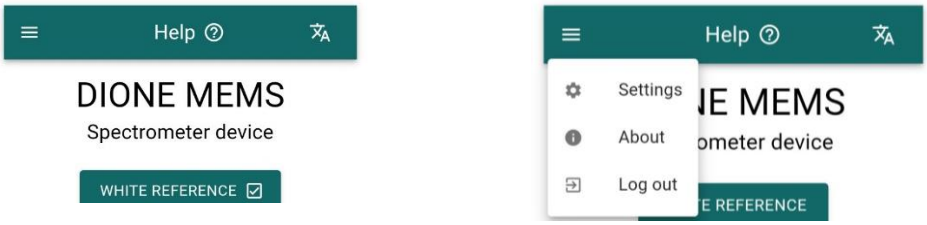
enabling the installation from unknown sources feature. When it comes to iOS devices, the installation is easily performed via the ipa file.

The users are authenticated either through DIONE's Single Sign On (SSO) where they can be self-registered to DIONE's toolbox[1], or by manually registering the user's Gmail account to the DBMS and using the Google Sign-in feature, which is used as a backup authentication and authorization method. Furthermore, since the device is distributed to a restricted number of authorized users, an extra imperviable layer of security is added, enabling the monitoring of the data flow between the central database and application holders.

By the time the Soil Spectrometer along with the apk file are delivered to the end user, the installation and first launch can be performed as depicted in the Figure 5:

Figure 5: First usage pathway

The interface is characterized from simplicity, aiming to guide the user to the final measurement step-by-step. The top of the screen contains the main ribbon where the user can access the main menu and the help section or to switch between the supported languages. The main menu is under the "sandwich menu" icon containing information related to the application release, providing access to a set of settings and enabling the user to log out and switch account. The help menu hosts a step-by-step guide on how to properly use the SSS during all measurement stages. As main language the English has been selected and at the top right corner the globalization (language selection) feature can be found; where the GUI's textual translations to Greek and Lithuanian are contained (Figure 6).



(a) (b)

[1] Soil Scanning System's users can register themselves through the unified sign on procedure accessed from: https://compliance.dione.gilab.rs/

(c)

(d)

Figure 6: (a)-(b) Main interface top ribbon. (c) Help section preview. (d) Globalization feature

The collected data are stored to the mobile device in a .json file including the sensed irradiance, its conversion to reflectance and all the relevant metadata (e.g., GPS latitude and longitude, timestamp etc.). It is noted that the photo captured by the cameras is stored as a base64 encoded string, while all the relevant EXIF metadata (including e.g., aperture value, exposure time, focal length, etc.) are also recorded. Since the SSS is intended to be used under field conditions, absence of stable internet connection may occur. To this end, the captured measurements are locally stored to the device's internal storage and transmitted to DIONE's DBMS asynchronously. For further information related to the collected data and metadata, please refer to D4.3 - "Implementation and development of systems; SSS, data processing and geo-tagged photos framework alpha versions".

2.1.4 Measurement protocol

Each commercial spectrometer comes with a usage protocol developed by its vendor, enabling users to operate the device, but these guide-through protocols are mainly general instructions that intend to cover all scenarios that the device can be used for. Since the SSS is intended to be used for the explicit use case of topsoil scanning, a simple protocol that permits the qualitative analysis of soil and can be followed by any user without any explicit technical skills has been development.

Before any usage, the latest application version must be installed to the mobile device (section 2.1.3). The device must be fully charged[2] and equipped with Bluetooth and GPS sensor which are both enabled. The Soil Spectrometer shall be also fully charged, having the protective glass cleaned from possible dirt or mud from previous usage. A simple soil spatula may help the user to prepare the topsoil prior the measurements, hence it is recommended.

With the launch of the application, the user can either sign in with their personal DIONE's account, as registered to DIONE toolbox, or with their pre-authorized Google account. After reading the terms and

[2] The device must be fully charged, meaning that it was unplugged just before departed for the field trip. The users are further advised to keep it charging (either via a power bank, or through a car charging kit) during the transition from one sampling point to another. It takes about three hours to perform a complete charging.

conditions, the user is requested to give permission to the application to have access to the device's storage space and GPS location.

The Soil Spectrometer can be switched on by steadily pressing the power button for 5 seconds Figure 7.

Figure 7: Device switch on by pressing the power button

The first time the Soil Spectrometer is used with each mobile device, it is required to perform pairing by setting the Soil Spectrometer to pairing mode through steadily pressing the Bluetooth button for five seconds until the Bluetooth led lights up Figure 8.

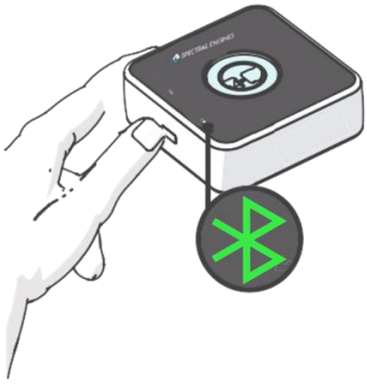


Figure 8: Enabling pairing mode by pressing the Bluetooth button

By the time the user successfully connects with their account, and before they perform any soil scanning, they must perform a white reference measurement. This can be done by placing the Soil Spectrometer with the sensor facing up, and the white reference disk centered on top of it. Then, the user can press the "WHITE REFERENCE" button on the mobile application and select their device from the list of available devices. Then, the device will initiate the procedure of taking the white reference measurement. When the white reference measurement's chart appears, by pressing the back button

the user returns to the main application screen. A new button will appear labeled as "SOIL SCAN", signifying that the device is ready to perform a soil scan[3] Figure 9.

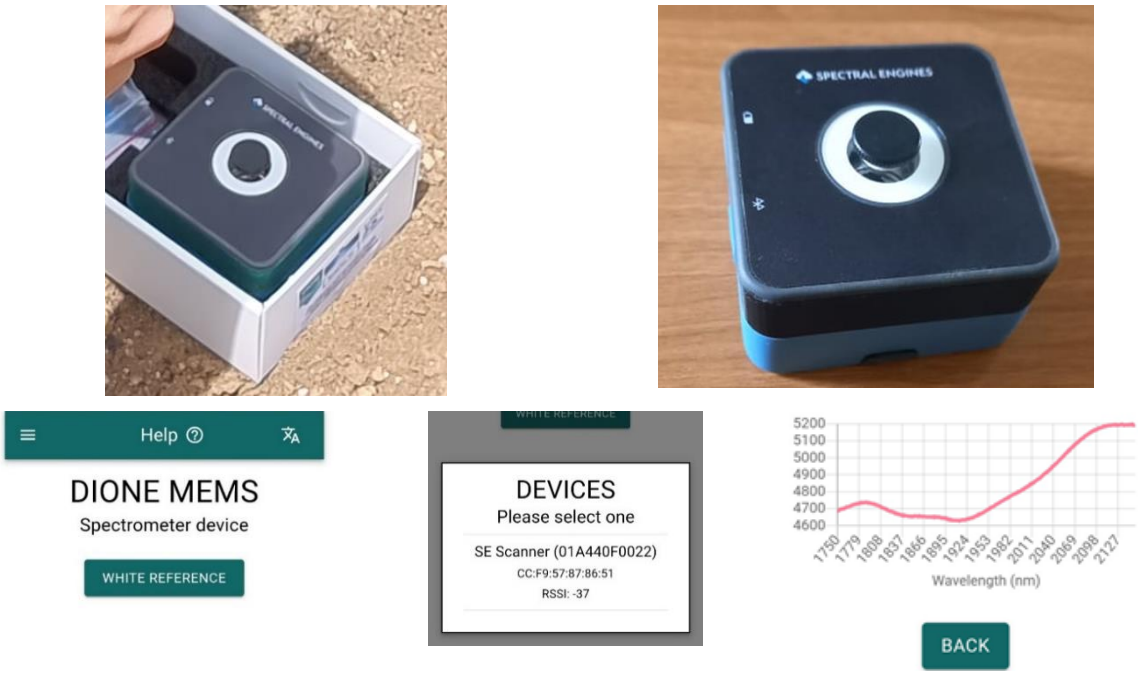


Figure 9: White reference measurement

For soil scanning, the user needs to find a spot that contains "pure" soil, meaning that there exists no vegetation, no stones or other non-soil materials and no stagnant water. Then with the use of a spatula it can be flattened. The soil scanner must be placed with the sensor facing the soil at the exact spot the measurement will be performed Figure 10.

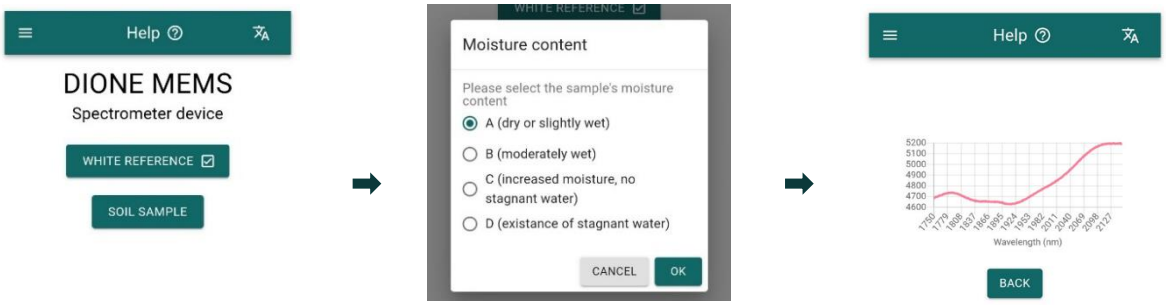
(a) (b)

[3] It is noted that the white reference measurement is the most important measurement from the procedure since any error that occurs to this measurement is transferred to the calculation of soil reflectance. The user is suggested to get to know their device by memorizing some characteristics of the curve shown in the graph (i.e. shape or the maximum value shown) and in case the white reference measurement is significantly different from the last one taken, they must repeat the measurement or contact with the application's support.

(c)

Figure 10: Soil preparation and device placement for soil scanning

By pressing the button "SOIL SAMPLE" at the application, the user is prompted to select a value from “A” to “D” which better describes the moisture that exists to the point that the Soil Spectrometer is placed, with “A” to be the optimal, completely dry case and “D” referring to the existence of stagnant water, setting the location as unsuitable for measurement. By pressing the “OK” button, the list of available devices will be shown where the user needs to select theirs. The soil measurement will instantly start, taking about five seconds to integrate. When completed, the camera will be launched, prompting the user to take a photo of the sample. After capturing the photo, by pressing the “BACK” button the device is ready to perform the next soil measurement. White calibration must be performed every time when the application is launched and after five consecutive soil measurements from last white measurement, or 10 minutes after switching on.



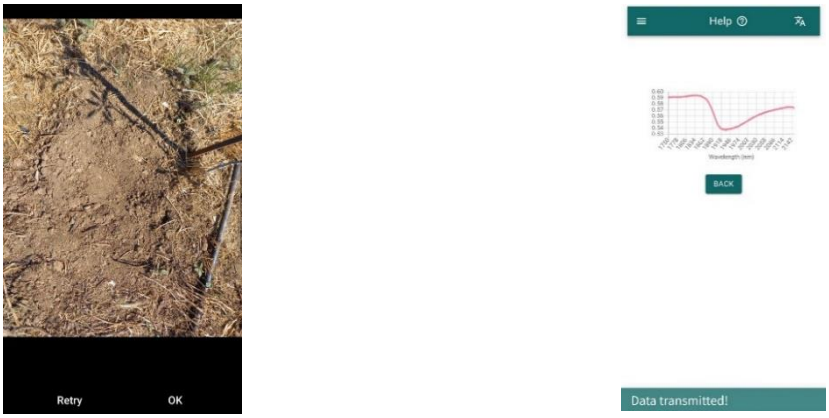


Figure 11: Soil measurement scheme

The user is suggested to perform a visual inspection to the collected reflectance by checking some characteristics of the shaped curve. The curve must be smooth; in case that it presents “steeps”, the measurement must be discarded and repeated. If the problem persists, it probably originates from low quality white reference measurement.

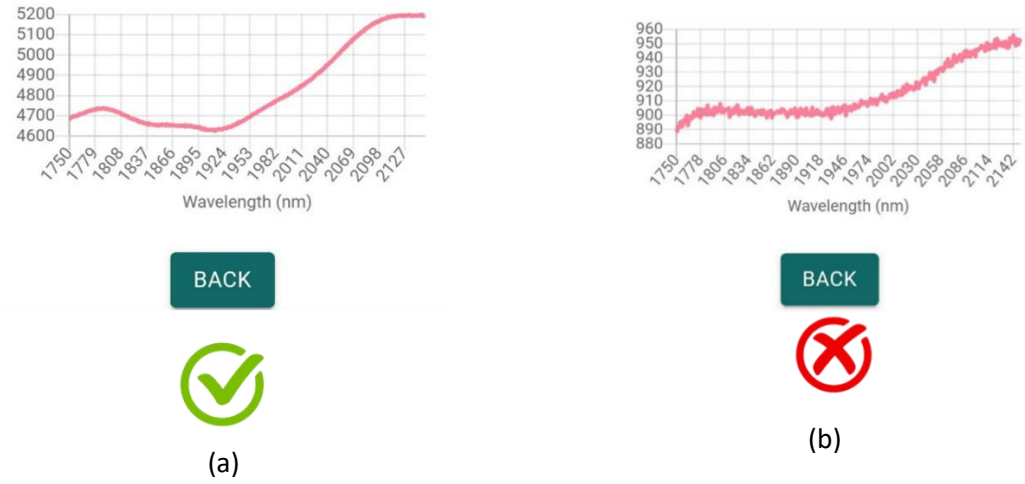


Figure 12: (a) Correct measurement - The user can proceed by tapping on “BACK”. (b) Measurement needs to be taken again

The application will try to post the collected measurements to the central database. If no stable internet connection can be established for the next 45 seconds, the “SYNC” button will appear, enabling the user to asynchronously send the measurements when internet is re-established Figure 13. A confirmation message appears when the data are successfully sent. In order to, avoid the accumulation of excessive amount of data, and thus increase the connectivity demands for the transmission to the DBMS, the user is suggested periodically synchronize the captures and avoid exceeding the local storage of more than five measurements.

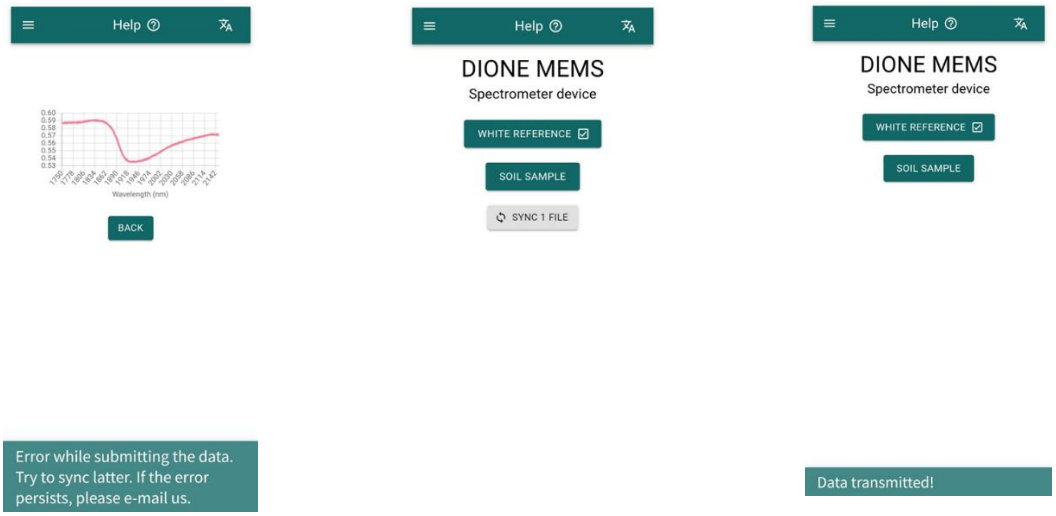


Figure 13: Asynchronous data transmission

The mobile device must have sufficient free storage space before taking any measurement. It is suggested that before any field excursion, 1 GB of storage space will be needed since every measurement along with the photo of the sample requires approximately 3 MB of space.

2.2 Farmer's geo-tagged photos framework

DIONE farmers' geotagged photos framework aims to complement the Earth Observation data sources with reliable ground-based information about agricultural parcels and thus facilitating CAP compliance monitoring. In this context, the framework comprises different components and technical innovations towards assisting and guiding users to capture efficiently representative photos of their parcels while adhering to current technical recommendations and ensuring the security, validity and reliability of the collected photos.

2.2.1 Geotagged photos mobile application

The data collection process is supported by a mobile application (frontend) that exposes to the user all the related content about their parcels while enabling the conclusion of the process and the provision of the final photos to the Paying Agencies. In this context, various processes are employed so as to enable among other things the provision of the necessary instructions for farmers to reach a given parcel, the reception of notifications about tasks they need to undertake as well as directions regarding the process of capturing appropriately a photo of a given parcel.

2.2.1.1 Geotagged photos mobile application

The application is developed in the Unity game engine[4] to benefit from the integrated Augmented Reality solution (ARFoundation[5]) that ships with the engine.

Unity is a cross-platform game engine developed initially as a Mac OS X-exclusive game engine. As of 2018, the engine had been extended to support more than 25 platforms. The engine can be used to create multi-dimensional, virtual reality, augmented reality applications.

AR Foundation is a cross-platform framework built for the Unity engine that allows to build augmented reality experiences once, then build for either Android or iOS devices. The package presents an interface for Unity developers to use, but doesn't implement any AR features itself. To use AR Foundation on a target device, separate packages are also needed to target platforms officially supported by Unity:

- ARCore XR Plugin on Android
- ARKit XR Plugin on iOS

Along with Unity, some native android plugins are developed, mainly to handle the low-level operations required for the raw measurements handling and integrity aspects. The augmented reality component aims to provide directions to farmers in order to enable retrieval of representative photos of a given parcel.

Further to this, a trip planning /navigation part is available in the mobile application, being responsible for providing the map tiles needed for the correct representation of the map while also providing the

[4] https://unity.com/
[5] https://unity.com/unity/features/arfoundation

required information for creating a route between the user location and the location of the parcel for which geotagged photos shall be collected.

During the data collection process, a time and location integrity module ensures that the application user cannot tamper with the device location and time by manually setting it. The DIONE platform has strict limitations since the photos taken must represent the real state of a parcel at a very specific time and day. Additionally, some preprocessing steps are taking place in the mobile application in order to enable at a later stage the server-side validation and integrity check of the collected photos. These include cryptographic techniques (signing the photo files with keys that are created for the encoding and decoding) as well as the extraction of relevant data the integrity checks algorithms (see section 2.2.2).

To communicate downstream to the geotagged mobile application, a service to facilitate the provision of push notifications is in place (OneSignal[6]). It provides an API to deliver messages across various platforms, abstracting details such as the platform the device is running on.

The mobile application also requires some functionalities from the DIONE Toolbox API. A farmer can register to use the DIONE services through the application by providing a unique identification originally provided by their Paying Agency. By registering, the user is provided with credentials with which they can authenticate themselves and access all the relevant data stored in the DIONE Spatial Data Infrastructure that is dynamically connected with the Paying Agencies' systems.

The photo taken, along with the required metadata, is uploaded to be verified by the data integrity, validation and anonymization component of the geotagged photos framework (section 2.2.2) and to be stored in the Central Database.

[6] https://onesignal.com/

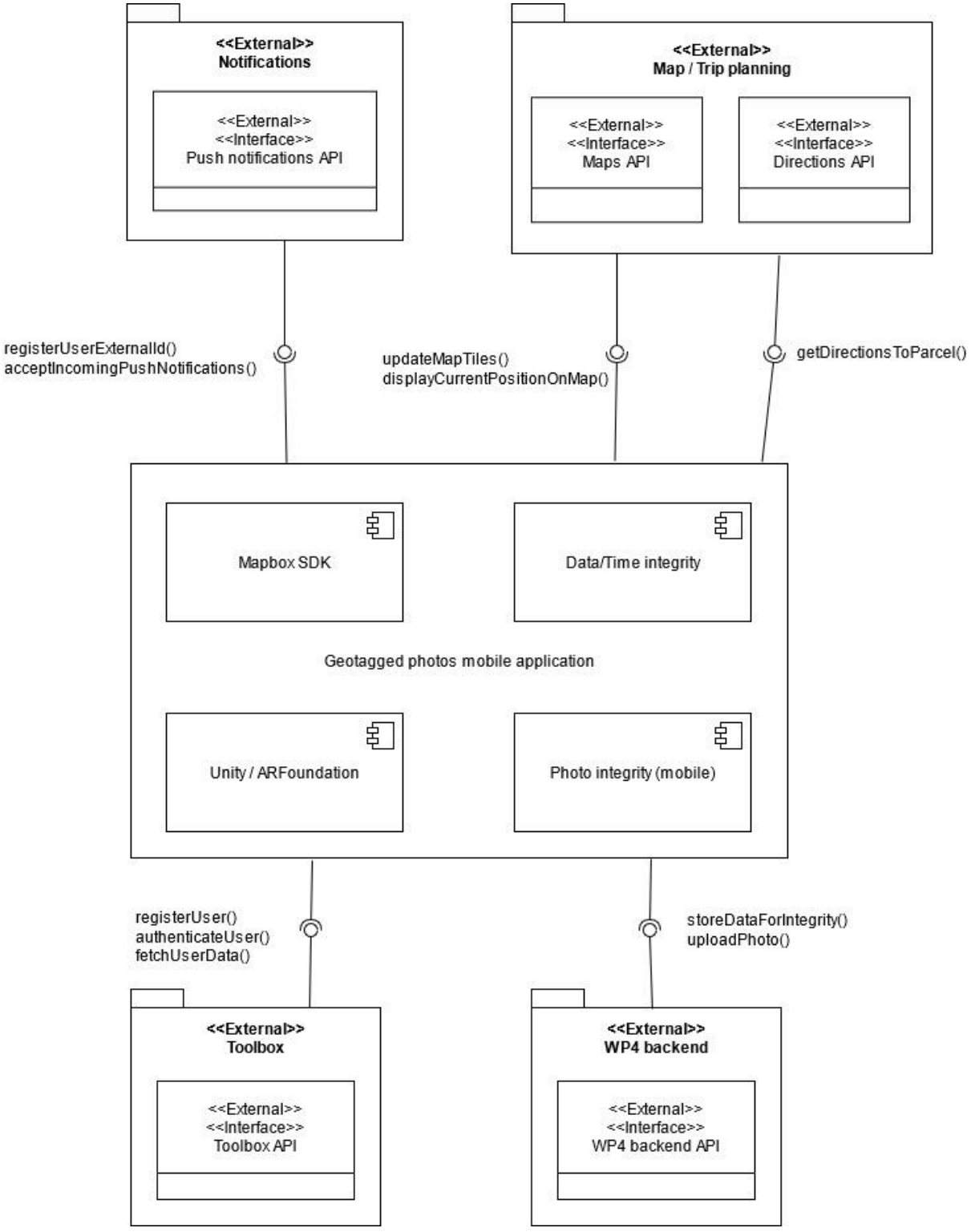


Figure 14: Geotagged mobile app connections with other components

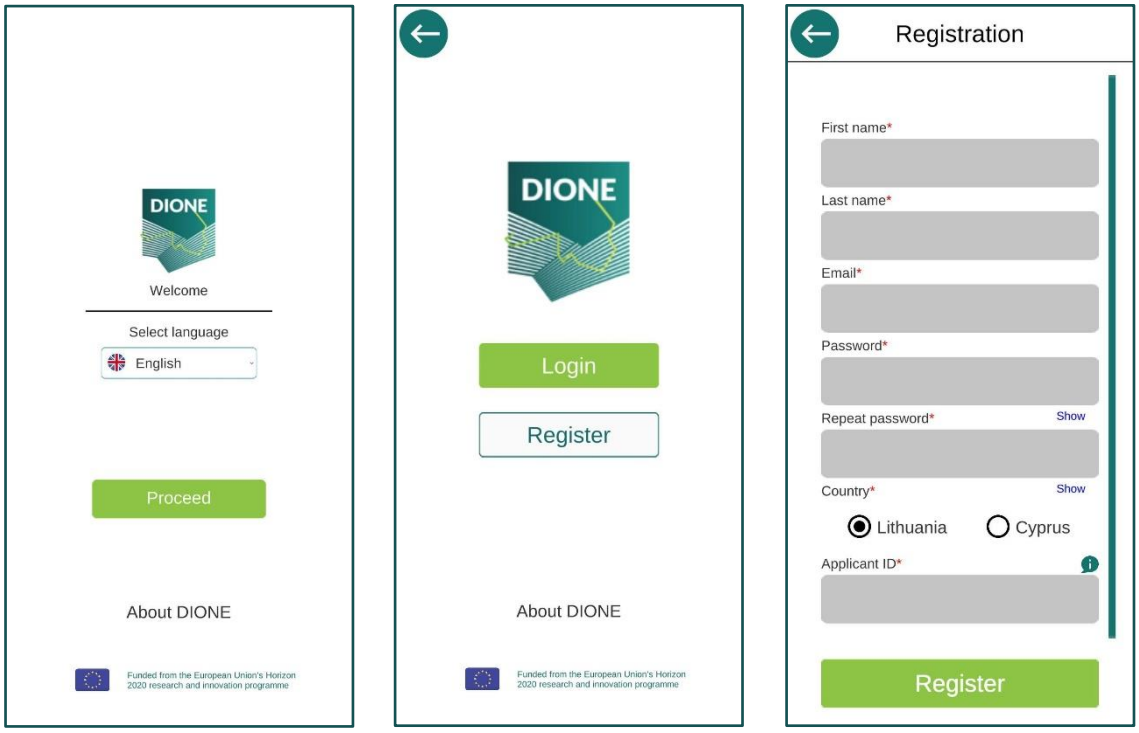
2.2.1.2 Mobile application features and functionalities

The DIONE geotagged photos mobile application covers all the essential steps of the data acquisition process. This ranges from the reception of notifications from other toolbox components (DIONE compliance monitoring tool) about the provision of additional evidence (parcels photos) to the visualization of related content and the provision of guidance towards the collection of suitable photos in the context of CAP monitoring. Since the release of the application's alpha version in April 2021, different interactions with the Paying Agencies and end-users (farmers) took place, based on which application processes and user interfaces have been properly adjusted and updated so as to further address user needs and ensure the efficient conduction of operations in a user-friendly way. The application is also accessible on Play Store. Details about the implemented features and application functionalities are presented below.

Home screen – Authentication/Authorization

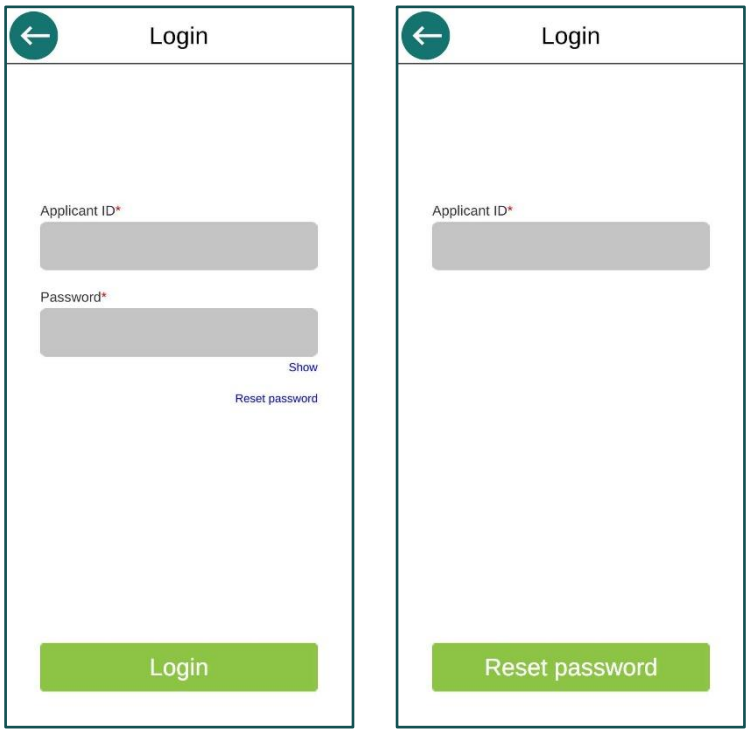
The geotagged photos mobile application provides translated text interface for English/Greek/Lithuanian to accommodate for the users in the respective countries. The language can be selected in the landing page of the application. It should be noted application is not meant for general use, rather being available only for registered users (farmers). In order to achieve this, the user must have registered themselves in the DIONE platform. The registration process requires the provision of relevant information to be stored for user security (password) and authentication (name, applicant identification). The process of verifying the user happens in two stages. The first step is the DIONE platform itself where verification happens by the user receiving an email in the declared address and clicking a verification link. The second step, is performed by the DIONE platform by checking if this user corresponds to some entry in the selected Paying Agency database. This happens by providing some form of applicant identification (i.e. applicant number and Id) required by each Paying Agency that ties the specific DIONE user to their data in the Paying Agency's backend infrastructure. During the application testing activities, the need to further assist users in the verification of their account was emerged. Thus, in order to further improve this process, user friendly messages have been added, so as to guide users in the conduction of the process.

Upon successful registration, the user may enter their credentials to login. By doing this, the app fetches all the relevant user data (parcels, locations, tasks etc), by querying the respective DIONE Toolbox API endpoints, that are crucial to populate all the different pages, menus and support all the functionalities. A session is created while the authentication performed is active and the user can interact with the DIONE platform. When this session ends or the user logs out from the application, all the temporary data stored for them are deleted. Additionally, the necessary interfaces to allow reset of the application password are provided through the login page.



(a) (b) (c)

Figure 15: a) Geotagged photos mobile application landing page, b) option for user authentication/registration, c) user registration page



(a) (b)

Figure 16: a) Geotagged photos mobile application login page, b) password reset page

Content visualization

Following user authentication, the application requires the enabling of location services in order to properly initialize the supported functionalities and allows the visualization of relevant content associated with the user. The main pages of reference in the mobile app are the "My feed/Home" page and the "My parcels" one. The former displays recent news from the respective Paying Agency (i.e. aspects of importance with respect to the application period, news on CAP implementation in their country, etc.) along with the latest pending Tasks for the user, while the latter displays all the user's declared parcels. Additionally, in the bottom of the Home page, access to the application's tutorial is provided. The tutorial is presenting the main steps that the user has to follow in order to complete a Task (Appendix A).

A Task is an action required from the farmer by the Paying Agency Inspector. It is related to a specific parcel and its location can be specified by the Inspector. Each parcel has its own unique page in the mobile app to host the various Tasks related to it.

However, a user can also act proactively and facilitate the compliance assessment process for their parcel, without receiving notice from the PA Inspector. For that cause, a Task ("Free Use") is added by default in every Parcel that can be used by the farmer to take photos and upload them towards the DIONE platform before any action may be required by the Agency.

Easy access and navigation between these two pages ("My feed", "My parcels") is supported through dedicated buttons in the bottom of the screen. Additionally, the parcel list and content provided (i.e. crop type) is dynamically extracted from the data transmitted via the DIONE toolbox API.

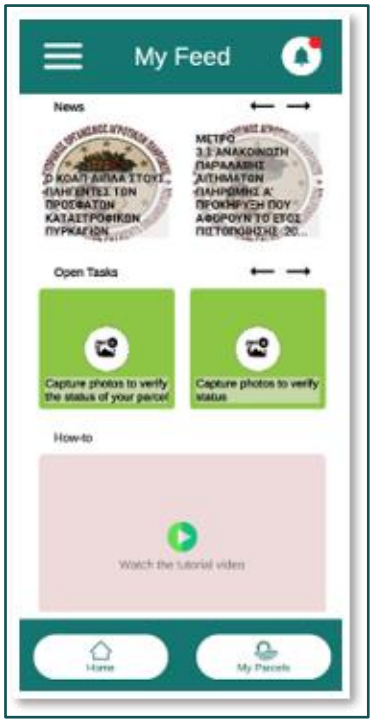


Figure 17: My Feed page (news feed, open Tasks, tutorial)

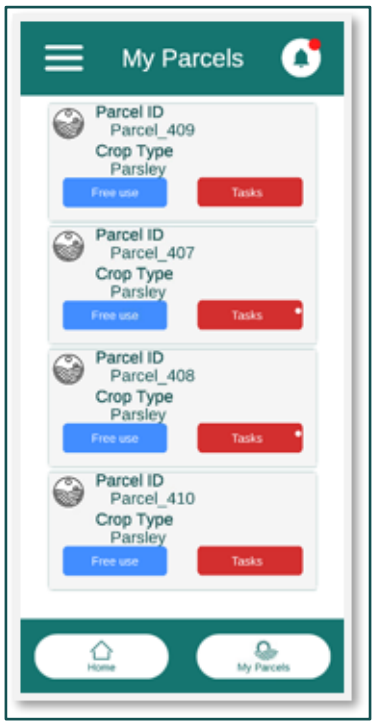
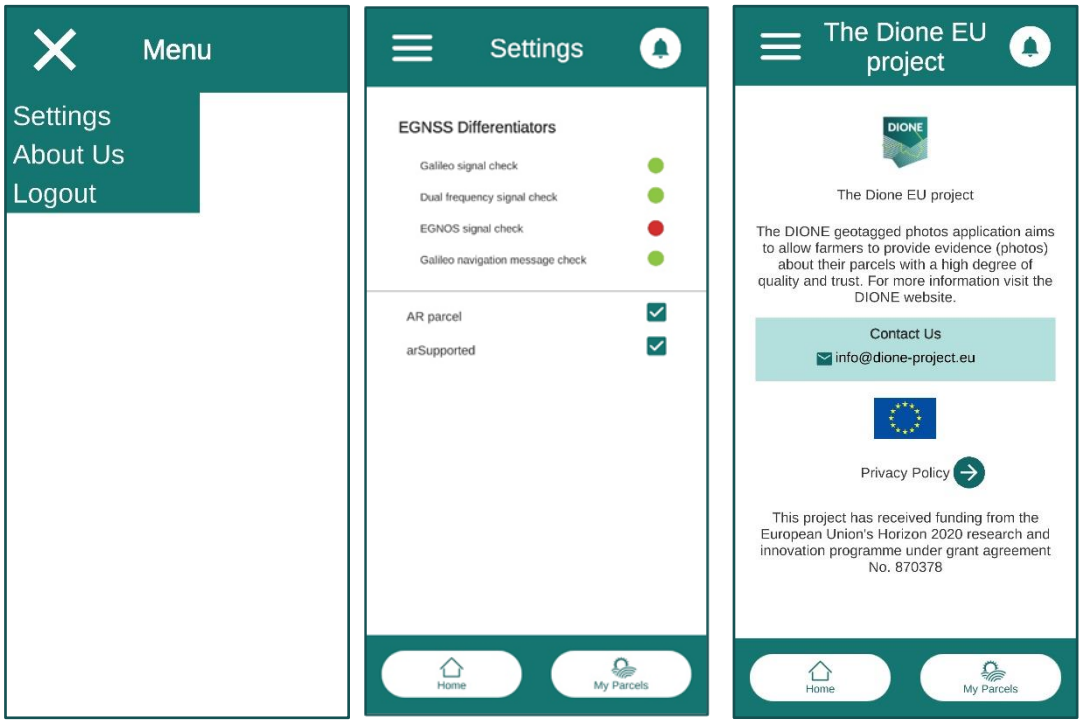


Figure 18: My parcels page (declared parcels list)

Menu

The user can also access the application's menu through the button in the upper left corner of the application. In this part, information about the scope of the implemented system and the privacy policy of the application can be found ("About Us" page), including also access to the "Settings" page. The latter aims to provide visual feedback in the form of a traffic light approach about the EGNSS differentiators supported by the device. More specifically, the application checks if the mobile device can support the particular property and specifically, Galileo capability, dual frequency support, EGNOS and Galileo navigation messages.



(a) (b) (c)

Figure 19: a) Geotagged photos mobile application menu, b) settings page, and c) information about the project page

Push notifications

A Paying Agency Inspector may create a new Task for the user at any time through the Compliance monitoring tool. In order to notify the user efficiently, a new Task in the DIONE Toolbox API, triggers a push notification to be sent towards the mobile device. The notification includes a brief description and if selected, opens the geotagged mobile application. Even if the user misses the push notification, they can find them later in the notifications page of the application. Also, while in the application, a red round icon on the notification icon and on the icons in the Parcels list also lets the user know that there are unread notifications or pending Tasks respectively. Each entry in the notifications page, can be dismissed and deleted or selected to open the related Parcel Tasks page.

The push notification system for the mobile application, relies fully on the OneSignal service. This service, primarily works by assigning player IDs to each application user. While this is the common way

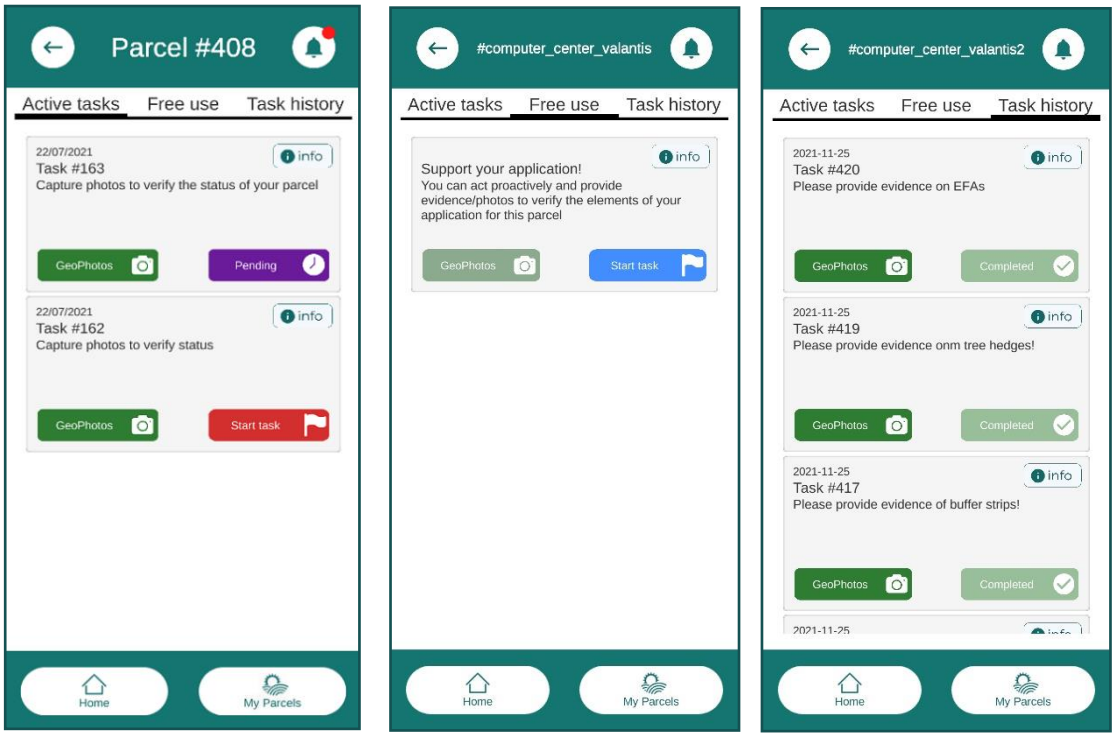
to go with, the Toolbox API authorization mechanism provides also unique user ids and thus it was utilised in the context of this application case. The OneSignal API, allows the developer to manually assign IDs that are created by an external entity (i.e. Toolbox API), hence managing to avoid multiple ids for different services.

Figure 20: Notifications' list

Parcel Tasks

Following the feedback received from relevant stakeholders, the model for managing the tasks has been updated, aiming to provide a clear representation of the respective information. For each parcel, three dedicated tabs are provided including 'Active tasks', 'Free use' and 'Task History'. The Tasks have also an associated status based on their progress (open, pending, completed etc.). This is denoted with different colours and relevant text on a button. For each Task entry of the list, the user can display its relevant information by opening the info panel.

Once a task communicated by the Paying Agency Inspector is completed, it is automatically removed from the 'Active tasks' tab and can be displayed from the 'Task History' tab. In the tab 'Free use', the user can similarly view a permanent task that allows the collection and upload of photos, without having a prior notice from the Paying Agency.



(a) (b) (c)

Figure 21: a) Active tasks, b) Free use, and c) Task history tabs

Navigation to parcel and defined spots

After pushing the button to launch a specific Task, the user is presented with a map view provided by Mapbox[7]. Its purpose is to guide the user to the Parcel this Task is associated with. In this screen several points can be identified:

- the user position is marked on the map
- the user's field of view to represent the camera orientation
- a route is rendered between the user position and Parcel position
- the parcel itself is drawn
- all other parcels with active tasks are also drawn but with a different layer to provide a quick reference to the user
- a button to change the map layers
- a button to focus the map view to the user position
- a button to display a page with text directions to the Parcel

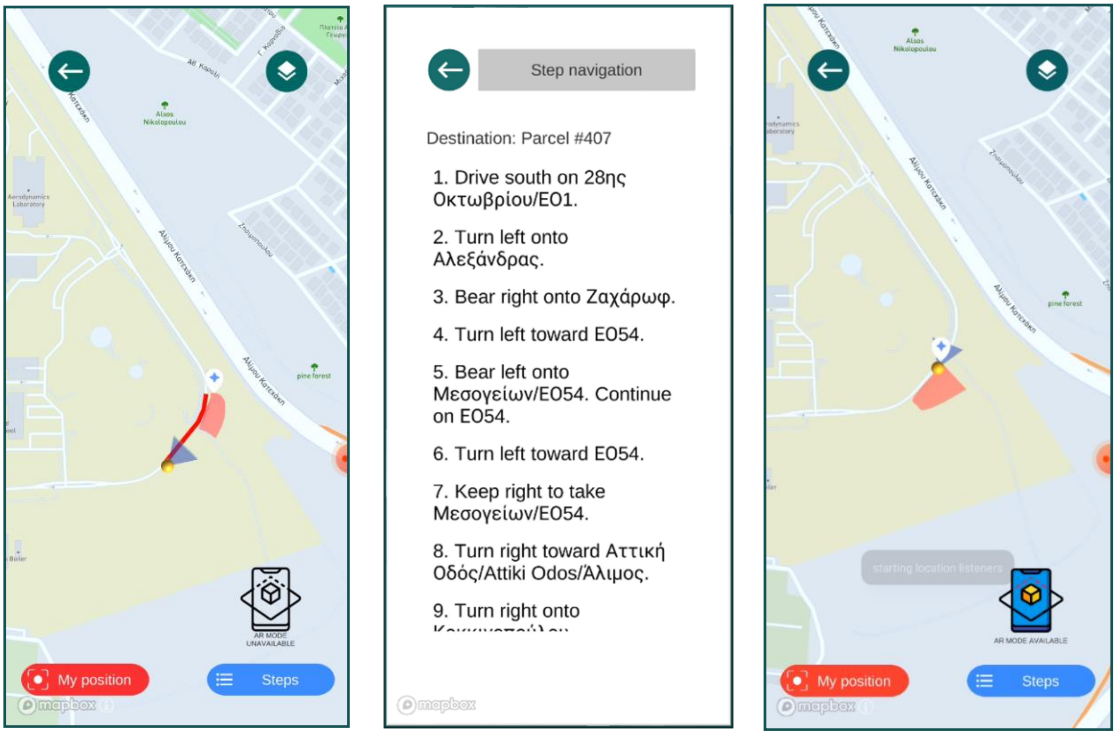
[7] https://www.mapbox.com/

- a button to switch to the AR session

The mapping platform of choice is the open source Mapbox which along with its very useful APIs, provides the building blocks for a complete position solution. Mapbox is a provider of custom online maps for websites and applications whose data is taken from open data sources, such as OpenStreetMap and NASA, and from purchased proprietary data sources. The Mapbox SDK that is used in the context of the mobile application, constitutes an open-source toolset for building mapping applications for Android devices. An essential part of the SDK if the Native Location Provider that allows the application to make use of the native Android positioning module. That way, the Unity mechanism for position can be overridden and along with it, the low precision it offers.

Users can identify their position on the Mapbox map by a marker and the related parcel is drawn as a colored layer. Also available, is the routing information from the user's position to the parcel location.

This functionality is provided by the Mapbox Directions API external service. The routing information is denoted on the map by connecting the user's position and destination, along with turn-by-turn text instructions.



(a) (b) (c)

Figure 22: a) Mapbox navigation to the Parcel, b) Mapbox text directions to the parcel, c) camera activation while approaching to the indicated location for photo acquisition

Augmented reality photo capture

When requesting geotagged photos from the farmer, the intension is to obtain sufficient information in order to avoid any physical field visits (on-the-spot-checks) by the Paying Agencies' Inspectors.

Therefore, the collected images should provide an overview of the parcel, but not necessarily cover its entirety and all the details. In order to assure a comprehensive view on the element and to limit the possibility of image manipulation, it is recommended to provide at least 2 photos of the element captured from different viewpoints or camera heading.

It is advised to capture photos in landscape format (horizontally) and point the camera so that the element to evidence is depicted in the image centre. Considering the image content, photos may be divided into two main framing categories of “overview” and “macro” photos. An overview photo should depict a larger part of the field and include landscape elements other than the main object (crop, activity etc.), if possible. This type of photo aims at reducing the uncertainty linked with the limited accuracy of the geotag and at providing an overview of the field condition. A macro photo must serve to enable the robust identification of the element to evidence. This subject could be a mixture of crop as Ecological Focus Area (EFA) cover, presence of rare crops that cannot be reliably discriminated in the Sentinel data etc.

In the context of a Task created by a Paying Agency Inspector, the number of photos that have to be collected including their types (macro/panoramic) and the preferred camera orientation, are specified in the Compliance monitoring tool.

(a) *(b)* *(c)*

Figure 23: a) Restrictions and guidance to allowed photo collection spot b) Parcel borders and allowed photo collection spots, c) Photo taken preview (without AR content)

Using ARFoundation, the user is instructed on what is required of them, like where to take the photo from, camera orientation etc. Since a user can take geotagged photos either proactively or in the context of a Task, the application restricts the location of a photo accordingly. If the farmer takes initiative without having a request, he/she is only allowed to take a photo from inside the parcel or

near each one of the parcel's corners (vertices). On the other hand, if there is a Task, the farmer is only allowed to take a photo within a radius from the location that the Inspector has selected when creating the specific Task.

Before any restriction can apply, the AR system needs to initialise properly. Also, due to the fact that the smartphone providers implement sensor fusion and smoothing algorithms in the determination of the geolocation and the compass readings, it is required to wait at the spot where the photo will be taken for several seconds before releasing the shutter.

The camera orientation is constrained by the tilt angles of the devices to only accept photos taken within a threshold. While in the AR session, the application continuously looks for active applications running in the background that may tamper with the GNSS signal by mocking the actual location.

ARFoundation is used to handle all the AR session content. AR Foundation can support the following concepts among many:

- Device tracking: track the device's position and orientation in physical space.
- Plane detection: detect horizontal and vertical surfaces.
- Anchor: an arbitrary position and orientation that the device tracks.
- Light estimation: estimates for average colour temperature and brightness in physical space.
- 2D image tracking: detect and track 2D images.
- 3D object tracking: detect 3D objects.
- Meshing: generate triangle meshes that correspond to the physical space.
- Raycast: queries physical surroundings for detected planes and feature points.
- Pass-through video: optimized rendering of mobile camera image onto touch screen as the background for AR content.
- Session management: manipulation of the platform-level configuration automatically when AR Features are enable or disabled.

In conjunction with this, the "AR + GPS Location" Unity asset is used to position 3D objects in real-world geographical locations via their GPS coordinates. This asset helps place all points of interest in the AR session so that they correspond to their real-world positions. The points of interests are considered to be the parcel shape vertices and conditionally the location picked by the Inspector for the Task request. These points are used to place AR anchors which are positions within the 3D world that are tracked for stability. A polygon mesh is drawn based on the parcel shape vertices to mark the actual parcel for visibility reasons.

To enhance the user experience, an AR session based on real coordinates should be as stable and accurate as possible. Various techniques have been employed towards this cause:

- The platform of choice for the development of the mobile application is the Unity game engine. Unity provides a mechanism to access location data, however this data is of low precision. This, in turn, leads to a lower position accuracy and a lower fidelity for the AR session in general. To overcome this, a method has been implemented to get the native location information directly from the Android system.
- Since the native location data can be accessed, the accuracy metric provided to calibrate the whole solution is utilised. The positional accuracy is being tracked and the digital content is drawn only when the accuracy is below a specific ceiling. This way, the spawned content is

placed accurately enough to begin with. On top of this, all the AR content is re-drawn whenever the accuracy is improved, so that the overall experience is improved as well.

- The geotagged mobile application does not use the traditional camera application to take the photos and this is because of the AR session occupying the camera hardware. Hence there is no mechanism to automatically embed the EXIF metadata to the file as it usually happens. Thus, a dedicated mechanism is employed, that takes the background of the image (i.e. no AR content) and in the resulting PNG file a custom method is applied to decode and embed the required metadata.

Photos' gallery and data upload

The user can view the photos taken, grouped by Task to know what they already have available and also delete or upload the desired ones to the backend for integrity checking. During the Task creation, the Paying Agency Inspector selects the number of photos required and sets a deadline for the completion of the Task. In this page the user can also see such information. When the Task is past its deadline or the max number of photos has been uploaded, no more photos are allowed to upload. Also, when at least the number of photos that are required have been uploaded, the Task status changes to Pending meaning that the uploaded evidence data is under inspection.

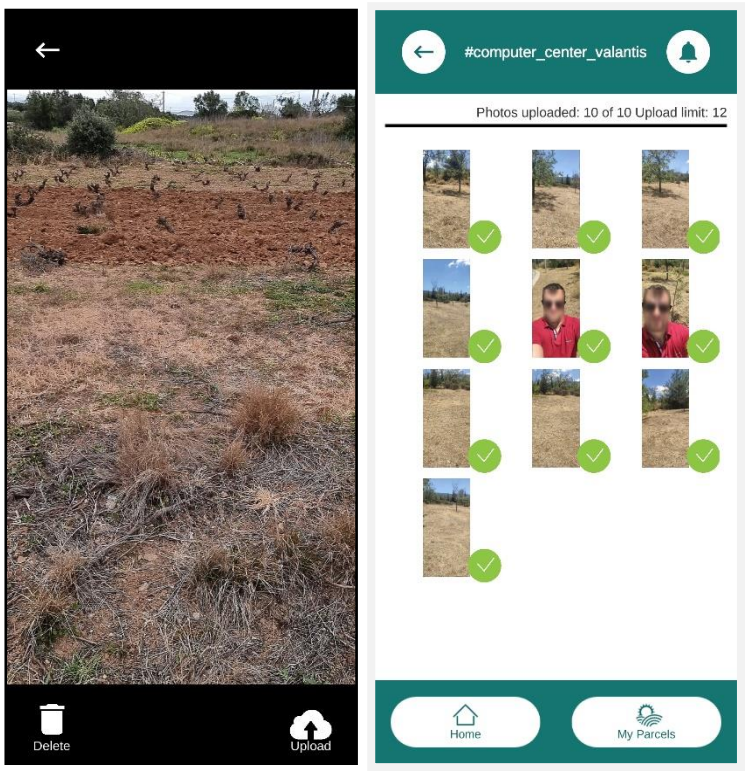


Figure 24: Upload of photos captured for a specific task

Offline mode

The nature of the farming activities and the geotagged mobile application's purpose mean that the most significant actions in the application's lifecycle will take place outdoors. It is not rare the fact of agricultural parcels being situated in remote and mountainous places that are not covered by mobile

network signal. Therefore, it is crucial that some of the functionalities of the application can be performed when no mobile signal reception is available.

Since the data presented to the user need to be the latest ones, the first steps of the process require an internet connection to fetch the relevant data (Parcels, Tasks). This data, in turn, is temporarily stored and is available while the application is "running" subsequently offline.

Also unavailable is the access to the map view via Mapbox. The map relies on getting tile information via the internet so no useful information can be presented otherwise. However, a user can download (automatic process) some initial map tiles in the map view, disconnect from the internet, and have these initial tiles as guiding reference to the parcel in question.

The entirety of the AR session is working offline. The AR content that is superimposed is based on the initial data fetched for the user from the Toolbox API. The photos taken are stored locally in the phone's internal memory so that when a network connection is available, the user can browse through them and upload the most appropriate ones. As mentioned, the time and date of a taken photo are very important to the project as they provide a timestamp for the snapshot of the evolving crops in a parcel. The time integrity component is working offline as well. It can provide a date and time irrelevant of the phone's settings or other external providers that require network to function. The only requirement is the reception of GNSS signals which is a trivial task when outdoors.

Location Accuracy

In the context of the application alpha version, different methods have been researched and analysed in order to ensure high location accuracy including the exploitations of multiple location differentiators (i.e. dual frequency devices) and the use of EGNOS Data Access Service (EDAS) for ground-based access to EGNOS data through the Internet..

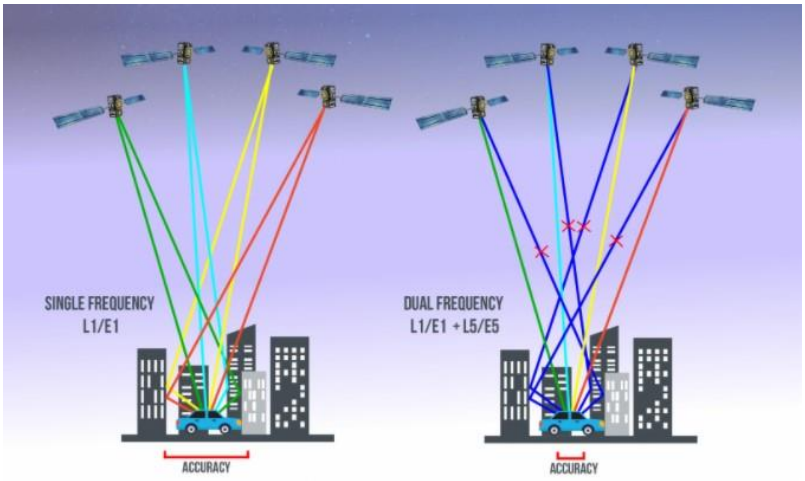


Figure 25: Dual frequency GNSS signals concept

Dual-frequency capability means that the GNSS receivers are able to receive two GNSS signals at different frequencies from a satellite. In the case of Galileo, these frequencies are E1 and E5a. Dual frequency provides increased reliability to users – if one of the frequency bands fails, the other can be used as backup.

The selection of this dual frequency combination is particularly appealing for two main reasons:

- Spectral efficiency: all the signals are with the same central frequency and bandwidth.
- Wide-band signals: most signals in the L5/E5a band are wideband with BPSK (10) modulation.

Users of dual-frequency smartphones are able to benefit from the reduced signal acquisition time and improved accuracy of positioning and timing. Dual frequency also reduces problems caused by obstructions such as buildings and other obstacles, thanks to the fact that the L5/E5a signals are lower in frequency, making them are less prone to multipath interference errors.

It is important to stress that the final positioning accuracy in mass-market devices is not only driven by GNSS measurements, either single or dual frequency. In addition to GNSS, a very important role is played by the smartphone integrated inertial sensors and additional terrestrial based signals, including for example cellular network (4G/5G), Wi-Fi, NFC, Bluetooth, etc. All these ingredients contribute to the fused location and its ultimate accuracy.

EGNOS

The European Geostationary Navigation Overlay Service (EGNOS) provides an augmentation service to global navigation satellite systems (GNSSs), such as GPS and Galileo. EGNOS[8] provides three services:

- Open Service (OS), freely available to any user.
- Safety of Life (SoL) Service, that provides the most stringent level of signal-in-space performance to all Safety of Life user communities.
- EGNOS Data Access Service (EDAS) for users who require access to specific GNSS data streams for the provision of added-value services, professional applications, commercial products, R&D, etc.

EGNOS is a Satellite Based Augmentation System (SBAS). SBAS systems are designed to augment the navigation system constellations by broadcasting additional signals from geostationary (GEO) satellites. The basic scheme is to use a set of monitoring stations (at very well-known positions) to receive the navigation signals from core GNSS constellations that will be processed in order to obtain some estimations of these errors that are also applicable to the users (i.e. ionospheric errors, satellite position/clock errors, etc.). Once these estimations have been computed, they are transmitted in the form of “differential corrections” by means of a geostationary satellite.

The EGNOS Data Access Service (EDAS) provides ground-based access to EGNOS data, through a collection of services, which are accessible to registered users through the Internet.

[8] https://www.gsa.europa.eu/egnos/what-egnos

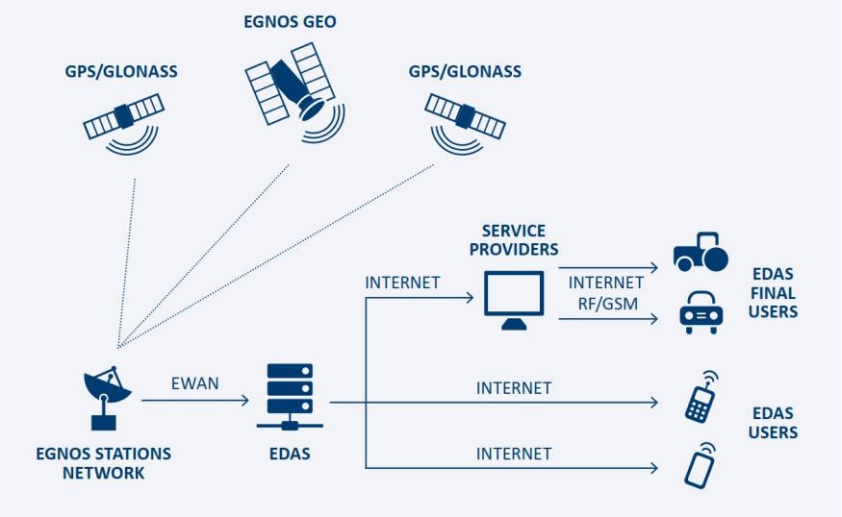


Figure 26: EDAS high level architecture

More specifically, he EDAS Signal In Space (SISNeT) service provides access to the EGNOS geostationary satellites messages transmitted over the Internet through the SISNeT protocol. Doing so, the EGNOS signal is available even if geostationary satellites are not visible from the user location.

Thus, in the context of DIONE, the EGNOS messages received from the EGNOS Data Access Service (EDAS) over the internet are used to augment the GPS signal received. This option requires an application capable of processing and applying the EGNOS corrections (obtained from the EDAS SISNeT service) to the computed position solution. This approach aims to also cover the mobile devices built with chipsets and firmware able to access GPS raw data but are not SBAS-compatible (not being able to access and process EGNOS raw due to their operative system/firmware).

2.2.2 Data integrity, validation and anonymization

2.2.2.1 Outline of integrity framework

In order to ensure that the file of the photo captured by the mobile application is not manipulated and is the same with respect to the location and time the photo was taken, an integrity framework has been developed. Different algorithms have been implemented for a multifaceted solution and better security. The framework consists of two parts, i) a mobile module which is an android library responsible for running some initial processing and ii) a backend module that validates the origin of the photo and on a successful validation creates a blurred image keeping potential personal information that may exist in the data private.

The chosen language for the implementation of the mobile library is Java. Java is one of the two languages with which native android applications are developed. Moreover, being one the most used programming languages provided a variety of possible APIs and modules that can be exploited in the context of the library. One equally used language is Python. By taking into consideration the good support for image processing and the security libraries available, it was considered as the right choice for the backend part of the framework.

It should be noted that since the initial release of the framework, implementation activities during this period were mainly focused on code refactoring and bug fixes. More specifically, the backend component was refactored, being able to process multiple photos in parallel without one intervening the others, instead of processing them in a serial manner. In this way, the response time of the requests was managed to be reduced.

2.2.2.2 Integrity algorithms implemented

2.2.2.2.1 Lightweight digital signature scheme

The mobile library is triggered in different times while using the application. The first time is during the authentication of the farmer, when a unique pair of a private and a public key is generated[9]. The public key is communicated and stored in the database, being bonded with a specific farmer. However, the private key never leaves the smartphone and is used only by the application. When a farmer takes a photo, the digital signing takes place, using the private key, and a signature is generated. The signature is produced and sent to the API (Central Database, see section 2.3) along with the photo. This signature cannot be verified even if a bit is changed in the photo. As a result, during the verification in the backend, it is ensured that the original photo has been sent.

For the generation of the keys and the digital signing in the mobile library, the Security library available in Java is utilised. The corresponding library in the backend framework is the Crypto Python library. For the DIONE scheme, the RSA with a SHA256 algorithm is chosen due to the smaller key sizes and the faster computations compared to the elliptic curves. It is crucial to mention here that the steganography technique and the metadata embedding (see sections below) are conducted before the digital signature takes place because otherwise the signature verification would fail.

2.2.2.2.2 Steganography

Another mechanism that it is used to ensure the authenticity of the photo is the steganography technique. The steganography technique uses the least significant bits of every pixel, in order to embed a secret message in the photo that can only be extracted by using the reverse algorithm, and is not identifiable by common tools. As only the least significant bit of a pixel is used there is no visual alteration in the photo.

For the implementation of the steganography technique a java algorithm has been implemented. For the extraction zsteg[10] was used, which is a bash command tool implemented in Ruby. The integration of this bash tool with the framework was straightforward using the OS Python module. A successful extraction of the message would mean that the file is not tampered.

2.2.2.2.3 Photo metadata

A number of information, like GPS, date and focal length, which are checked for their integrity (see next paragraphs) are collected during the usage of the mobile application. This information is

[9] https://ieeexplore.ieee.org/document/507642
[10] https://github.com/zed-0xff/zsteg

embedded in the photo in the form of metadata[11]. Later the metadata are extracted in the backend and cross checked with the farmer's declared data stored in the database.

For the embedding, the pngj[12] library and for the extraction the imagemagick[13] suite are utilised. The integration with the framework is identical to zsteg tool described above.

2.2.2.2.4 Copy Move Forgery Detection

One popular digital modification technique is the copy move forgery. The copy-move forgery is the process of copying and pasting from one region to another location within the same photo. An algorithm has been implemented which is able to detect such cases. The algorithm is based on principal of component analysis, but sensitive to noise and post region duplication process[14].

This is pure python implementation, using Pillow module for the photo processing.

2.2.2.2.5 Error level analysis

Every photo file is compressed using a specific algorithm (i.e. JPEG etc.). Taking this into consideration, the entire photo should be roughly at the same level, if a difference is detected, then it likely indicates a digital modification. This technique is called Error level analysis.

For this technique a pure python script was implemented. The result of the algorithm is photo with the parts of the photo which are at different compression levels in brighter colour. Different parts of the photo could be in different compression levels, but if these parts form a meaningful figure (i.e. a face or a tree), this implies that the photo is modified.

However, the automation of this process didn't yield satisfactory results, and thus this algorithm wasn't included eventually in the geotagged photos integrity framework.

2.2.2.2.6 Camera identification

To further ensure the origin of the photo, each user is paired with his mobile device. This is achieved with a module which calculates the pattern noise and identifies the camera used to take the photo. The identification becomes possible due to the fact that each component in a digital camera leaves intrinsic fingerprints in the final image output, which due to manufacturing choices are unique for each device. The component of the camera that makes possible the identification is the complementary metal oxide semiconductor (CMOS) image sensor, which inserts a pattern noise in the photos which is unique for each device.

The method used for the development of this module is presented in the "Estimation of Gaussian, Poissonian–Gaussian, and Processed Visual Noise and Its Level Function"[15]. However, due to the fact photos are collected through the mobile application AR features, the characteristics of the camera

[11] https://en.wikipedia.org/wiki/Exif
[12] https://github.com/leonbloy/pngj
[13] https://imagemagick.org
[14] https://www.semanticscholar.org/paper/Exposing-Digital-Forgeries-by-Detecting-Duplicated-Popescu-Farid/b888c1b19014fe5663fd47703edbcb1d6e4124ab
[15] https://ieeexplore.ieee.org/document/7506318

cannot be imprinted in the photos and thus the pattern noise cannot be computed properly. To encounter this, the focal length is read which constitutes a fundamental characteristic of the camera. After the photo is taken, using the Android API, the focal length is extracted and embedded in the photo so as to cross check it later in the backend.

2.2.2.2.7 Time Integrity

The time integrity module uses the Android API and based on the equation below, manages to recreate a near-precise clock independent of network access. By doing this, a user cannot fool the DIONE platform and all the photos taken are tagged with their true date and time. The date and time issue are sensitive in the context of in-situ checks since the evidence data is only relevant at the time period that they are requested.

```
local estimate of GPS time = TimeNanos - (FullBiasNanos + BiasNanos)
```

2.2.2.2.8 Location integrity

One critical piece of information that has to be verified is the location from which the photos are being captured.

For this purpose, a module has been developed, being capable to detect any external process/application that attempts to alter the position information/GPS of the mobile device. In such cases, the data collection is not permitted.

Furthermore, in order to enhance the location integrity verification, a dedicated algorithm has been developed allowing the exploitation of the open service navigation message authentication scheme (OSNMA). OSNMA allows a GNSS receiver to verify the authenticity of the GNSS information and of the entity transmitting it, and ensures that it comes from a trusted source. By exploiting the OSNMA, the geotagged photos integrity framework shall be able to ensure that navigation message received is identical to the message transmitted and that was generated by a trusted source.

From development perspective, the different components of the receiver algorithm have been implemented, including HKROOT and MACK parsing from the Navigation message and their validation as well as the realisation of the TESLA protocol.

Regarding the implementation, the C++ programming language was used due to the faster computation times. For the cryptographic primitives the CryptoPP[16] library is utilized. This application will be integrated as a library in the next version of the DIONE geotagged photos application using the Java Native Interface (JNI) of the Android platform. For the parsing of the Navigation message there is a dedicated java class in Android, called GNSSNavigationMessage. The specific bits regarding OSNMA can be received using the getData function of the class.

Last but not least, it should be noted that currently smartphones based on BROADCOMM GNSS chipsets provide navigation messages and carrier phase messages and thus can exploit OSNMA. Until

[16] https://www.cryptopp.com/

recently, QUALCOMM GNSS chipsets could not be used for this cause, since they do not provide such messages. However, the new generation of QUALCOMM Snapdragon chipsets[17] will be able to meet also this requirement. For the purposes of DIONE geotagged photos framework, the location integrity for devices that do not support the needed hardware will be able to be ensured through the first module that is able to detect and trap fake GNSS signal.

2.2.2.3 Anonymization component

The last component of the framework aims to address privacy considerations. Personal information that may exist in the collected photos should not be communicated to other users of the data (i.e. PA Inspector) in order to comply with the provisions of the General Data Protection Regulation (GDPR). Thus, an anonymization component has been implemented, which is responsible for blurring any faces or license plates exposed in the photos.

The anonymization component is basically a pre-trained convolutional neural network. For the training of the network various core libraries have been used such as tensorflow, numpy, scipy and Pillow.

Figure 27: Example of anonymized photo

The overall process (both mobile and backend) is depicted in the next photo.

[17] https://www.gsc-europa.eu/news/qualcomm-launches-snapdragon-with-dual-frequency-and-5g

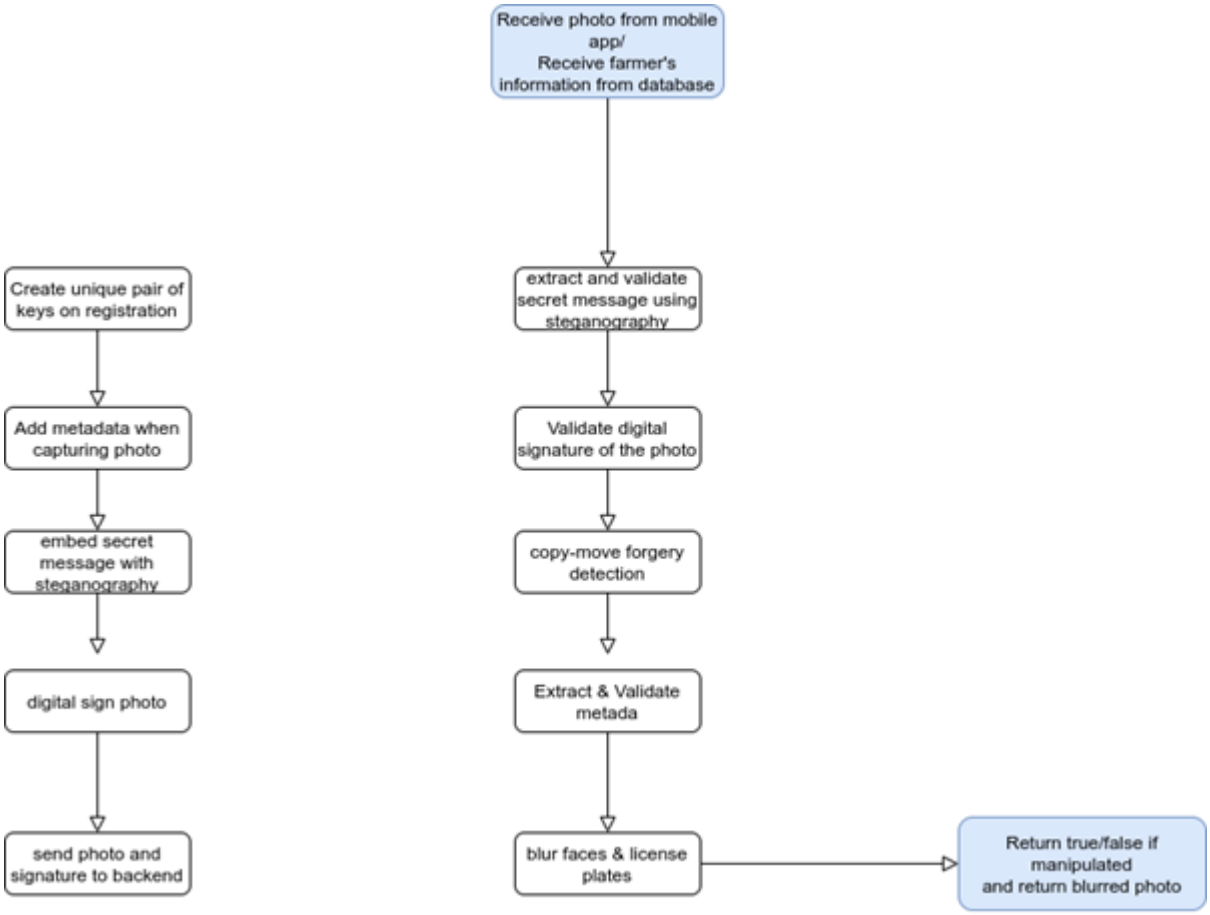


Figure 28: Schematic overview of the mobile and backend process of the geotagged photos integrity, validation and anonymization framework

One critical consideration of the development phase was the smooth communication between the application and the forensics library. This task was successfully done, due to the AndroidJavaClass class in the Unity game engine, which is basically a wrapper for an android class.

2.2.3 Data transmission

The photo related data is stored in the Android device in a .json file as follows:

Key	Value
`userId`	Unique identifier of the user
`uuid`	Unique identifier of the image
`parcel`	Unique identifier of the Parcel the photo is taken for
`task`	Unique identifier of the Task the photo is taken for
`date_time`	Date and time directly decoded from GNSS signals
`coordinates`	Coordinates of the photo location from native Android

bearing_angle	Compass reading from the device sensor at the time of the photo
image	The photo itself in base64 PNG format
blurred_image*	The resulting image after the integrity checks and the blurring applied
validated_image*	Boolean flag
signature	The hash of the file used for integrity
publicKey	Public key for signature check
focalLength	Device camera focal length
status	Image status (uploaded, not uploaded, dismissed by integrity)

2.3 Data processing and storage system

The goal of this section is to present the central data processing and storage system for the two mobile applications presented in Sections 2.1 and 2.2, namely the SSS and the farmers' geotagged photos framework. The high-level overview of this subcomponent is presented below and in Figure 29.

In essence, the data collected by those two mobile applications are transmitted over-the-air to a server that hosts a) the database and b) the RESTHeart service required to expose this database to the public. Moreover, the server further contains automated hooks which are executed on specific instances (e.g., when a new datum is received) to check the integrity and validity of the data, ensuring a) the integrity of the data is not comprised during the transmission, b) that the data are error-free and have not been tampered with by external actors, c) the data are sound (contain no outliers or are of the expected input). Should the aforementioned conditions be met, then further data-specific processing pipelines are also automatically triggered; these entail for example the estimation of soil properties from the spectral data and the automatic posting of the geo-tagged photos to the DIONE toolbox.

2.3.1 Architecture of the database management system

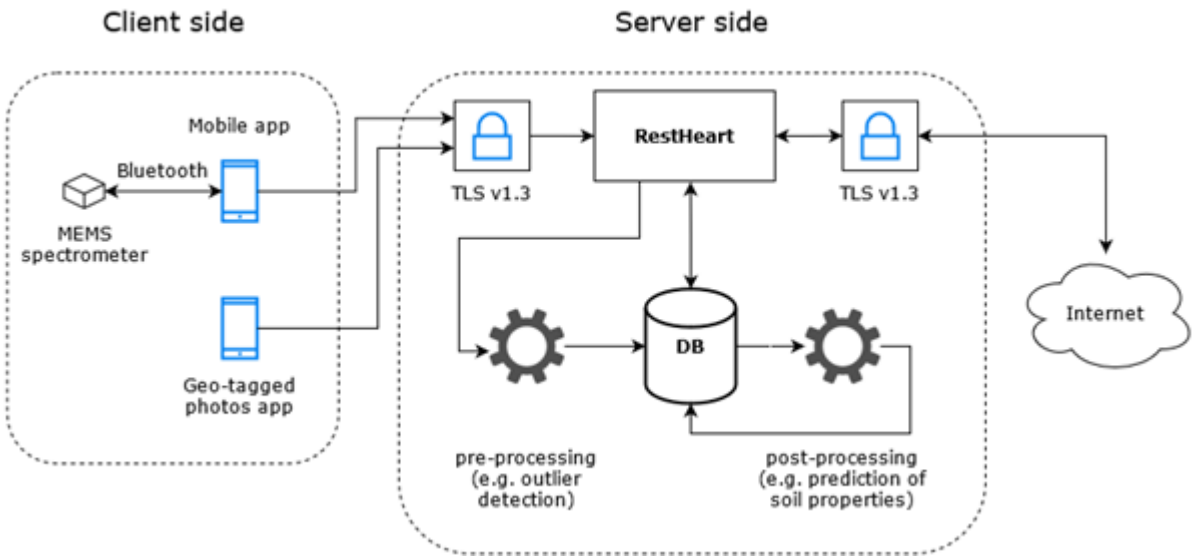


Figure 29: A simplified view of the data processing and storage system

At a higher level, the database management system deployed (Figure 29) uses:

- The **MongoDB** document-oriented database, which stores data records as BSON documents. BSON is a binary representation of JSON documents, though it contains more data types than JSON. It can thus handle the heterogeneous data that are required within the scope of the DIONE project. Moreover, MongoDB also includes GridFS which is a specification for storing and retrieving files that exceed the BSON-document size limit of 16 MB; it is used in DIONE for storing the photos.
- **RESTHeart** which connects to the MongoDB instance and exposes the data using a RESTful interface over Transport Layer Security (TLS). It is written in Java and is built on top of RedHat's

undertow non-blocking HTTP server. Moreover, it is extendable via plug-ins enabling developers to implement web services in minutes.

- The **TLS** components to connect to the outside world securely. It is a cryptographic protocol designed to provide communications security over a computer network.

With respect to the underlying database, MongoDB is a NoSQL database which differentiates from classical relational database systems (RDBMS Relational database systems (RDBMS) and NoSQL databases have different strengths and weaknesses:

- In RDBMS, data can be queried flexibly, but queries are relatively expensive and don't scale well in high-traffic situations.
- In a NoSQL database such as MongoDB, data can be queried efficiently in a limited number of ways, outside of which queries can be expensive and slow.

These differences in turn make database design different between the two systems:

- In RDBMS, one designs for flexibility without worrying about implementation details or performance. Query optimization generally doesn't tend to affect schema design; however, normalization is an important parameter.
- In MongoDB, the schema is designed specifically to make the most common and important queries as fast and as inexpensive as possible. Put in another way, the data structures are tailored to the specific requirements of the different use cases.

MongoDB uses a database to hold one or more collection of documents. A document can be seen as a single measurement, analogous to a line of a traditional relational database, but it can hold more complex information. The documents are arranged into collections which are analogous to tables in RDBMS. This is illustrated in Figure 30.

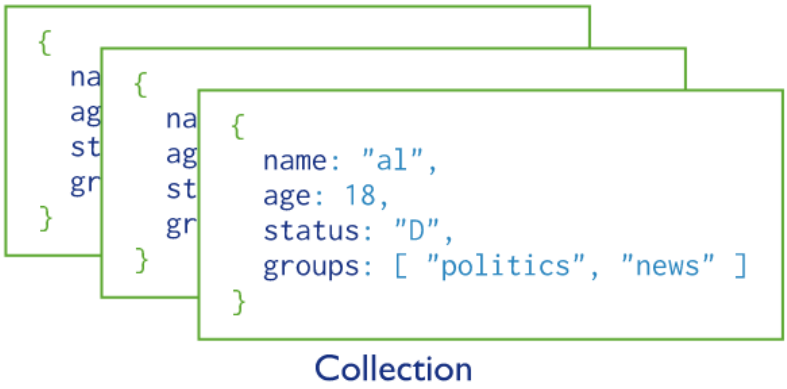


Figure 30: MongoDB stores the data in a collection of documents

2.3.2 Server-side validation of the incoming traffic – Security protocols

Following the technical specifications, the data integrity was focused on three aspects:

- Error detection / data validation to overcome errors in data transmission;
- Integrity check of the transmitted documents (to ensure conformity) with a JSON Schema Validator;
- Data backup and duplication to prevent data loss prevention.

Error detection in data transmission

The security protocols installed on the DBMS are to ensure that the incoming traffic is securely transmitted without any loss of information. This covers all traffic (both in- and out-going) of the server. To do so, the DBMS makes use of TLS v1.3.

It should be noted that this is distinct from the integrity and validation checks that take place specifically for the geo-tagged photos (see Section 2.2 - Farmer's geo-tagged photos framework). The validation described herein is general and concerns every transaction that takes place.

The TLS is a widely adopted security protocol designed to facilitate privacy and data security for communications over the Internet. A primary use case of TLS is encrypting the communication between web applications and servers, such as web browsers loading a website. In a nutshell, it encrypts data sent over the Internet to ensure that eavesdroppers and hackers are unable to see what you transmit which is particularly useful for private and sensitive information such as passwords, credit card numbers, and personal correspondence. TLS 1.3 is the latest specification which has made TLS faster and more secure. It is faster since TLS handshakes now only require one round trip (or back-and-forth communication) instead of two, and it is safer because it removes obsolete and insecure features from TLS 1.2.

HTTPS is an implementation of TLS encryption on top of the HTTP protocol, which is used by all websites as well as some other web services. Any website that uses HTTPS is therefore employing TLS encryption. A website with an HTTPS address has a legitimate SSL certificate issued by a certificate authority, and traffic to and from that website is authenticated and encrypted with the SSL/TLS protocol.

There are three main components to what the TLS protocol accomplishes: Encryption, Authentication, and Integrity.

- Encryption: hides the data being transferred from third parties.
- Authentication: ensures that the parties exchanging information are who they claim to be.
- Integrity: verifies that the data has not been forged or tampered with.

It should thus be noted that TLS does not secure data on end systems. It simply ensures the secure delivery of data over the Internet, avoiding possible eavesdropping and/or alteration of the content.

The essential principles that govern how TLS works are the following:

- Secure communication begins with a TLS handshake, in which the two communicating parties open a secure connection and exchange the public key.
- During the TLS handshake, the two parties generate session keys, and the session keys encrypt and decrypt all communications after the TLS handshake.
- Different session keys are used to encrypt communications in each new session.
- TLS ensures that the party on the server side, or the website the user is interacting with, is actually who they claim to be.
- TLS also ensures that data has not been altered, since a message authentication code is included with transmissions.

With TLS, both HTTP data submitted by the users to the server and the HTTP data that the server sends to users are encrypted. Encrypted data must be decrypted by the recipient using a key.

The digital certificate installed in the server was issued by Let's Encrypt. This is a free, automated, and open certificate authority, run for the public's benefit, provided by the Internet Security Research Group. The TLS v1.3 was installed using the SHA-256 with RSA Encryption signature algorithm. The automatic renewal was enabled to ensure the certificate is continuously updated.

The DBMS is exposed to the internet via the https://dione.iccs.gr/ API endpoint. The SSL/TLS report for this endpoint is given in the Appendix B.

JSON Schema validation

JSON Schema specifies a JSON-based format to define the structure of JSON data for validation, documentation, and interaction control. A JSON Schema provides a contract for the JSON data required by a given application, and how that data can be modified. The benefits are:

- It describes the existing data format(s).
- It provides clear human- and machine- readable documentation.
- It validates data which is useful for:
  - Automated testing.
  - Ensuring quality of client submitted data.

On top of supporting JSON schema validation of MongoDB, RESTHeart also provides a general approach for validation based on Interceptors that can verify write requests based on any condition. That is, when new data are transmitted to the server and are about to be inserted in a given collection, the interceptor validates the body of the write request against a JSON schema. The additional advantages are that:

- Schemas are stored in a special collection, the schema store /_schemas and are validated
- Schemas can and can be reused on multiple collections
- Complex schemas can be defined using sub-schemas, i.e., using the $ref keyword
- Documents can be validated using schemas that available online
- Schemas are cached

RESTHeart returns a 400 Bad Request error code if a schema violation is found.

Figure 31 provides a quick graph overview of the schema used for the JSON data of the MEMS measurements. It is mostly concerned with the structure of the document and with the validity of the data (e.g., reflectance values cannot be negative, min and max wavelengths should be positive integers, etc.).

Figure 31: Quick graph overview of the schema used for the MEMS JSON data

Data backup and duplication

The DBMS is hosted on a Linux VM which uses the ZFS file system that in effect combines both the file system and a volume manager. One major feature that distinguishes ZFS from other file systems is that it is designed with a focus on data integrity by protecting the user's data on disk against silent data corruption caused by data degradation, power surges (voltage spikes), bugs in disk firmware, phantom writes (the previous write did not make it to disk), misdirected reads/writes (the disk accesses the wrong block), DMA parity errors between the array and server memory or from the driver (since the checksum validates data inside the array), driver errors (data winds up in the wrong buffer inside the kernel), accidental overwrites (such as swapping to a live file system), etc. Moreover, instead of hardware RAID, ZFS employs "soft" RAID, offering RAID-Z (parity based like RAID 5 and similar) and disk mirroring (similar to RAID 1).

In terms of data backup, there is an automated process ensuring that the data are regularly duplicated to a separate physical location over the Internet. As soon as data are ingested into the main database, they are automatically duplicated to a server hosted in i-BEC premises.

2.3.3 Authentication and authorization

Authentication and authorization are the two notions that might sound similar but have significant differences between them. Authentication is the procedure of recognizing the user's identity, whereas authorization is the procedure of providing partial or foul access to particular resources. In simple terms, authentication is the process of verifying who a user is, while authorization is the process of verifying what they have access to. The authentication determines whether users are who they claim to be and challenges them to validate their credentials (for example, through passwords, answers to security questions, or facial recognition). On the other hand, authorization determines what users can and cannot access and verifies whether access is allowed through policies and rules.

When a new request reaches the RESTHeart interface, the first step is to ensure that the request is authenticated and authorized (Figure 32). The Authenticator verifies the credentials extracted from the request by Authentication Mechanism. The authentications mechanisms used are the token authenticator from the DIONE toolbox (for single sign on purposes), and Basic Auth (username / password combination). The Authorizer is responsible of checking if the user can actually perform the request against an Access Control List. For example, if the authenticated user is indeed authorized to perform the requested action on a specific collection.

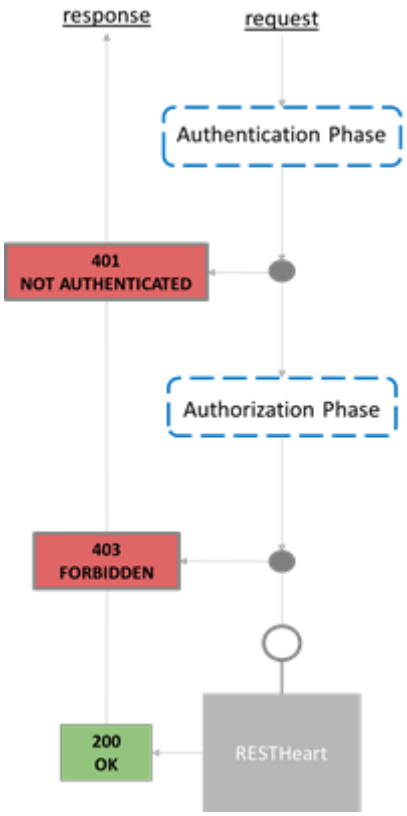


Figure 32: Requests in RESTHeart should be both authenticated and authorized.

In short, access to a resource is protected by both authentication and authorization. If you can't prove your identity, you won't be allowed into a resource. And even if you can prove your identity, if you are not authorized for that resource, you will still be denied access.

Authentication

To authenticate users, the DIONE toolbox API was used to enforce single-sign-on. Basically, when data are transmitted to RESTHeart it makes sure that the transmitted token key is valid.

To this end, a custom Authenticator java plug-in was developed and installed and was used in all the data collections. The workflow of this plug-in is depicted in Figure 33.

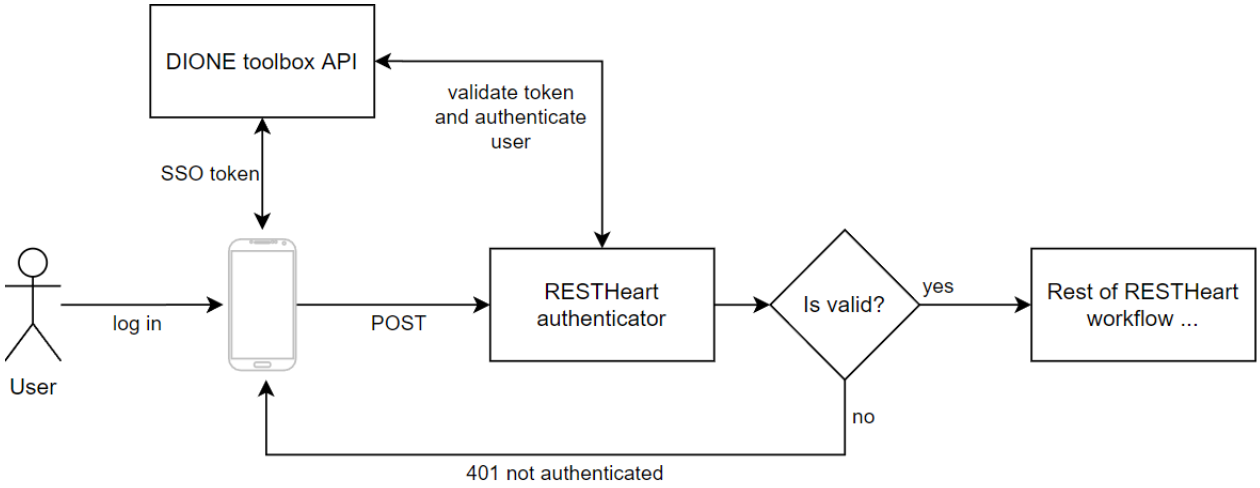


Figure 33: Custom authenticator plug-in to use DIONE's SSO approach

Authorization

For authorization purposes, the following restrictions were applied:

- Users authenticated via a DIONE toolbox token are allowed to POST, PUT and PATCH documents in the respective collections for DIONE. They are not allowed to GET their submitted data directly. Users are allowed to PUT and PATCH only documents that belong to them.
- A specific user instance has been created who is the only one (excluding the administrator) that can be authenticated via the basic authentication mechanism (using a username / password combination). This user has only GET permissions on the collections generated for DIONE and can be used to retrieve the data from the DB.

These were implemented using the File ACL Authorizer of RESTHeart to define the Access Control List in each MongoDB collection.

2.3.4 Statistical analysis for outlying behaviour

Outlier detection is an analysis for identifying data points (outliers) whose feature values are different from those of the normal data points in a particular data set. Outliers may denote errors or unusual behavior. Detection and removal of outliers in a dataset is a fundamental pre-processing task without which the analysis of the data can be misleading. Furthermore, the existence of anomalies in the data can heavily degrade the performance of machine learning algorithms.

Generally speaking, outlying detection techniques may be categorized under unsupervised and supervised. In the former, the training data contain only known valid data (inliers) whereas in the latter the training data contain both inliers and outliers. The unsupervised approaches may still be further categorized in further sub-categories depending on the nature of the technique used.

Considering that for DIONE's case, as data with outlying behavior can be characterized all non-soil objects that a user might "accidentally" try to post to the database, we created a collection of 195 spectral signatures sourcing from every-day objects. Both supervised and unsupervised classifiers have been assessed for their potential to distinguish soil and non-soil targets, with the following models tested:

- KNN classifier
- Support Vector Machine
- Gaussian Process Classifier
- Decision Tree
- Random Forest
- MLP
- AdaBoost classifier
- Quadratic Discriminant Analysis

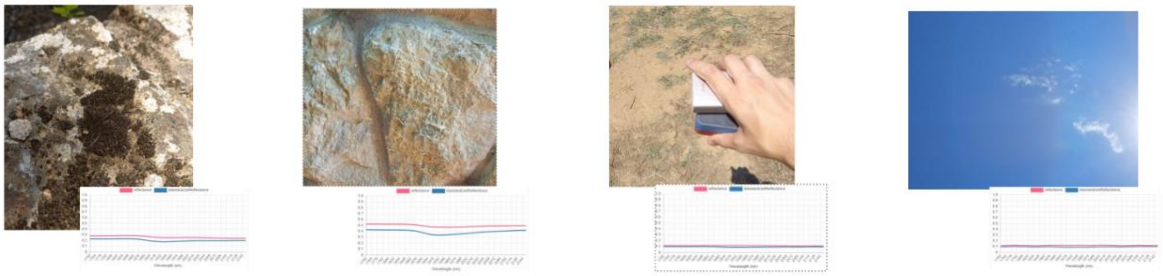


Figure 34: Collection of non-soil objects for outlier dataset creation

Each model was evaluated through common classifying accuracy metrics such as precision, accuracy, and recall with the ensemble of the above to provide best results with an accuracy of 94.47%, a precision of 95.27% and recall score of 97.58%. The deduced confusion matrix contains 9 erroneous classifications, from which 6 are false positives and 3 are false negatives as shown at Table 1, while the classification outcomes from the ensemble method are presented at Figure 35.

Table 1: Confusion matrix of ensemble method for identifying measurements with outlying behavior.

		Actual value	
		True	False
Predicted Value	True	33	6
	False	3	121

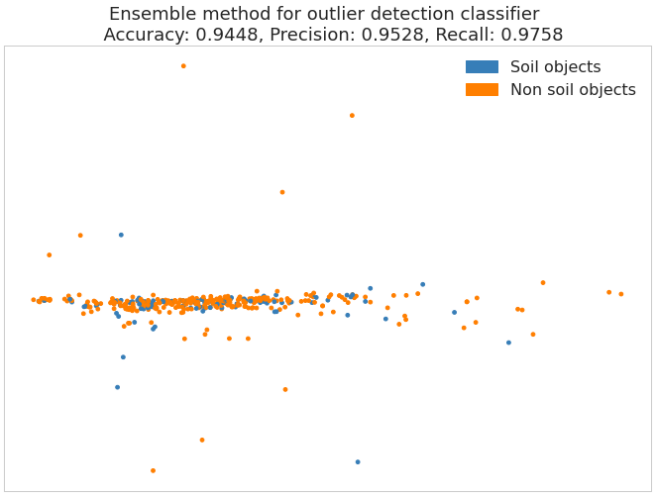


Figure 35: Classifier's estimations

2.3.5 Novelty detection mechanism

Novelty detection is different from outlier detection in the sense that in the former the goal is to determine whether new observations fit within the existing data set, whereas in the latter it is to determine if they are from a different domain. It thus refers to the identification of novel or abnormal patterns embedded in a large amount of normal data. To put it simply, in the particular use case employed herein, outlier detection can identify spectra that are captured from non-soils (e.g., vegetation, water) that may be recorded due to operator error, whereas novelty detection is about determining whether a given soil spectrum is represented well enough in the dataset of known soil spectra. This means that soils from a new region in the world that may be completely different from the ones contained in the libraries that are used in DIONE's models will be flagged as novel. This is particularly useful because the ML models will estimate the soil properties from the spectra are agnostic as to the origin of the recorded spectrum and will extrapolate to account for this discrepancy – which may lead to precarious estimations.

Put in a way that is more familiar to data scientists, consider a data set of observations from the same distribution. Assume now that we add one more observation to that data set. Is the new observation so different from the others that we can doubt it is regular (i.e., does it come from the same distribution?)? Or on the contrary, is it so like the other that we cannot distinguish it from the original observations? This is the question addressed by the novelty detection tools and methods.

The techniques usually employed to identify novel patterns may be broadly summarized as:

- The ML approach that uses one-class classifiers, aiming at capturing characteristics of training instances, to be able to distinguish between them and potential novel patterns to appear. This is different from and more difficult than the traditional classification problem, which tries to distinguish between two or more classes with the training set containing objects from all the classes.

- The nearest neighbour approach, based on the assumption that normal points tend to have close neighbours while novelty points are located far from the distribution.
- The clustering-based approach, which assumes that normal data belong to large and dense clusters, whereas novel data do not belong to any of those clusters. Basically, for each data point a degree of membership is assigned to each of the clusters; novel patterns are samples that do not belong to any of those clusters.
- The statistical approach which makes use of stochastic distributions to model the data.

For DIONE, the following methodologies were tested:

- One-class SVM. SVMs are max-margin methods, i.e., they do not model a probability distribution. Here the idea is to find a function that is positive for regions with high density of points, and negative for small densities. Compared to the traditional SVMs that use a hyperplane to separate two classes with the largest possible margin, the one class-SVM uses a hypersphere to encompass all the instances. The margin here refers to the outside of the hypersphere — so by "the largest possible margin", in this case we mean "the smallest possible hypersphere".
- The elliptic envelope approach that fits a robust covariance estimate to the data and thus fits an ellipse to the central data points, thereby ignoring points outside the central mode.
- The isolation forest approach which is an efficient way of performing outlier detection in high-dimensional datasets by using random forests. In essence, the forest isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.
- The local outlier factor algorithm which computes a score (called local outlier factor) reflecting the degree of abnormality of the observations. It measures the local density deviation of a given data point with respect to its neighbours. The idea is to detect the samples that have a substantially lower density than their neighbours. In practice the local density is obtained from the k-nearest neighbours.

We tested the methodologies by utilizing the GEO-CRADLE and LUCAS open soil spectral libraries and creating various scenarios whereby the known dataset belonged to one (or more) country (-ies) and the novel data originated from different country (-ies). In essence, a simulation was conducted to create the scenarios where soil spectra from a new region should be identified as novel, based on the information contained in the dataset. It should further be noted that the spectral resolutions were constrained so as to much those of the in situ soil scanning system. For all the models we tested various hyperparameters including:

- The gamma parameter of the RBF kernel used in SVM, selected from {0.025, 0.05, …, 0.20}
- Random states in isolation forest
- Number of neighbors in the local outlier factor algorithm selected from {5, 10, …, 50}

The average accuracy across the datasets per each methodology (after hyperparameter optimization) is presented in Table 2. Based on these results the local outlier factor methodology was selected as it presented the more robust performance metrics. Of course, given that novelty detection is a difficult task, the output of this best model should only be an indication to the modeler / expert, who will ultimately decide if this sample's properties may be accurately estimated by the model that follows.

Table 2: Average performance metrics in the independent test set for the novelty detection module per each learning methodology.

Method	Avg. Accuracy	Avg. Precision	Avg. Recall
Robust covariance	0.7469	0.9853	0.6678
One-class SVM RBF	0.7199	0.9561	0.6512
Isolation Forest	0.7224	0.9653	0.6478
Local Outlier Factor	0.7813	0.9015	0.7907

2.3.6 Cross-device standardization

A group of different devices will be involved to the acquisition of in situ spectral measurements. To this end, an internal standard needs to be used to minimize systematic effects induced by different devices. For this purpose, two soil samples were used as Internal Soil Standards (ISS) as proposed at Ben Dor et al. (2015) that have been proven to be stable in space and time. These two sand dune samples, which are collected from Wylie Bay and Lucky Bay coastlines of South-Western Australia, are made of about 90% quartz, are fine soil samples that is proven to maintain their reflectance characteristics both at Vis and NIR spectral region along time and can be used both for radiometric and spectral calibration. Their spectral signatures are featureless, meaning that any correction performed to any other soil spectral signature will not introduce artifacts related to the existence of non-soil materials (i.e., minerals). Their spectral curves are shown at Figure 36.

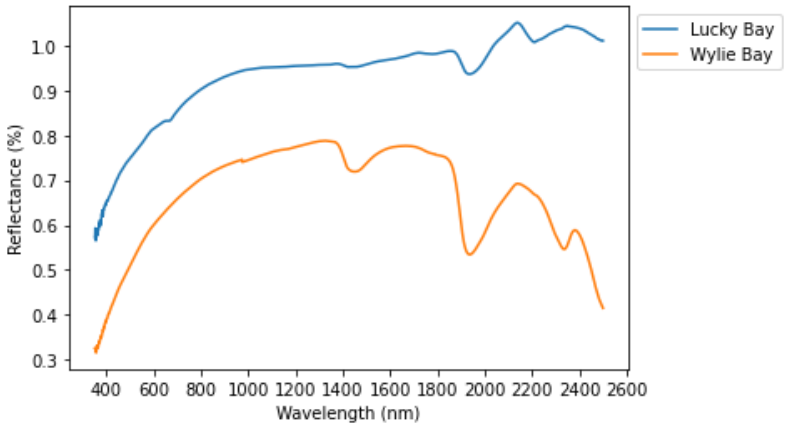


Figure 36: Spectral signatures of Wylie Bay and Lucky Bay fine dunes

A self-standardization approach has been followed for each of the portable soil spectrometers, according to which, Wylie Bay and Lucky Bay samples were measured with each spectrometer. Then the correction factors (CF) for each device were obtained through the calculation of ratio of standards' MEMS reflectance (S_ρ) to benchmark standards' reflectance (SBM_ρ) per wavelength (λ), according to the following formula:

$$Cf(\lambda) = \frac{SBM(\lambda)}{S_\rho(\lambda)}$$

The corrected reflectance ($R_c(\lambda)$) will be the result of the multiplication of soil sample's reflectance ($R_0(\lambda)$) with the derived CF as follows:

$$R_c(\lambda) = R_0(\lambda) \times CF(\lambda)$$

The calculated correction factors shaped a collection at the DBMS and for every incoming measurement, the corrected reflectance is calculated and stored to the measurement's record. Each instance of the collection containing the CFs is uniquely connected to a single device through matching the unique device ID labelled as "instrument". The ISS' spectral curves as measured from each presented very low variability per wavelength (Figure 37), indicating that they were well selected for their purpose.

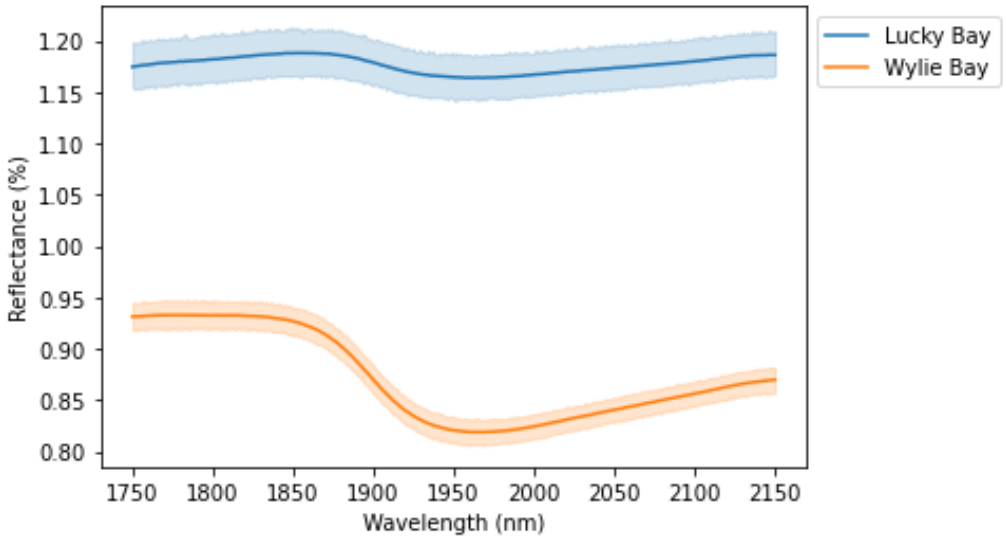


Figure 37: Wylie Bay and Lucky Bay spectral signatures

The application of the self-standardization was proved to decrease by an order of magnitude the spectral distance between MEMS collected spectral signatures to reference spectral signatures collected by standard laboratory equipment. Specifically, a collection of 100 archived samples were measured with and without standardization and compared to reference spectral[18] (Figure 38), since the raw MEMS reflectance error compared to reference values ranged from 0.11% to 0.18% while the error corresponding to standardized MEMS measurements dropped significantly from 0.01% to 0.04%.

[18] Reference spectral signatures were acquired with Spectral Evolution PSR+ 3500 instrument (Spectral Evolution Inc., Lawrence, MA, USA) covering the electromagnetic spectrum from 350 to 2500 nm and with spectral resolution of 2.8 nm @ 700 nm; 8 nm @ 1500 nm; 6 nm @ 2100 nm

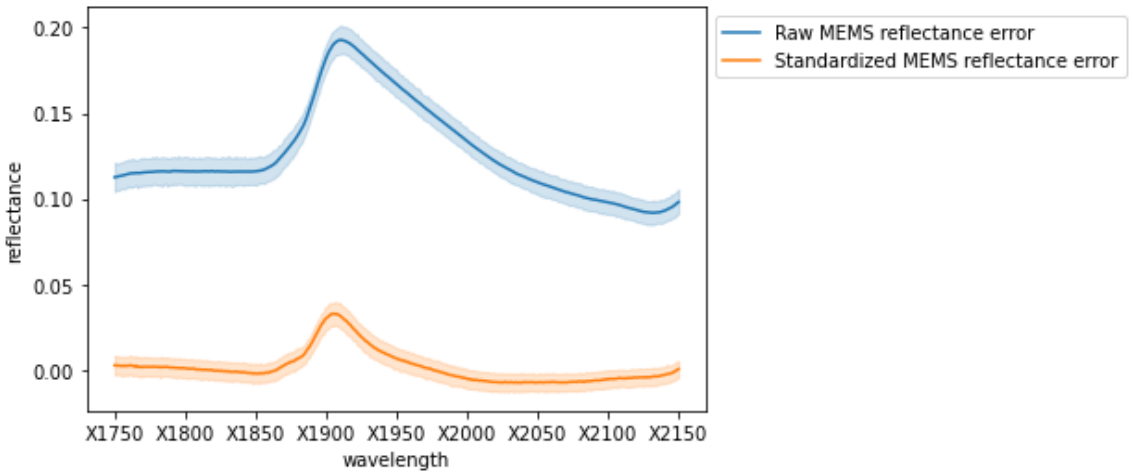


Figure 38: Mean differences from reference spectrum (PSR+3500) to MEMS measurements before (Raw) and after standardization (Standardized)

2.3.7 Restheart API Custom plugins

RESTHeart comes with a complete MongoDB API and a security implementation that allows to authenticate users and authorize requests according to a role-based policy. These core functionalities are provided and distributed with RESTHeart as a set of standard plugins. To further extend the functionalities and meet DIONE user requirements, a wider set of custom plugins have been developed and included to the data processing and storage system, covering Geo-tagged photos plugin and Spectra interceptor plugin.

Interceptors allow to snoop and modify requests and responses at different stages of the request. At the stage of json post from the mobile application, the following jsonInterceptors have been implemented:

Standardization interceptor

Transmitted reflectance from the mobile application to the DBMS will be corrected according to ISS as described at section 2.3.6. To achieve this, through the Standardization interceptor, the device identifier denoted as “name” from the “device” schema is matched to the corresponding device from the collection containing the correction factors (labelled as “mems.standardization”) and retrieves the correction factors for the calculation of the standardized reflectance. Then, derived array of standardized reflectance is stored to the “scans” schema as an array to the “standardizedReflectance” schema.

Outlier detection interceptor

The outlier detection classifier was developed with the use of Keras library that is deployed as a custom plugin. It retrieves the value of the “standardizedReflectance” schema and classifies the sample as an inlier or an outlier with the mechanism described at section 2.3.4 by assigning a logical value to the “outlier” key under the “Analysis” schema. Samples that are recognized as outliers or as results of erroneous measuring methodology are not further processed by the analysis pipeline.

Novelty detection interceptor

The novelty detection architecture is similar to the outlier detection interceptor. It was developed with Keras library classifying the analyzed samples as novels for instances that are originated from soils that are not similar to the ones included to the developed Soil Spectral Library (SSL). As previously described, the array stored to the "standardizedReflectance" schema is used as input to the Novelty detection classifier as described at section 2.3.5 and the classification is stored as a logical value at the "novel" key under the "Analysis" schema.

ML prediction interceptor

The targeted properties are derived as results of machine learning models having as set of predictors the bands over which the topsoil reflectance is captured. A set of models along with different pre-processing techniques have been evaluated and as described at section 4.2.4 the most suitable ones have been developed as regressors with the help of Keras library. The interceptor after validating that the incoming sample is not an outlier and not a novel instance, with the help of the developed model performs an estimation of each property and stores the results to the "Analysis" schema under the corresponding property. An output example of the custom plugins interceptors to the JSON document follows

Analysis: {"outlier": False, "novel": False, "SOC": 5.3, "Clay": 13.2, "Sand": 43.9, "Silt": 42.9, "pH": 7.56, "CaCO3": 4.32}

3 In situ tools for complementing EO data

3.1 In situ component as a crowdsourcing campaign

With the development of the SSS and after thorough laboratory test in terms of measurements' quality and reproducibility, an extensive field mission was planned in coordination with Lithuanian and Cypriot paying agencies (NPA and CAPO), and was conducted during the growing season of 2021 for the application of the developed methodology. A set of soil scanners was delivered to various end users, (soil surveyors) with no prior explicit expertise in soil spectroscopy along with the mission to crowdsource the developed database management system. All methodologies to be followed from end users were initially assessed by local farmers during a testing field mission that took place between May 18th and May 31st 2021, around Imathia region, Western Greece. During the conducted testing activities, five farmers were trained on how to operate the SSS and they were assigned with 50 measuring locations each. With the completion of this task, each of the farmers were asked to provide feedback regarding the operation of the SSS, and general aspects of the procedure. Their responses were taken in consideration in developing the final methodology that would be implemented by NPA and CAPO soil surveyors.

Figure 39: Training of local farmers at a Kiwi farm - Dio Gkortsies region, Wester Macedonia, Greece

The two Paying Agencies indicated three sub regions each with different soil characteristics located at Lithuania and Cyprus respectively. The pilot areas' extent is 8.32km^2 and 9.30km^2 for Cyprus and Lithuania respectively and are shown at Figure 40.

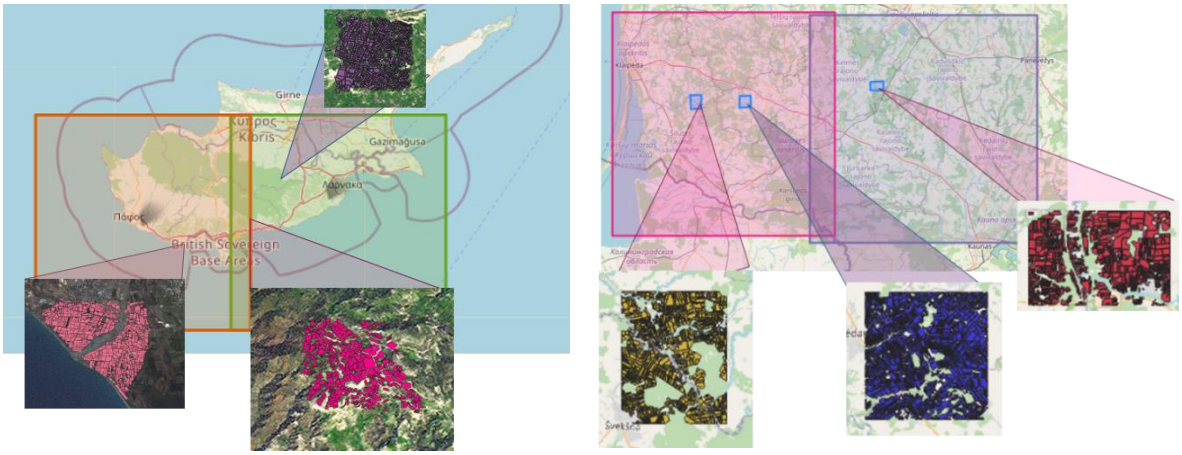


Figure 40: Pilot regions

The objective of the preparatory tasks of the mission planning was the selection of sampling points to be visited for the soil scanning. The methodological framework that was developed and followed is based on Sentinel-2 composites analyses for pixel clustering based on spectra related similarities. To this end, a three year long timeseries of Sentinel-2 reflectance observations was compiled for each pilot sub-region and according to the methodology proposed by Demattê et al. (2009) the bare soil composite was extracted and used as a regional mask. The Sentinel-2 multi-temporal values corresponding to bare soil were then converted to median values per pixel and combined with ancillary variables related to physicochemical and topographical variables retrieved from open access data sources. The derived cube shaped dataset was then clustered with the application of k-means clustering algorithm in order to derive sub-area clusters with similar spectral, physicochemical and topographical properties. The conclusive part of the pilot area preprocessing was the application of Conditioned Latin Hypercube Selection (CLHS) algorithm (Minasny & McBratney, 2006), which is commonly applied for subset selection among different sampling points when information about ancillary variables is available. This algorithm was applied twice to the extent of each sub-area, once for the selection of the points to be visited from soil surveyors and the second time for the selection of a validation sub-sample among the points selected during the first iteration of CLHS algorithm.

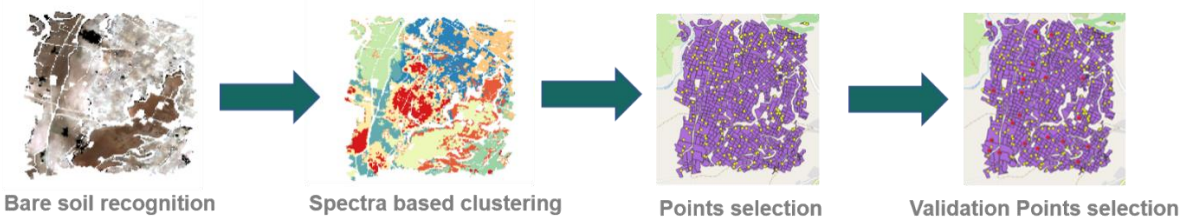
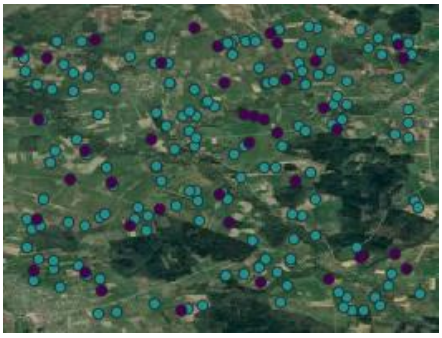


Figure 41: Sampling locations selection methodology

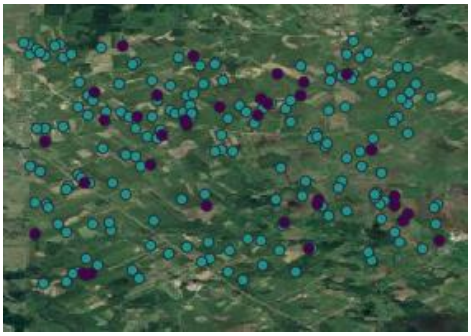
With the completion of the procedure, 200 point locations per region were indicated with 20% of them to be characterized as calibration points. All points were visited from soil surveyors and scanned with the SSS while a small bag with topsoil sample was collected from the validation points for chemical analysis (see section 4.2.3). Each point location had been assigned a unique id to facilitate data cataloguing of later stage. The assigned id is of the form XX_YY_ZZZZ where:

- XX denoting the country code (LT for Lithuania and CY for Cyprus)
- YY denoting the subarea code. Cypriot sub areas are Episkopi (EP), Leukara (Le) and Agia Varvara (AG) while Lithuanian are Western (W) Central (C) and Eastern (E) regions.
- ZZZZ denoting a unique location's number ranging from 1 to 600 for Cypriot locations and 1000 to 1600 for Lithuanian ones.

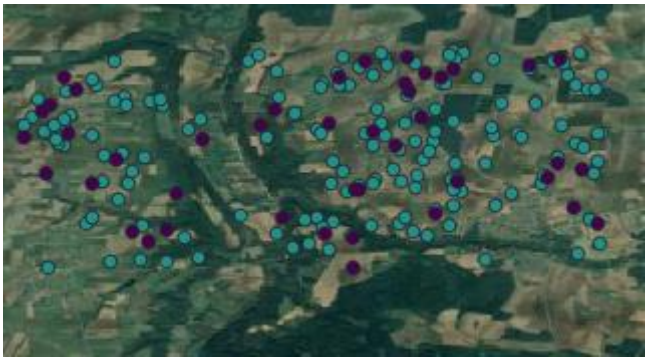
The spatial distribution of sampling locations can be found at Figure 42.



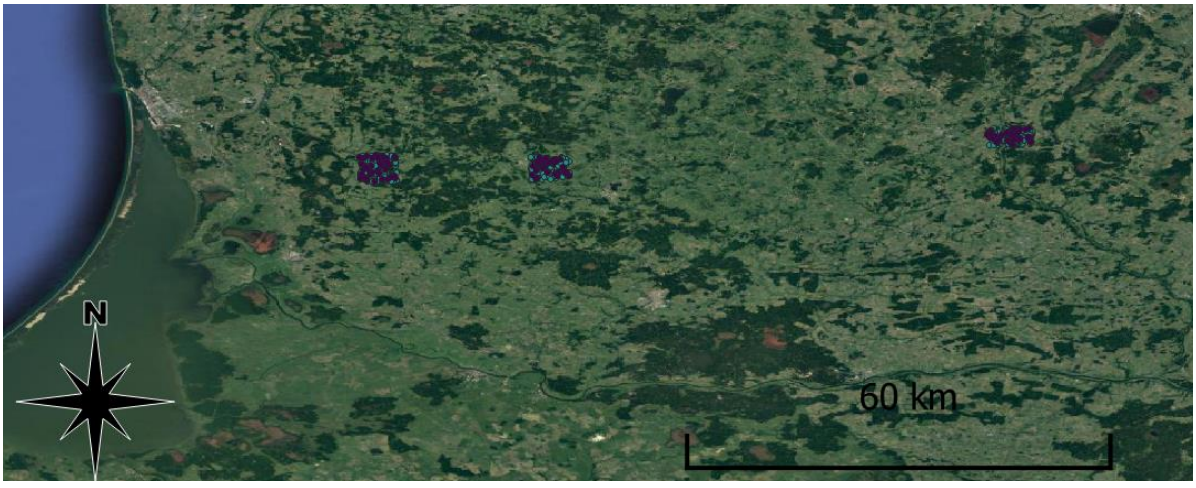
(a)



(b)

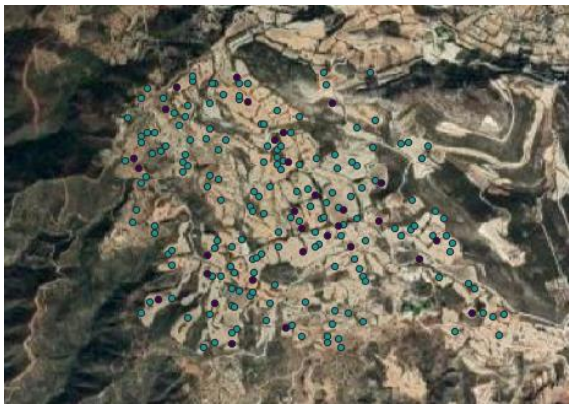


(c)

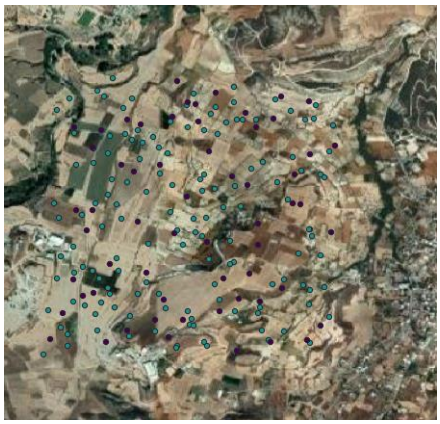


(d)

(e)



(f)



(g)

(h)

Figure 42: Spatial distribution of point sampling locations. With purple are denoted the calibration points and with green the points that were measured only in situ with the SSS. (a) Western Lithuania – (b) Central Lithuania – (c) Eastern Lithuania – (d) General aspect of Lithuania – (e) Episkopi region, Cyprus – (f) Leukara region Cyprus – (g) Agia Varvara region, Cyprus – (h) General aspect of Cyprus

Prior to the initiation of the pilot activities, two training webinars, one for each paying agency, were conducted, dedicated to the demonstration of SSS use case scenarios and installation of the mobile application to each soil surveyor's mobile device. During the implementation of the pilot activities, real time support was provided via the creation of a group chat over a commercial cross-platform voice over IP and instant messaging software application. The pilot activities were concluded at September 30, 2021 with the shipping of the portable spectrometers and soil samples back to i-BEC. The devices' post usage measurement stability was tested while the soil samples were analyzed at the chemical laboratory. For the above-described pilot activities, 15 different SSS were utilized under the same in situ usage protocol, as described at section 2.1.4.

(a) (b)

Figure 43: (a) In situ usage of SSS, (b) Soil sampling at Lithuanian Eastern region

3.2 Historical Soil Spectral Libraries scheme

Open-access SSLs have enabled a data-driven approach to effectively describe soil, both qualitatively and quantitatively, finding robust statistical models between laboratory spectral signatures and soil properties (Ballabio et al., 2016; Nocita et al., 2015). To this end, the collection of in situ topsoil reflectance measurements combined with chemical analyses of the soil properties analysed to the laboratory shaped an SSL that can be used for the developed of such models. This SSL was extended through merging the most up-dated and historical European datasets, which are the European Soil Data Centre (ESDAC) Land Use and Cover Area frame Survey (LUCAS) and the GEO-CRADLE Soil Spectral Library.

LUCAS is the result of the triennial survey of land use as conducted by the statistical office of the European Union (EUROSTAT), which further to a wide set of key soil properties, it also contains since 2009 topsoil reflectance analyses of more than 20.000 (Tóth et al., 2013) samples distributed over 23 EU member states. The topsoil sampling locations were selected using a Latin hypercube-base stratified sampling design from the LUCAS master sample grid of 2 km by 2km, (Castaldi et al., 2019). This topsoil survey was an attempt to build a consistent database using standard sampling and

analytical techniques, where the analysis of all soil samples was carried out by a single chemical laboratory.

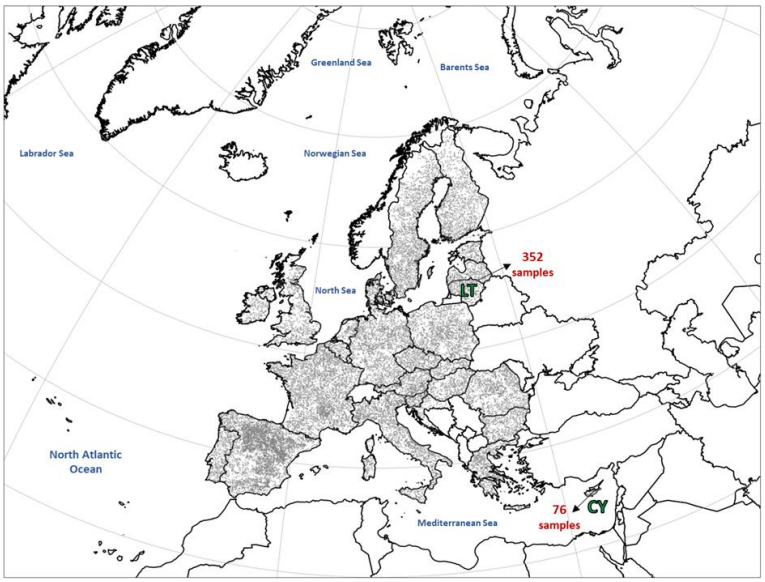


Figure 44: Overlap of LUCAS 2015 points around Europe region. DIONE pilot areas (Lithuania and Cyprus) are also appear in the map with the total number of samples.

The GEO-CRADLE SSL was developed in the context of the GEO-CRADLE EU funded project[19]. For the first time an open and standardized regional SSL was developed, as a complement to the EU Soil Sample Data Base. The GEO-CRADLE SSL was created quite recently and contains 1754 soil samples and their corresponding soil properties (SOC, Texture, $CaCo_3$, CEC, NO_3, pH) in a very well spatial distributed area across nine countries in the North Africa, Middle East and Balkan region, comprising soils of 18 soil classes of the world (Figure 45). The samples were selected from national soil data archives or collected through field surveyors during the project's lifetime, following general guidelines, standards, and protocols to ensure consistent data collection and analysis. The GEO-CRADLE SSL is in compliance with GEOSS data principles and Open Database License standards and is publicly available through the GEO-CRADLE's project regional datahub[20].

[19] http://geocradle.eu/en/
[20] http://datahub.geocradle.eu/dataset/regional-soil-spectral-library

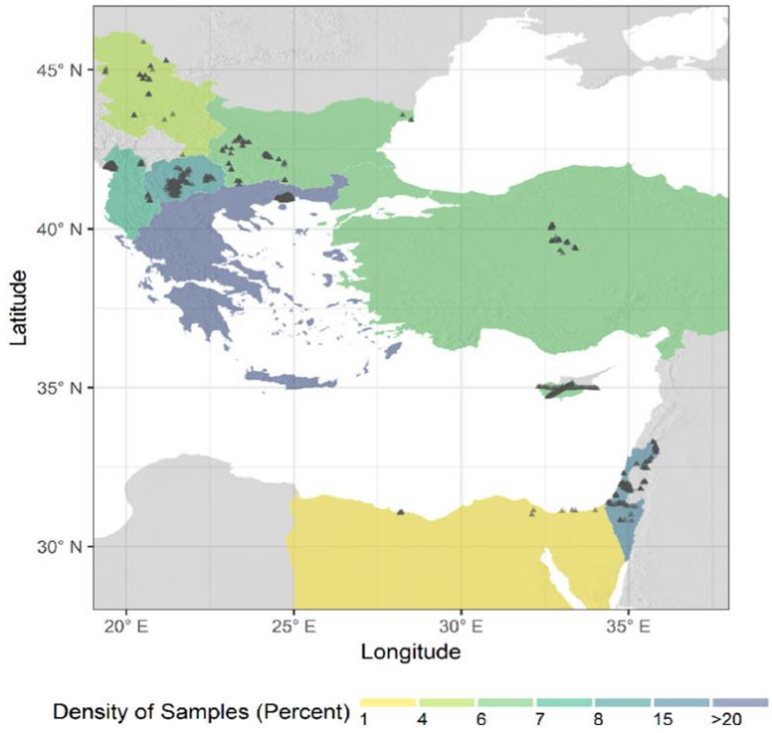


Figure 45: Location of the 1754 sample sites with reflectance spectra in the GEO-CRADLE SSL

Given the sufficient spatial distribution and representatives of LUCAS and GEO-CRADLE data archive across the European continent, we leverage on the reference soil data as the source for the soil texture information. In addition, these datasets include laboratory soil spectral samples (VNIR) that cover the same spectral range of the Soil Spectrometers, and enabling the model calibration for soil properties estimation. In the same context, the modelling procedure is largely supported, by allowing the resampling process to be performed in order to achieve the spectral configuration of the MEMS in accurately and reliably way.

3.3 EO data scheme

Topsoil reflectance measurements collected with the Soil Spectrometers can be transformed to point estimations through the calibration of models taking advantage of the underlying correlations between soil spectra and physical and chemical values measured in the laboratory. These estimations are intended to be up-scaled over the parcels that lie inside the pilot areas and there is lack of reference measurements. Towards this direction, a collection of heterogeneous EO multi-temporal data was set together over the centroids of declared parcels of the pilot areas, containing both satellite optical imagery, radar and existing soil property layers. The rationale behind this data collection is to develop machine learning calibration models over the point estimations derived from in situ spectra of soil spectrometers, and thus based on few physicochemical analyses performed in lab produce soil indicators over wide areas, which in case of DIONE is covered by 3987 parcels covering more than 22 km^2 in total.

3.3.1 Data sources and mining mechanisms of historical EO data

Data were extracted from data hubs provided free of charge by space agencies or specific departments that developed, monitored and managed the satellites and instruments, such as Copernicus, USGS, NASA, among others[21]. This wide set provided observations with various spatial and temporal resolutions that were fused to a unified dataset. The data sources and mining mechanisms used are described to the following sub-section.

3.3.1.1 Third party data hubs

Copernicus is the European Union's Earth Observation Programme that is funded, coordinated and managed by European Commission and implemented in partnership with the Member States, the European Space Agency (ESA), and European Organization for the Exploitation of Meteorological Satellites (EUMETSAT), among others.

The Copernicus program contains a group of dedicated satellites and contributing missions. The Sentinel satellites are specifically designed to meet the needs of the Copernicus information services and their users. Based on satellite and in situ observations, the Copernicus services deliver near-real-time data on a global level.

Today, the European Union through the Copernicus program has seven satellites in orbit, being the Sentinel family of satellites. The first satellite, Sentinel-1, was launched on 3 April 2014 and Sentinel-1B in 25 April 2016 for land and ocean services. The Sentinel-1 provides data from a dual-polarization C-band Synthetic Aperture Radar (SAR) instrument.

Sentinel-2A was launched on 23 June 2015 and Sentinel-2B on 7 March 2017. Sentinel-2 is a wide-swath, multispectral high-resolution for land monitoring and provide imagery from vegetation, soil and water cover with a global 5-day revisit frequency. This satellite has 13 spectral bands, such as visible and NIR (10 meters of spatial resolution), red edge and SWIR (20 meters) and atmospheric bands (60 meters). Due these characteristics, Sentinel-2 provides high quality data, covering all planet and high spatial-temporal resolution. Afterwards, five more versions were launched (Sentinel 3,4,5,6), specific for atmospheric variables, such as gases and aerosols and oceanographic variables.

Landsat is a program that has been collecting Earth data since 1972 (Landsat 1). Landsat 2 was launched in 22 January 1975, Landsat 3 in 1978, Landsat 4 in 1982, Landsat 5 in 1984, Landsat 6 in 1993, Landsat 7 in 1998 and the last, Landsat 8 was launched in 11 February 2013[22]. It is a joint of United States Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA) program that captures Earth images with a spatial resolution of 30 meters and temporal resolution once every 16 days.

[21] https://www.copernicus.eu/sites/default/files/Brochure_Copernicus_2019%20updated_0.pdf

[22] https://pubs.usgs.gov/fs/1997/0084/report.pdf

The Landsat 8 acquires around 740 scenes a day with dimension of 185 km x 180 km has eight spectral bands such as visible, near-infrared and short-wave infrared and a thermal infrared band, processed for orthorectified surface temperature[23].

The Moderate Resolution Imaging Spectroradiometer (MODIS) instrument is operating on two satellites, on the Terra and Aqua satellites and is coordinated and managed by Land Processes Distributed Active Archive Center (LP DAAC) that operates as a partnership between USGS and NASA. The LP DAAC processes, archives, and distributes land data products to the users in the earth science community.

The Terra satellite was launched on December 18, 1999 and the Aqua on May 4, 2002. It has a temporal resolution between 24 and 48 hours and three spatial resolutions, 250 meters, 500 m and 1000 m. These instruments contains 36 bands, being 11 bands in the visible region, 6 in near-infrared region, 3 in short-wave infrared, 10 in thermal-infrared and 6 in Mid-wave infrared. They are designed to provide measurements in large-scale global dynamics. It was used The MOD11A1 V6 product that provides daily Land Surface Temperature (LST) and emissivity values in a 1200 x 1200 km grid.

Another product utilized was the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS), which contains a series of land surface parameters. There are several different FLDAS datasets and this was simulated from the Noah 3.6.1 model with CHIRPS-6 hourly rainfall that has been downscaled using the NASA Land Surface Data Toolkit. FLDAS is a custom model of NASA Land Information System that was built and adapted to support developing countries with regard to food security[24]. The aim of FLDAS is to ensure a more democratic and effective use of existing and available hydroclimatic data, with this, the adoption of LIS allows FEWS NET to take advantage of these existing models and, based on meteorological models, generate sets of soil temperature, soil moisture and evapotranspiration, among others[25].

The Digital Elevation Model was collected through the Shuttle Radar Topography Mission (SRTM) that started operated at 11 February 2000. The SRTM reached at space aboard the space shuttle Endeavour. This project was developed and managed by NASA and National Geospatial-Intelligence Agency (NGA) to create the first near-global set of land elevations.

SRTM provides a breakthrough in the accessibility of high-quality elevation data (Digital elevation models – DEM) from around the world. Aboard Endeavour, the SRTM orbited Earth 16 times each day the 11 days, collecting radar data over 80% of the Earth's land surface.

3.3.1.2 Datacube

The abovementioned data sources provided multi-temporal layers of observations that were collected and stored as a datacube over each study area. In the context of computer science, a datacube is a data structure containing a multidimensional array of values. When it comes to EO data

[23] https://www.usgs.gov/core-science-systems/nli/landsat
[24] https://disc.gsfc.nasa.gov/datasets/FLDAS_NOAH01_C_GL_MA_001/summary
[25] https://hydro1.gesdisc.eosdis.nasa.gov/data/FLDAS/FLDAS_NOAH01_C_GL_M.001/doc/README_FLDAS.pdf

representation, datacube might contain one or more spatial and temporal dimensions. The term cube might be abusively used for EO data, especially when the temporal dimension exist, since the graphical representation does not match a typical 3-dimenssional cube, but is more accurately represented as a time-series of cubes as shown in *Figure 46*. The main advantage of datacube structures is that they provide interpretable depictions of spatiotemporal data that are easily handled through the execution of various operations over them. For DIONE's EO data case, the data shaping the datacube structure are described from four dimensions:

- Geospatial (latitude)
- Geospatial (longitude)
- Temporal
- Observation layers which are:
  - Sentinel-1 or Sentinel-2 bands
  - Digital Elevation Model layers
  - Landsat bands
  - MODIS layers
  - FLDAS topsoil moisture

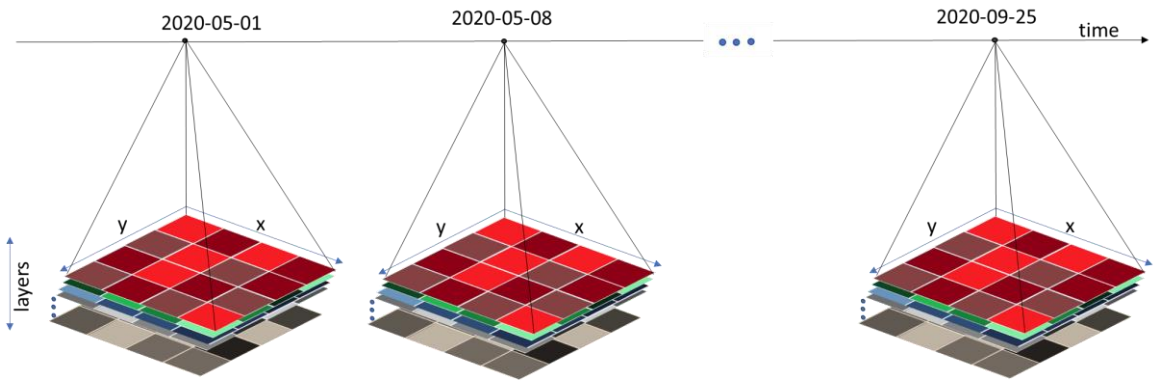


Figure 46: Datacube structure

The retrieval and fusion of the required multi-temporal layers was performed with the Open Data Cube (ODC) that is hosted on a dedicated virtual Ubuntu server at i-BEC premises, under the Apache License, Version 2.0. ODC is an open-source[26] web based UI for handling spaceborne EO data and has been developed from Committee on Earth Observation Satellites[27]. It assists the handling, pre-processing, visualization and combination of ingested data. The UI in combination with the ODC core can utilize various different frameworks such as Python, Javascript and PostgreSQL.

[26] ODC source code and documentation can be found at: https://github.com/opendatacube/documentation
[27] CEOS: The Committee on Earth Observation Satellites is an international organization created in 1984 around the topic of Earth observation satellites

3.3.2 Data filtering and handling

Sentinel 2 data downloads were performed based on the centroids of each polygon calculated using the QGIS software. (QGIS Development Team, year). The downloads are composed of image properties referring to the coordinates of the collection points in Lithuania and Cyprus extracted from Sentinel 1 and 2, Landsat 8, MODIS, FLDAS and DEM (SRTM) The indices were extracted from the images collected between the dates 03/01/2018 and 09/20/2021.

Table 3: Outline of the downloaded EO datasets for each examined point; data were collected from March 2018 to September of 2021

Dataset	Origin	Description
Sentinel-1	Copernicus Open Access Hub	Radar data; VV, VH and angle
Sentinel-2 L2A	Copernicus Open Access Hub	Optical imagery; 10 bands at 10/20/60 m, aerosol optical thickness (AOT) and scene classification (SCL)
Landsat-8	USGS	Optical imagery; 8 bands at 30 m, aerosol attributes (SR_QA_AEROSOL)
MODIS	USGS	Climate variables at 1 km; land surface temperature
FLDAS	NASA Earth Data Search	Climate variables at 12 km; soil moisture, soil temperature, total precipitation
SRTM	USGS	Digital elevation model at 90 m; elevation, aspect, slope

To get the **Sentinel-1 SAR** Ground Range Detected scenes, provided by European Space Agency – ESA/Copernicus, we used the COPERNICUS/S1_GRD image collection. The bands downloaded were VV - Single co-polarization, vertical transmit/vertical receive and VH - Dual-band cross-polarization, vertical transmit/horizontal receive, with 10 meters resolution.

To the **Sentinel 2**, we utilized the Sentinel-2 MSI: MultiSpectral Instrument, Level-2A. The dataset was provided by European Space Agency – ESA/Copernicus through the COPERNICUS/S1_SR image collection. A cloud filter was applied to download images with less than 10% cloud coverage. The selected bands were all 13 bands, available in the visible, near infrared and short-wave infrared plus AOT (Aerosol Optical Thickness), WVP (Water Vapor Pressure) and QA60 (cloud mask). The spatial resolution of the bands B1, B9 and QA60 is 60 meters, for bands B5, B6, B7, B8A, B11 and B12 it is 20 meters and for bands B2, B3, B4, B8, AOT and WVP it is 10 meters.

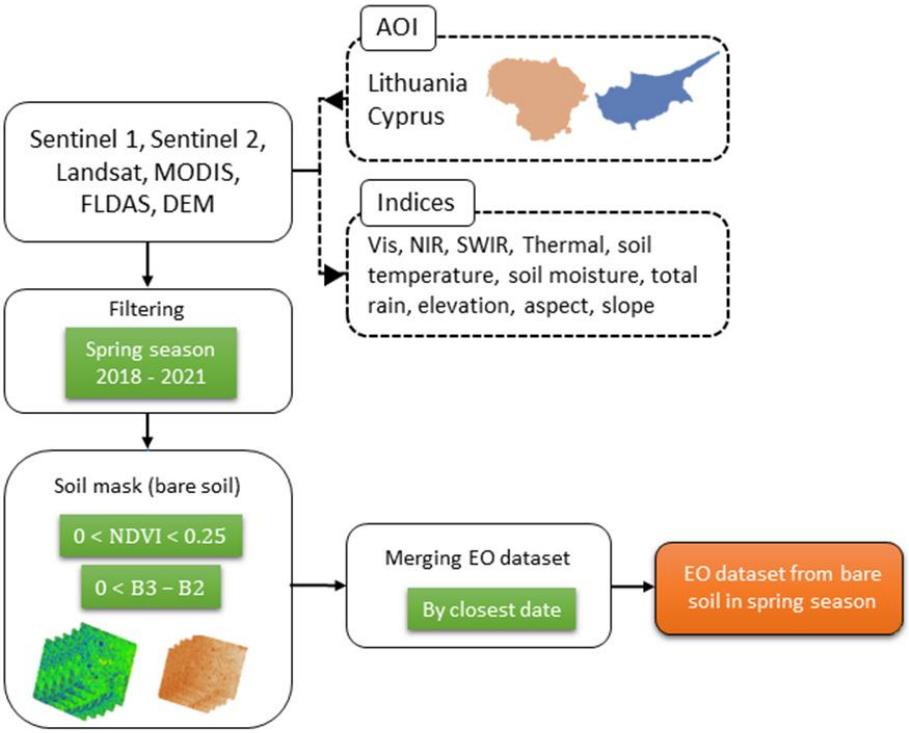


Figure 47: Flowchart of the downloads of Earth Observations (EO) dataset

We used the USGS Landsat 8 Level 2, Collection 2, Tier 1 to download the **Landsat 8** images. The dataset was provided by United States Geological Survey – USGS. The images were downloaded between the years of 2018 and 2021 through the LANDSAT/LC08/C02/T1_L2 image collection. A cloud filter was applied to download images with less than 10% clouds. The selected bands were all 8 bands available in the visible, near infrared, short wave infrared and thermal infrared spectrums plus SR_QA_AEROSOL (Aerosol attributes). All bands have a resolution of 30 meters. Both collected Sentinel and Landsat images, contains atmospherically corrected surface reflectance.

The **MODIS** (Moderate Resolution Imaging Spectroradiometer) image from 2018-2021 was retrieved on 21/09/2021 from https://lpdaac.usgs.gov provided by NASA and USGS Earth Resources Observation and Science (EROS). The images were downloaded between the years of 2018 and September of 2021 through the Image collection MODIS/006/MYD11A1 - Aqua Land Surface Temperature and Emissivity Daily Global 1km - MYD11A1.006. The selected band was the Daytime Land Surface Temperature (LST_Day_1km) with a 1 km spatial resolution.

The **FLDAS**: Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System dataset, was provided by NASA. The images were downloaded between the years of 2018 and September of 2021 through the NASA/FLDAS/NOAH01/C/GL/M/V001 image collection. The selected bands were the Qair_f_tavg (specific humidity), SoilMoi00_10cm_tavg (soil moisture), SoilTemp00_10cm_tavg (soil temperature), Rainf_f_tavg (total precipitation rate) and Tair_f_tavg (near surface wind speed). The spatial resolution is around 11 km and the temporal resolution of 1 month.

To download the Digital Elevation Model (**DEM**) dataset were utilized the SRTM Digital Elevation Data Version 4 with 90 meters of resolution. The dataset was provided by NASA and CGIAR. The image was downloaded through the CGIAR/SRTM90_V4. The relief features collected were elevation, aspect and slope.

All data downloaded and used here is free of charge and freely accessible for all users.

After the download and filter the images in spring range (March to May), we merged all data. To perform this task, we first recognized areas of bare soils and flagged out the areas covered from vegetation by analyzing Sentinel 2 data. Two indices were used to perform this selection, being the vegetation index - NDVI (normalized difference vegetation index) and the difference between bands B3 and B2.

The NDVI is used to define the greenness and healthiness of vegetation (Kumar and Pavan, 2018) following the equation:

$$NDVI = \frac{B8 - B4}{B8 + B4}$$

Where B8 refers to a maximum reflectance due to cellular structure (853.1 nm) and B4 that refers to the maximum absorption by the leaf pigments (664.5 nm) (Rani et al., 2018). The threshold used to select bare soils with NDVI are values NDVI < 0, to exclude water bodies and NDVI > 0.25, to exclude vegetation cover (Guo et al., 2014; Wang et al., 2006).

After the selection of the bare soil areas, the merge with the other data was carried out in the following order: S2> S1> LS> DEM> FLDAS> MODIS through the closest date, following as reference the date of Sentinel 2 dataset. For data analysis and merge, the R software was used, using the packages dplyr and data table (Wickham & François, 2014).

3.3.3 Data overview

After applying the mask to filter the bare soils areas, 2142 instances (unique ID's) were found for Cyprus and 1845 for Lithuania, that is, considering the NDVI values, at Cyprus 94.4% of the areas (1928 parcels) analyzed are considered as bare soil. For Lithuania 75% of the areas have no vegetation (1391), that is, bare soil. Even though the bare soil exposure might be low for the Lithuanian case, which might be related to high cloud coverage, or low decomposition rate of plant residuals, it might have a positive effect in soil erosion mitigation and decrease nutrients runoff.

The data extracted from MODIS was the LST_Day_1km, that is, daytime land surface temperature in kelvin. Data were scaled with the given value of 0.02 and transformed to Celsius degrees. Regarding to the data extracted from FLDAS, the index called SoilTemp00_10cm_tavg were also converted from Kelvin unit to Celsius degree as to Tair_f_tavg (near surface air temperature) index as well.

For other indexes the default units were kept. For FLDAS, the SoilMoi00_10cm_tavg (Soil moisture (0 - 10 cm underground)) index, the soil moisture format is in the volumetric water content m3 m-3. To the Rainf_f_tavg index (total precipitation rate), the unit is in kg m-2 s-1. To the Digital Elevation Model (DEM) indices, unit for elevation is meters, aspect and slope is degree.

The environment variables showed a small difference between the two regions, Lithuania and Cyprus. According the SoilMoi00_10cm_tavg index from FLDAS, that is, the soil moisture from 0 to 10 cm deep,

in Cyprus the mean of the surface soil moisture is of 0.25 m^3m^{-3}, in Lithuania this value is of 0.33 m^3. The soil moisture has minimum value of 0.14 m^3m^{-3} for Cyprus and 0.24 m^3m^{-3} for Lithuania and with maximum value similar for both regions (0.40 m^3m^{-3} for Cyprus and 0.41 m^3m^{-3} for Lithuania).

Regarding the soil temperature, the SoilTemp00_10cm_tavg index from FLDAS show values collected by month, for Cyprus the mean temperature found in soil surface was of 25.6 C and Lithuania of 28.9 C. The minimum and maximum for both countries presented similar values (Table 4), differently from the data reached for LST_Day_1km index from MODIS, with minimum of 18.7 C and maximum of 52 C. This difference between both indices is due the temporal resolution, being the index from MODIS collected daily. To the region of Lithuania, the maximum temperature found in soil surface was of 14.84 C and the minimum of -3.47 C to LST_Day_1km index.

Table 4: Descriptive statistic of climate variables in Cyprus and Lithuania

	Soil Moisture $m^3\ m^{-3}$	Soil Temperature C	Rain total $kg/m^2/s$
Cyprus			
Minimum	0.15	1.5	0.00E+00
Median	0.25	27.1	1.00E-05
Mean	0.25	25.6	1.00E-05
Maximum	0.40	50.8	3.00E-05
Lithuania			
Minimum	0.24	0.7	1.27E-06
Median	0.33	30.9	1.65E-05
Mean	0.33	28.9	1.92E-05
Maximum	0.41	0.9	6.05E-05

Cyprus has the greatest range of elevation, ranging from 2 to 706 meters, with an average of 297 meters. As for Lithuania, the minimum was 38 and the maximum 155 meters, showing a much smaller variation compared to Cyprus.

4 Novel Machine learning tools for the generation of spatial explicit soil indicators

4.1.1 Post usage measurement validation

With the completion of the pilot activities, each device's accuracy was assessed in order to quantify the deviations introduced due to usage under probable extreme conditions. To this end, the ISS were measured by each device and the Mean Absolute Error (MAE) was calculated again between the measured values and the values stored to the correction factor collection of the JSON schema. All used devices passed the post usage benchmarking; thus, no further action was needed to be taken. Furthermore, for each measurement that was collected and posted to the DBMS, the corresponding white reference measurement was tested for validity and accuracy. To this end, each measurement that was referenced by a white measurement that presented an outlying behavior was not included to any further analysis. As outlying behavior was defined the distance from mean that was greater than two standard deviations per nanometer. This led to the discarding of the following measurements that were characterized erroneous probably due to misplacement of the white reference disk during the white reference measurement:

- LT_W_1142
- CY_AG_555

4.1.2 Outlier analysis and filtering

As described at paragraph 4.4, each measurement that was posted to the database was classified as inlier or outlier according to the estimation of the ensemble classifier. A group of measurements was labelled as outliers and were deducted from the analysis pipeline. An example of what the classifier recognized as erroneous can be found at *Figure 48*. The depicted measurements present reflectance more than 100% which is the significant indicator that something went wrong during the procedure. Furthermore, the strong water absorption spectral region around 1900nm cannot be identified, since there is no observable reflectance drop at these examples. Furthermore, the blue curve is featureless with reflectance around 100%, thus there is a strong indication that the targeted object was not soil but the white reference panel, and was transmitted by false operation. The other two signatures may have been produced due to faulty white reference measurement. In any case, the above-described examples along with a set of 10 so classified outliers were removed from the in situ spectra collection.

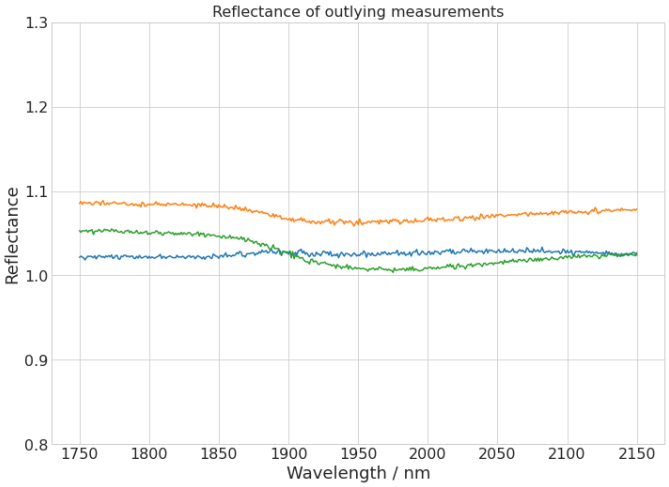


Figure 48: Examples of soil measurements characterized as outliers

4.2 Point predictions

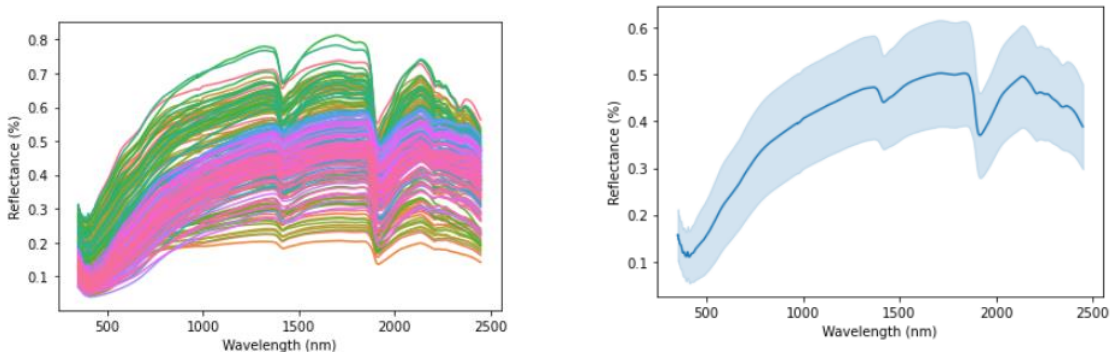
4.2.1 Vis-NIR-SWIR analysis

As described at section 3.1 a soil sampling mission was performed to a subset of the locations that the field surveyors visited, and an amount of approximately 200 gr of topsoil was collected. Each sample was air-dried, crushed, and passed from a 2 mm sieve to obtain fine earth. Then it was separated into to two equal sub-samples, the first was used for diffuse reflectance analysis and archiving, while the second one was used for wet-lab analysis as described in section 4.2.3. For the VNIR analysis the protocol proposed at (Ben Dor et al., 2015) was followed, and the measurements were performed inside an opaque dark box with the use of Spectral Evolution PSR+ 3500 instrument (Spectral Evolution Inc., Lawrence, MA, USA) covering the electromagnetic spectrum from 350 to 2500 nm. According to the followed protocol, all measurements were standardized through the calculation of correction factors of two the two ISS buffer samples, Willey Bay and Lucky Bay. Thus, the induced SSL is a standardized collection of spectral signatures that can be merged with existing libraries that were compiled under the same protocol, such as GEO-Cradle regional SSL.

(a) (b) (c)

Figure 49: (a) Crushing trough mortar and pestle (b) Sieve for soil preparation– (c) Dark box with Spectrometer placed in it

The spectral signatures of the analyzed soil samples, alongside with the average reflectance are shown at *Figure 50*. It can be observed that the reflectance curves extend to a wide range signifying that the sampling locations selected are capable of producing an SSL with high variability covering various soil types.



(a) (b)

Figure 50: (a) Reflectance of Vis-NIR-SWIR laboratory spectra - (b) Average reflectance per nanometer with an interval equal to a standard deviation

4.2.2 Ambient factor removal from in situ spectra

The SSL that will contain the training data for the modeling part of the EO component needs to contain spectra acquired according to a universal scanning protocol. Meeting this requirement is futile since different instruments and protocols were used for the development of each SSL. Furthermore, the most significant factor that affects the homogeneity of the compiled SSL is that spectra acquired from the in situ component present artifacts and effects of non-soil elements such as plant residues, rocks, shadows or moisture. The removal of these factors was performed via the "translation" of in situ soil scans to spectral signatures collected at the laboratory. The rationale behind this transformation lies to the fact that the instrument used for the laboratory spectral analysis is equivalent to the spectrometer that was used for the development of the GEO-Cradle regional SSL and also has the same characteristics with the instrument used for the development of LUCAS 2009 and 2015 SSLs. The signatures were then cropped to the spectral range covered by the portable soil scanners (1750-2150nm). The schematic depiction of this workflow along with the calibration transfer described in 2.3.6 is shown at Figure 51.

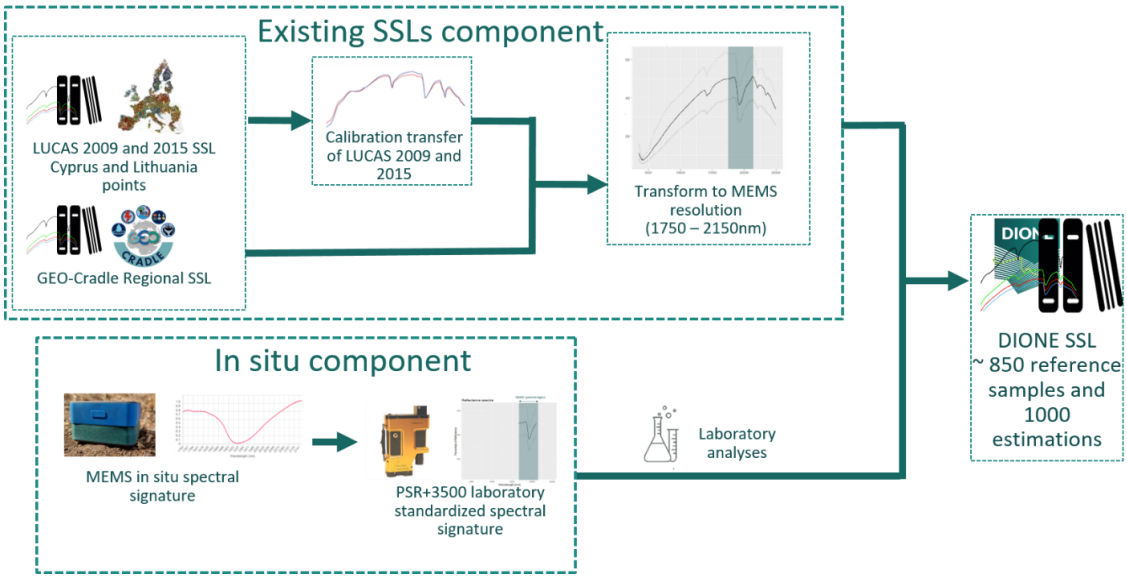
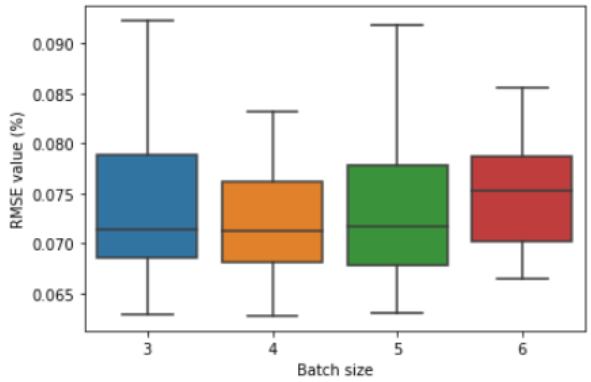


Figure 51: In situ measurements fusion to existing SSL developed under different protocols to a unified SSL

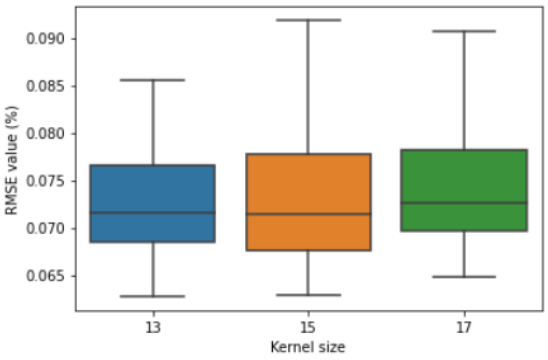
This procedure resulted in analyzing the physical soil samples collection under two different protocols:

- In situ measurements with the MEMS spectrometer
- Laboratory-conditions measurements with full Vis-NIR-SWIR spectrometer

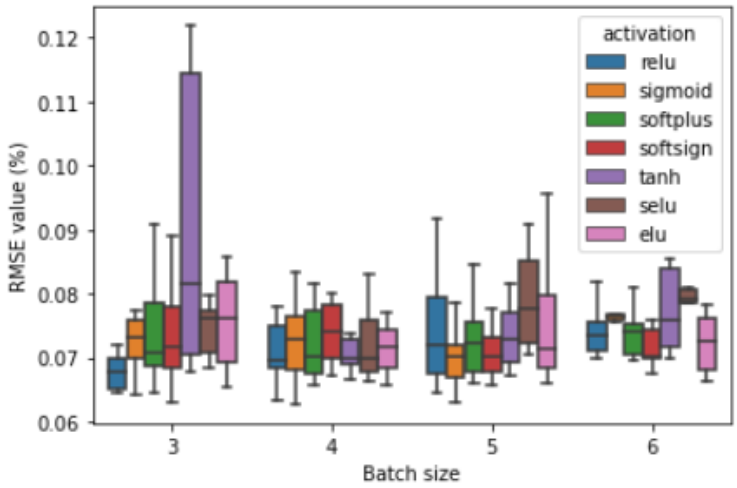
The protocol translation was performed via deep learning, and especially a 1-dimensional Convolutional Neural Network (CNN), which is a common technique in signal transformation, was trained to produce a transformation from MEMS in situ spectra to the corresponding VNIR spectrometer's measurements. The CNN was tuned after a greedy grid search, covering various combinations of activation functions, kernel and batch sizes and different filters. The combination that presented the minimum error in terms of Root Mean Squared Error (RMSE) between projected and reference spectra was then selected, resulting to the development of the desired transformation. An observable slightly increasing trend of the median RMSE for the trials performed for various batch size can be shown at Figure 52 (a) while the selection of the activation function affects the interquartile range of calculated RMSE - Figure 52 (b) – thus different combinations of kernel and batch sizes induce deviations from reference values with high variability. Kernels of varied sizes induce also different RMSE values with the minimum corresponding value to be the one of the greater kernels as shown at Figure 52 (c). A kernel size of 15, batch size of 5 and Scaled Exponential Linear Unit as activation function were selected, resulting to a CNN with induced RMSE value equal to 0.05% calculated over an independent. This transformation applies a correction to the spectra acquired in situ that results to a significant spectral distance drop, which can be reflected as a reduction of Mean Absolute Error from 0.20% to 0.025%.



(a)



(b)



(c)

Figure 52: Boxplots of RMSE values over: (a) different batch sizes, (b) different kernel sizes and (c) different batch sizes and activation functions

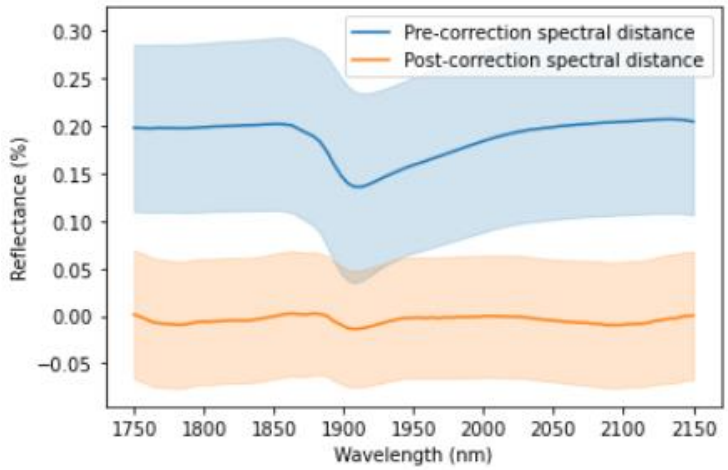


Figure 53: Absolute difference between in situ spectra and their transformation to laboratory reference spectral values

The correction as depicted at Figure 53 has a significant effect in reducing the spectral distance around the strong absorption band of 1900 nm with the weakest effect observed at the bounds of the MEMS spectral range due to the absence of many neighboring bands. Throughout the above translating technique, all in situ acquired spectra can from pilot areas can be regarded as translatable to laboratory ones and in combination with cross-device standardization that was applied, the in situ collection can be merged with the historical SSLs.

4.2.3 Physicochemical analysis

The soil samples originate from regions with different soil characteristics thus a wide range of variation has been covered over the set of monitored parameters, and mainly over the soil texture. At *Figure 54* we can easily distinguish two groups of soils, where soils originating from Cyprus are mainly Clay soils with few of them to be characterized as Clay Loam, Silty Loam and Silty Clay Loam according to United

States Department of Agriculture (USDA) taxonomy, while the Lithuanian soils are mostly characterized as Sandy Clay Loam, Clay Loam and Loam, based on soil texture. The soil texture was determined through the Bouyoucos-hydrometer method (Bouyoucos, 1962).

Electrical conductivity and pH were measured through Saturated Paste extract, with the pH mean value to be 7.84 implying that the collection also includes alkaline soils. This mainly refers to Leukara region of Cyprus, which also presented high concentrations of CaCO3 (measured with calcimeter) and low concentrations of SOC (measured with Walkley-Black method (WALKLEY & BLACK, 1934)). Soils over the Lithuanian pilot areas are characterized as neutral or slightly alkaline since the mean pH value is 7.64. All attributes presented positive skewness (except pH and sand) meaning that higher frequency is observed at lower values than higher ones. Sand and Clay present negative kurtosis (less than 3) which indicates that the distribution is flat with thin tails, meaning that more observations are located near the mean and less of them are located on the tails. The other soil attributes present a positive kurtosis, thus we expect higher frequencies around the mean value.

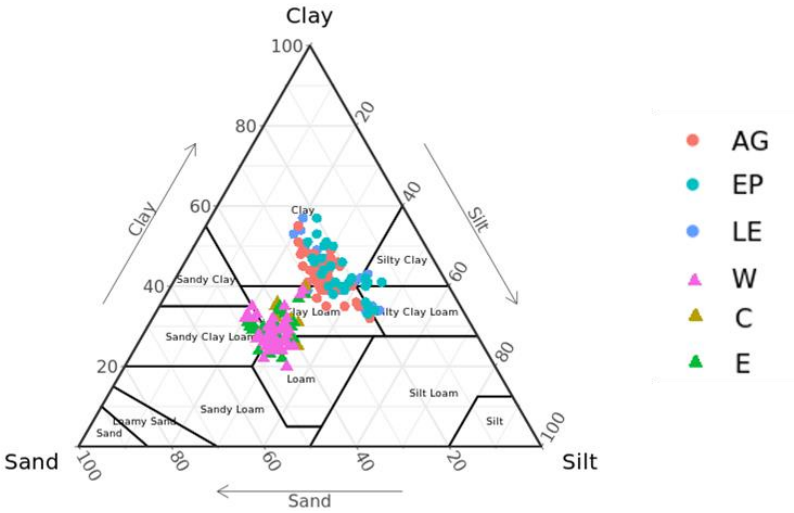


Figure 54: Textural characterization of soil according to USDA

Table 5: Descriptive statistics of soil analyses

Soil analyses									
Soil Attribute	Min	Q25	Median	Mean	Q75	Max	SD	Skew	Kurtosis
Sand (%)	14	25	32	33.55	44	49	10.29	-0.01	1.44
Silt (%)	20	27	29	30.04	33	48	5.93	0.90	3.8
Clay (%)	20	29	35	36.05	42	57	8.18	0.31	2.16
SOC (%)	0.7	1	1.325	1.39	1.67	3.8	0.50	1.39	5.99
pH	7.04	7.668	7.95	7.84	8.06	8.27	0.29	-1.03	3.16
CaCO3 (%)	0.3	1.1	3.65	15.62	31.25	59	17.74	0.76	2.14

EC (mS/cm)	0.52	3.96	4.36	4.39	4.82	6.04	0.65	-0.07	2.23
Cyprus									
Soil Attribute	Min	Q25	Median	Mean	Q75	Max	SD	Skew	Kurtosis
Sand (%)	14	21	25	23.99	27	32	3.68	-0.36	2.86
Silt (%)	20	29	32	33.19	37.25	48	6.60	0.41	2.59
Clay (%)	32	39	42	42.82	46	57	5.25	0.34	2.93
SOC (%)	0.70	0.91	1.10	1.20	1.50	1.79	0.31	0.24	1.63
pH	7.70	7.98	8.06	8.04	8.11	8.27	0.9	-0.55	4.01
CaCO3 (%)	1.10	21	31.50	29.94	39.25	59	14.73	-0.22	2.41
EC (mS/cm)	3.16	3.90	4.18	4.27	4.69	6.04	0.55	0.41	3.17
Lithuania									
Soil Attribute	Min	Q25	Median	Mean	Q75	Max	SD	Skew	Kurtosis
Sand (%)	31	41	44	43.11	46	49	3.85	-0.96	3.96
Silt (%)	20	26	28	27.62	30	35	3.36	-0.35	2.90
Clay (%)	20	26.75	29	29.27	31.25	40	3.79	0.40	3.33
SOC (%)	0.71	1.1	1.48	1.57	1.88	3.80	0.58	1.08	4.39
pH	7.04	7.45	7.66	7.64	7.89	8.19	0.28	-0.36	2.21
CaCO3 (%)	0.30	0.80	1.1	1.29	1.40	8.90	1.12	4.14	2.34
EC (mS/cm)	0.52	4.06	4.65	4.51	4.93	5.98	0.58	-1.40	2.45

4.2.4 Point estimations of soil indicators

The use of indirect, non-destructive methods, such as soil spectroscopy, for analyzing top-soil in support of traditional laboratory analysis facilitates the development and sourcing of dense soil property libraries, enabling the production of detailed maps of spatial depiction of the monitored soil parameters. When compared to standard wet-lab approaches, analyzing the spectral response of soil samples for the determination of soil condition is a widely used method with low time and cost demands. The impacts of one or more soil characteristic parameters on certain spectral regions are reflected in a vector made up of thousands of reflectance measurements. The difficulty in making this transition (from spectral signature to soil property value) is figuring out how to link the shape of the spectral signature to the physical or chemical property. Extensive attempts have been made to build adequate models for linking spectral signatures with the desired attributes, and the estimations are extremely accurate in research regions with great homogeneity in soil structure and mineralogical composition. When this homogeneity declines, or, to be more precise, when attempting to create models on a scale larger than regional (i.e. national, or global), model uncertainty rises. Several

statistical and machine learning techniques have been developed. Several statistical and machine learning approaches, such as Partial Least-Squares Regression (PLSR), Random Forests (RF), Support Vector Machines regression (SVM), or the Cubist algorithm, have been successfully tested for this purpose, achieving high accuracy when it comes to predicting SOC, texture, pH value, Calcium Carbonate concentration, or other soil physical and chemical properties. As of now, the PLSR algorithm remains the most prevalent technique, giving estimates that accurately reflect real-world soil conditions based on a small number of spectral areas as predictors. We evaluated the abovementioned algorithms for the DIONE case and for each region and each soil property, the one producing the best fitting estimations in terms of standard performance metrics was selected. Furthermore, both regional and national scale models have been evaluated for each region in order to investigate whether the range increase of each property will enable the assessed models to better reveal underlying correlations between them and spectral signatures. Added to that, a series of widely employed scatter corrective and spectral derivation pre-treatment techniques were applied for spectral preprocessing. In brief, the reflectance spectra (Ref) were converted into absorbance (Abs) spectra through the transformation

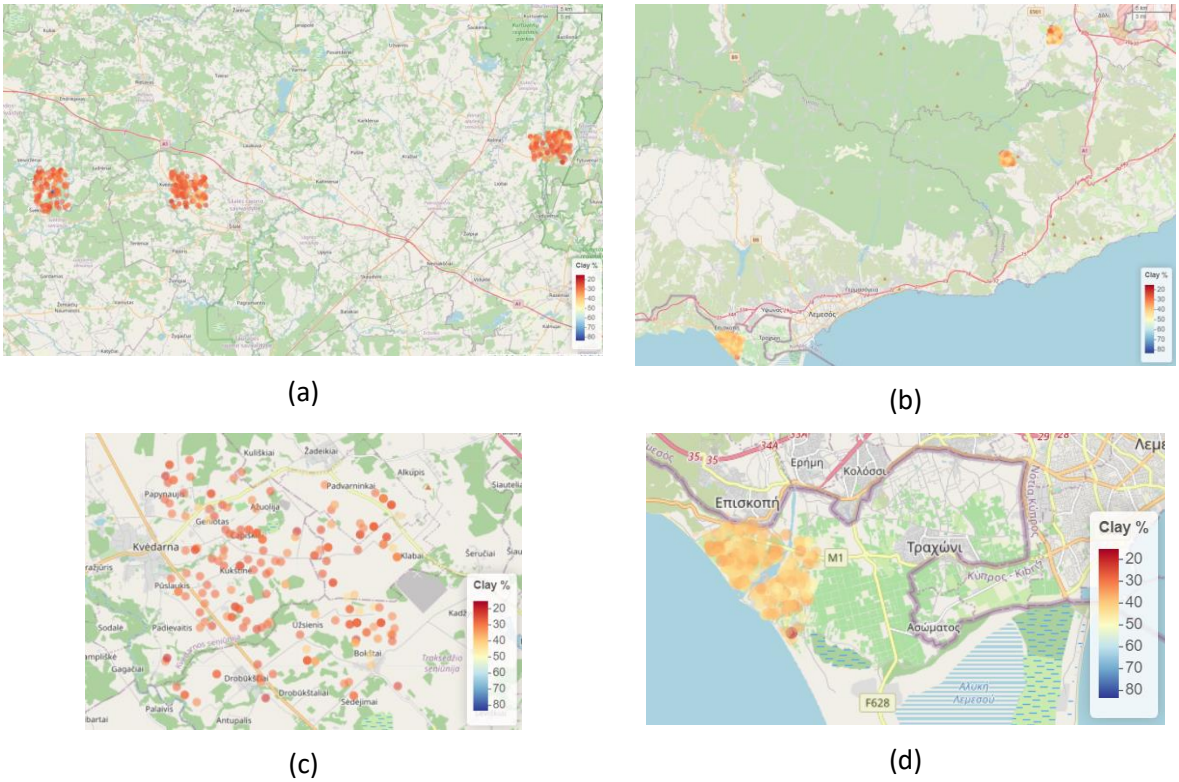
$$\log_{10} 1/R$$

In addition, the Savitzky – Golay (Savitzky & Golay, 1964) filter smoothing combined with standard normal variates (SNV) transformation (Barnes et al., 1989) was used as pretreatment before the spectroscopic modelling to remove unwanted background noise from the spectra (SG). Further pre-processing techniques such as first (SG-1) and second order derivative (SG-2) of the spectra were calculated[28]. The best fitting technique was selected for each iteration. The accuracy results for each pilot case can be found at Appendix C and further visualizations of the properties derived for each pilot area at Appendix D.

Before proceeding to any model fitting, a representative subset equal to 20% of the observations contained at the dataset was selected through the CLHS algorithm and was used as an independent set for assessing the accuracy of every developed model.

Soil texture properties have been estimated through the development of models covering both Lithuanian and Cypriot pilot areas. Especially, Sand content was predicted with RMSE below 11.06% and R^2=0.62 after fitting the Cubist model to the spectral signatures dataset after applying absorbance transformation and SG-1. The result is in accordance with (Cho & Sudduth, 2015) where the reported RMSE is 11.93%. The same approach was followed for the prediction of Clay content which presented higher accuracy (RMSE = 5.89% and R^2=0.77) since Clay minerals are known to have identifiable correlations with SWIR regions. Silt percentage was determined as the proportional residual of the Sand and Clay sum from soil unit. To this end, it presented the highest inaccuracy compared to the other two components with RMSE = 8.98% and R^2 = 0.40. Both Sand, Clay and Silt were representatively split to train and test datasets as shown at *Figure 56* (a), (c) and (d) respectively.

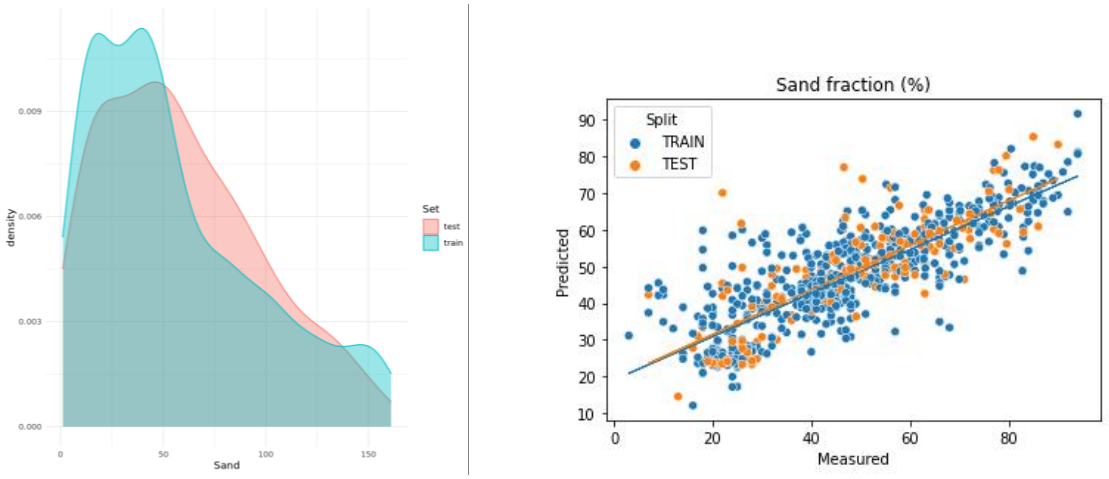
[28] An overview of those techniques is presented by (Rinnan et al., 2009)



(a) (b)

(c) (d)

Figure 55: Clay % estimations for Lithuanian pilot areas (a), Cypriot pilot areas (b), Lithuania Central region (c) and Episkopi region of Cyprus



(a) (b)

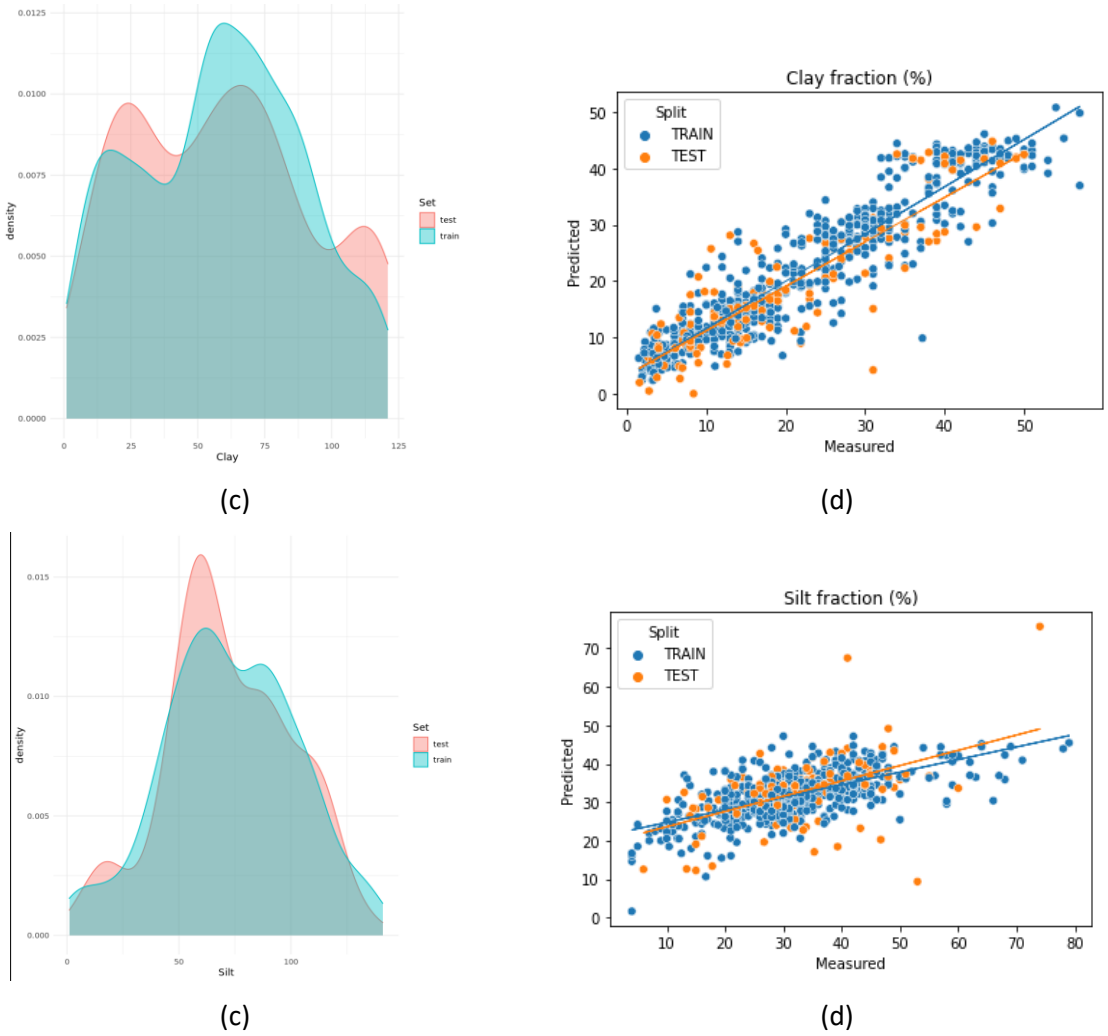


Figure 56: Probability distribution of (a) Sand, (c) Clay and (e) Silt fraction for train and test dataset. Scatterplot of measured and predicted values for (b) Sand, (d) Clay and (f) Silt fraction

For the estimation of SOC content, two national scaled models have been developed that are based on Random Forest regression. SG-1 scattering removal was applied and the two combined models presented joint high accuracy which in terms of RMSE and R^2 is equal to 0.41% and 0.75 respectively. The produced estimations were further upscaled to the extent of each pilot region through the EO component, as described at chapter 5.

As for the pH, the predictability is significantly higher for Lithuanian soils compared to Cypriot's which probably lies to the fact that pH is not directly identified via soil spectroscopy, but the SWIR spectra can be expressed as a linear combination of the decrease of kinetic energy of the acids that correlated to soil's pH value. The acids responsible for pH value of Cyprus might not be triggered by applying electromagnetic energy at the range the Soil Spectrometer operates. To this end, for the estimation of pH two national models were attempted to be developed, one for Cyprus and one for Lithuania respectively. For the Lithuanian case, Cubist model was fitted after the correction of absorbance values with SG-2 and provided an accuracy of RMSE = 0.38 and R^2= 0.76. For the Cypriot case, all

attempts were equivalent to assigning the average value as a prediction, thus no underlying correlations between SWIR spectra and pH were unveiled.

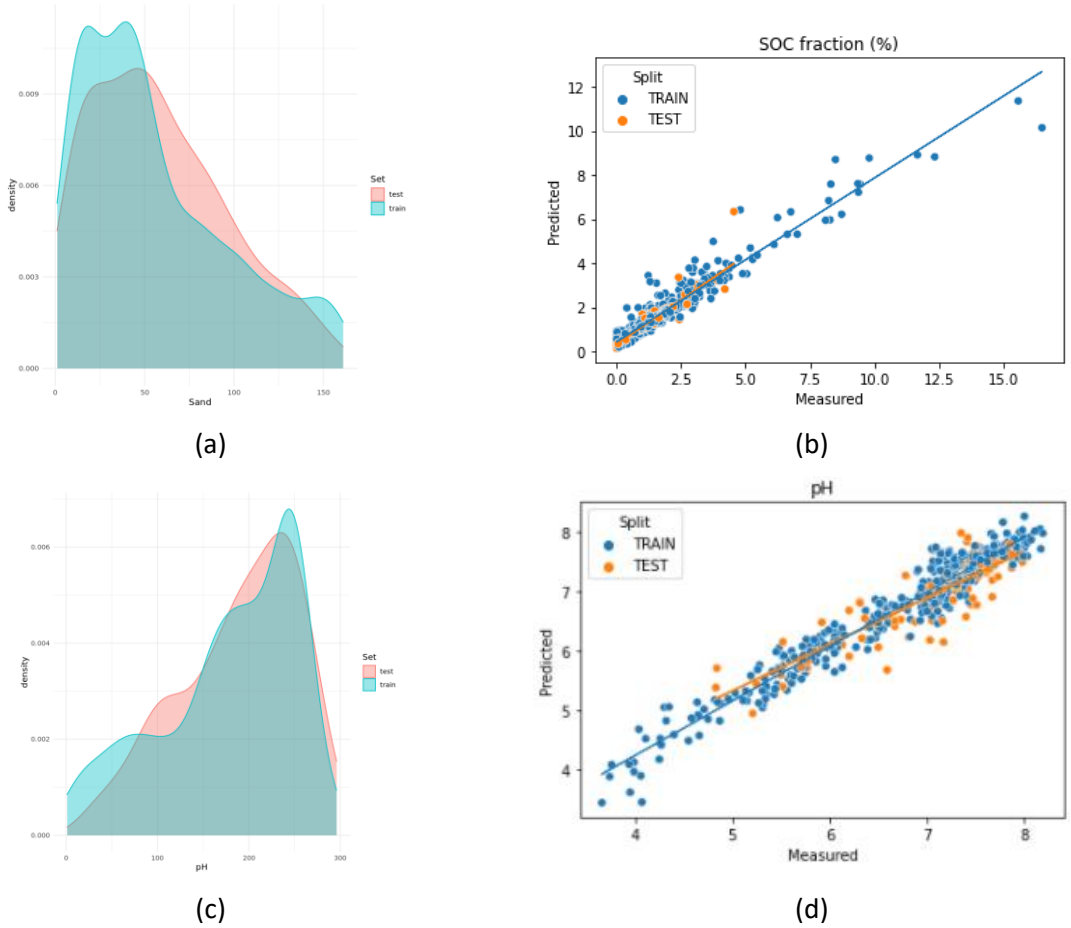


Figure 57: Probability distribution of (a) SOC content and (c) pH values for train and test datasets. Scatterplot of measured and predicted values for (b) SOC content and (d) pH values.

Calcium Carbonate concentration was predicted by fitting Random Forest regression model to the absorbance spectra of both Lithuania and Cyprus, after applying SG-2 filtering with an accuracy of RMSE=10.71% and R^2= 0.75. Electrical Conductivity was estimated through Cubist model with the same preprocessing method as CaCO3 (RMSE=61.95 µS/cm, R^2= 0.80). As previous, the CLHS algorithm provided a representative split to train and test subsets Figure 58 (a) and Figure 58 (c).

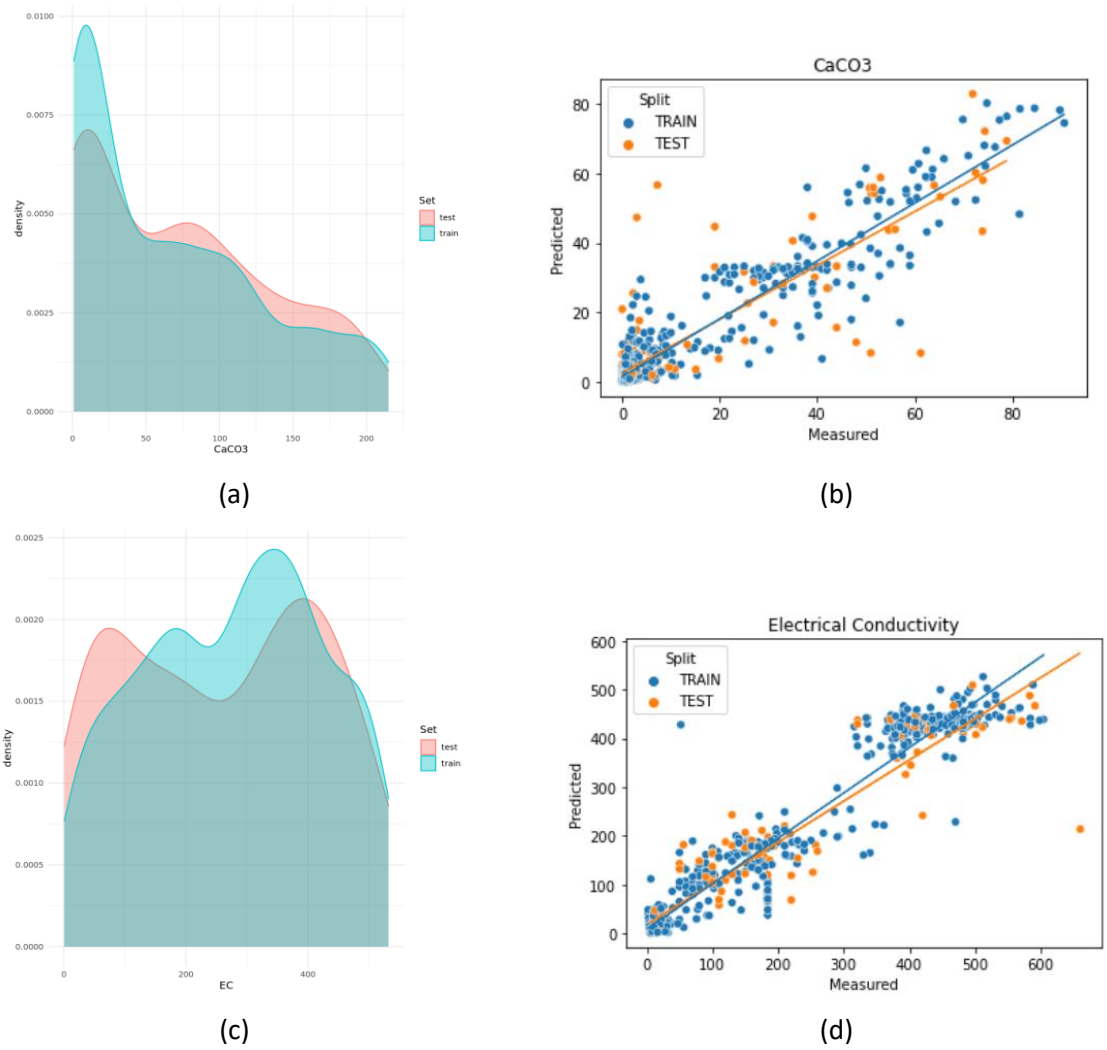


Figure 58: Probability distribution of (a) CaCO3 content and (c) EC values for train and test datasets. Scatterplot of measured and predicted values for (b) CaCO3 content and (d) EC values.

5 Operational provision of in situ component for complementing EO data

The potential of data from satellites is unveiled day after day through the development of EO digital services. To this end, the necessity of design effective integration strategies of data from satellite and airborne with ground-based measurements is highlighted. To address this, a spiked bottom-up approach[29] methodological framework has been developed, leveraging on the data mining techniques with the integration of in-situ and remotely sensed imagery data. This methodology aims to minimize the costs of laboratory measurements by combining existing measurements of open SSLs with few new local soil samples collected from the study area, which in our case are the samples collected from the pilot excursions described at section 3.1. Initially SOC concentration was estimated through calibrating a model over the local soil samples which was then augmented by including the observations of the open SSLs. The comparison between the models developed over the local samples and the augmented set (local samples and open SSLs) is shown at Table 6. The selected model was RF after applying scatter removal through SG-2 to absorbance spectra since it provided the overall best results for the spiked bottom-up approach.

The conclusive step of this work was to derive SOC concentration estimations over the parcels identified as bare soil at 3.3.3 and were not selected for visiting during the pilot activities Figure 60. For this purpose, the collection of EO data as described in the previous section acted as predictors of the SOC concentration target variable, reference of which was the point estimations of the spiked bottom-up approach. The models attained satisfactory results with RMSE=0.75%, R2=0.63 and RPIQ=1.45.

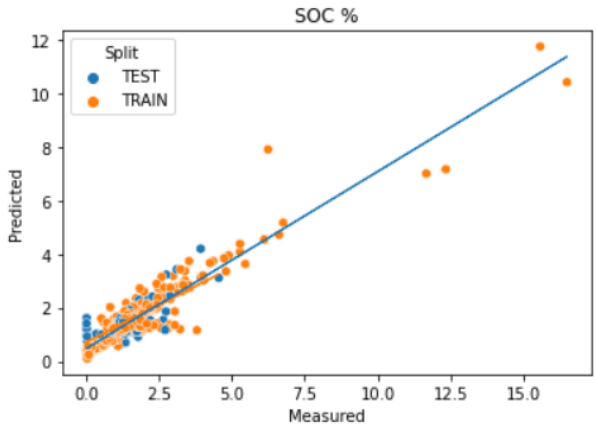


Figure 59: Scatterplot of measured and predicted values for SOC concentration based on estimations of EO data

[29] Extensive description of the spiked bottom-up approach can be found at DIONE public Deliverable D4.2 and at (Tziolas et al., 2020b).

The induced deviations from reference values in EO derived estimations are slightly higher than the ones expected from traditional EO modeling performed over ground truth data since the developed methodology employs as reference set the estimations of SSS spectra. This trade-off between error transfer and the scalability of the studied area is in favour of the second part, with most important factor the option of operating the SSS (or any other spectrometer) in lab. This would eventually diminish this shortcoming by increasing the accuracy of point estimations, by physically removing the ambient factors' effect to spectral response. Furthermore, with the introduction of new missions with better spatial, spectral and temporal resolution, the overall SOC predictability through calibrating models based on EO data is expected to increase.

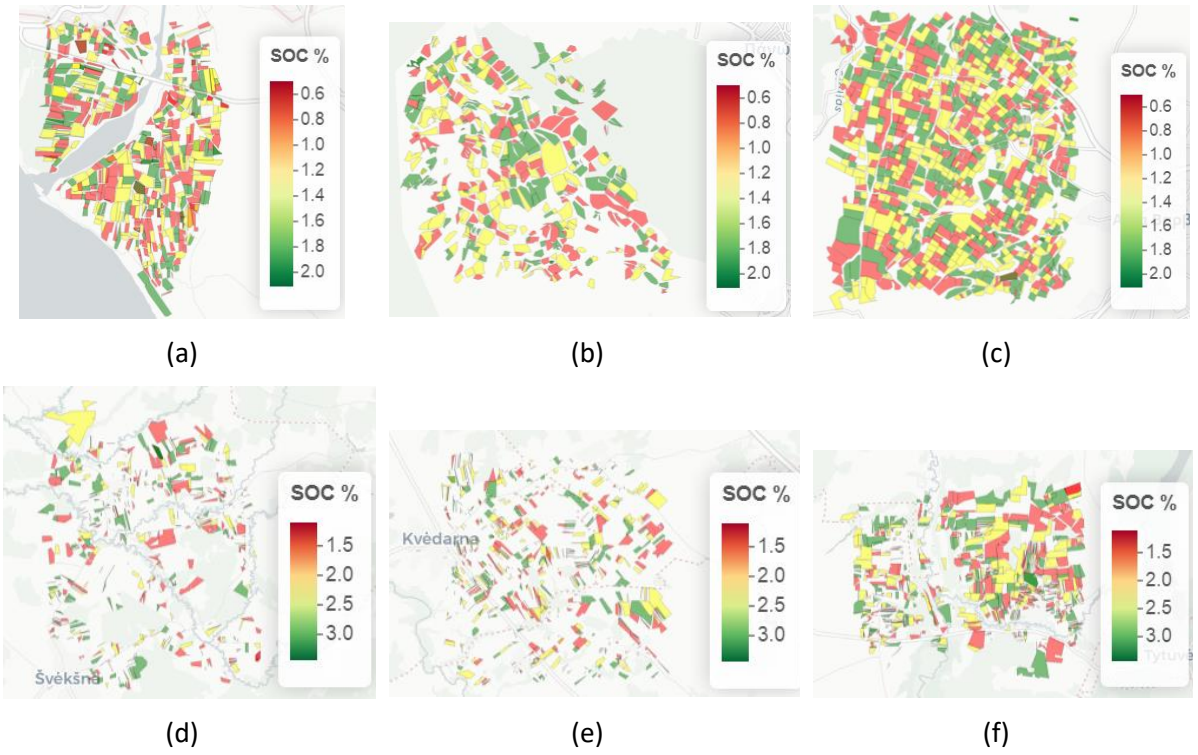


Figure 60: SOC content estimations over declared parcels of pilot areas - (a) Episkopi, Cyprus, (b) Leukara Cyprus, (c) Agia Varvara Cyprus, (d) Western Lithuania, (e) Central Lithuania and (f) Eastern Lithuania

Table 6: Accuracy metrics of the performance of different approaches for the estimation of SOC concentration

Cyprus						
Approach name	Calibration dataset	Model	Preprocessing	RMSE	R^2	RPIQ
Bottom-up	GEO-CRADLE and LUCAS	RF	ABS, SG-2	0.55	0.39	1.41
Spiked bottom-up	Local samples, GEO-CRADLE and LUCAS	RF	ABS, SG-2	0.42	0.68	2.52
Standard	Local samples	RF	ABS, SG-2	0.47	0.61	2.47

Lithuania						
Approach name	Calibration dataset	Model	Preprocessing	RMSE	R^2	RPIQ
Bottom-up	LUCAS	RF	ABS, SG-2	0.54	0.45	1.35
Spiked bottom-up	Local samples and LUCAS	RF	ABS, SG-2	0.50	0.64	2.44
Standard	Local samples	RF	ABS, SG-2	0.52	0.59	2.24
Global						
Approach name	Calibration dataset	Model	Preprocessing	RMSE	R^2	RPIQ
Bottom-up	GEO-CRADLE and LUCAS	RF	ABS, SG-2	0.60	0.36	1.37
Spiked bottom-up	Local samples, GEO-CRADLE and LUCAS	RF	ABS, SG-2	0.55	0.52	1.45
Standard	Local samples	RF	ABS, SG-2	0.67	0.24	1.14

6 Overall achievements

DIONE's main objective is to design and implement to high technological readiness level and demonstrate in two distinct pilots (Lithuania, Cyprus) an integrated green direct payment monitoring toolbox. The toolbox addresses the new regulations inserted by the Modernised CAP 2021-2027 regarding the use of automated technologies for a gradual substitution of on-the-spot-checks and at the same time assists the paying agencies to quantify and tangibly demonstrate the relevant environmental impact of their payments. In this context, there has been developed a set of quantified Key Performance Indices (KPIs) to evaluate the performance of the developed DIONE's components, while there have been concretely identified the system specifications (D2.2: DIONE toolbox specifications and software architecture) that each component needs to meet, based upon the identified user scenarios (D2.1 DIONE stakeholders, personas and co-designed scenarios). The developed KPIs enable end users to assess the quality of the provided EO services and tools within the framework of project[30] while the system specifications include a variety of elements that attempt to define the intended functionality required to satisfy their different users. The consistent set of specifications identified was used to ensure that the developed components respect the intended use and can effectively provide value to the various stakeholders.

As for WP4, a set of ground-truth data has been collected to assess the status of the soil ecosystem with respect to key physicochemical properties. These data were up-scaled using EO means, aiming to contribute to mitigation of soil degradation through promoting obligations to keeping agricultural land in good agricultural and environmental condition. Current level of land degradation can be quantified through DIONE's in situ component, and especially through monitoring related physicochemical properties with the implemented low-cost in situ SSS. The work carried out and the components developed, met the pre-defined users' specifications as shown at the following traceability matrix (Table 7) of system specifications.

Table 7: Traceability matrix of user requirements for developed components

Unique Id	Description	Fit criterion	Verification status	Comments	Priority
SYS_F_SPEC_1	The smartphone application must connect via Bluetooth to the spectrometer and operate it; i.e. perform scanning measurements and record the	The smartphone application must establish a two-way Bluetooth connection with the spectrometer device. The smartphone device must support Bluetooth connectivity,	Connection between spectrometer and Android device is established via Bluetooth. Operation is performed from the Android device where the	Included in alpha version	Must

[30] KPIs were developed in the context of D6.1: DIONE Impact Assessment Methodology and Key Performance Indicators

	associated metadata (i.e. timestamp, GPS) which will be retrieved from the smartphone	have built-in GPS, and internet connection.	measurements are locally stored along with the GPS recording before the transmission to the DBMS		
SYS_F_SPEC_2	The smartphone application must include a component to capture a photo of the soil sample using the camera integrated on the smartphone.	The smartphone application must be able to acquire a photo.	By the time the soil reflectance is captured and automatically the built-in camera is enabled, prompting the user to capture a picture of the scanned soil. The capture is locally stored and transmitted to the database along with the EXIF file included in a JSON format	Included in alpha version	Must
SYS_F_SPEC_3	The smartphone application must be able to transmit the data over-the-air to a database using an Internet connection. In case no internet connection exists, the data should be stored locally and transmitted when a connection is established.	The smartphone app should have a component to transmit the data through e.g. HTTP POST according to the specifications of the RESTful API of the server.	All data and metadata are transmitted via a single JSON file by the time the soil reflectance is captured. In case of unstable internet connection, the user can asynchronously transmit the data by the time internet connection is established through the “sync” button	Included in alpha version	Must
SYS_NF_USE_1	The smartphone application must support recent Android versions (e.g. >= 5.0.0, Lollipop) and be easily installable through the standard package manager i.e. without requiring root permissions.	Provide apk files with support for Android API >= 21	Application can be installed via APK file distributed –	Current release v2.1.0-2021/10/05	Must
SYS_NF_USE_2	The smartphone application should support multiple languages.	All text should be encoded as variables and the translation should be easily implemented through e.g. XML files	Globalization feature enables the user to select as screen language on of English, Greek and Lithuanian	Included in current release	Must

SYS_F_SPEC_4	The smartphone application should be able to visualize the acquired spectrum.	The smartphone application should include a visualization component using e.g. a plotting library.	A graph of the acquired spectrum is shown as a continuous curve representing the reflectance per wavelength when the measurement is completed	Included in alpha version	Must
SYS_F_INTEG_1	UR_C21, UR_C22	The server must validate the incoming data to ensure their integrity using an error correction code.	A specific error correction protocol should be decided upon and implemented client- and server-side, plus the relevant checks should be made by the server as soon as the data are received.	All data are stored as a JSON Schema which specifies a JSON-based format to define the structure of JSON data for validation, documentation, and interaction control. A JSON Schema provides a contract for the JSON data required by a given application, and how that data can be modified.	Must
SYS_F_INTEG_2	UR_C21, UR_C22	The server must (a) statistically analyse the data for outlying behaviour, and (b) flag potential data as (bi) outliers or (bii) novel or (biii) describe their qualitative nature.	The components for outlier and novelty detection should be implemented and transferred to the server. Similarly, the ML model which classifies the images should also be implemented.	(a) a classifier that labels measurements as outliers or inliers has been included. Instances flagged as outliers are not processed (b) a classifier that flags instances as novel has been developed, recognizing measurements sourced from soils classes that have not been measured before. Instances flagged as novels are stored but not processed	Must
SYS_F_INTEG_3	UR_C21, UR_C22	The server must store both the raw (transmitted) as well as the cleaned & validated values to the central DB.	The server should have a method to flag erroneous data.	A single json file is transmitted to the DBMS containing the labels assigned from outlier and novelty detection	Must

SYS_F_INTEG_4	UR_C25	The server must be able to serve the data through a RESTful API.	A RESTful API must exist to extract and insert data.	All data are stored to a MongoDB and via a RESTful API data can be accessed through a RESTHeart interface	Must
SYS_NF_INTEG_1	UR_C21, UR_C25	The server must be always online and handle incoming traffic, including the transmission of geo-tagged photos from the field to the central database, as well as outgoing traffic in a timely manner.	Minimum 99% uptime from the start of the pilot activities till the end of the project. Minimum 200GB of data storage. Minimum 100Mbit downlink and 10Mbit uplink OS version under Long Term Support to ensure security; updates should patch the system live and not require frequent reboots.	Server has been active by the beginning of pilot activities. Furthermore a backup server located at I-BEC premises mirrors all incoming data and can be used when the main server is down (for maintenance, power failure or other reasons) – Both servers' capacity exceeds 1TB - Back up server has been active by the time the pilot activities initiated	Must
SYS_F_INTEG_5	UR_C21	The DBMS must be able to handle heterogeneous input (including geo-tagged photos, spectra, soil maps with their associated metadata).	Use of a NoSQL DB to facilitate the easy integration of data	All data are transmitted as a JSON file and are stored to a document-oriented database (MongoDB) that is able to handle heterogeneous data	Must
SYS_NF_INTEG_2	UR_C21	Data should be transmitted to the database securely over TLS to provide communications security	TLS/HTTPS should be enabled in the API	RESTHeart which connects to the MongoDB instance and exposes the data using a RESTful interface over TLS	Must
SYS_F_INTEG_6	UR_C25	The DBMS must provide data security; access to the data should be restricted, i.e. data must only be edited and deleted by the administrator and new	C operation only by authorized entities R operation by all	Data flow can be sourced only from authorized users. Users can login either through DIONE's SSO or preregistered to	Must

		data must be inserted only by authorized parties. However, data should also be open and easily accessible with a RESTful API.	UD only by admin	the server via their email account. Furthermore, any data flow involves the usage of a physical device (soil scanner) which add an extra security level. Furthermore, communication is performed via secure protocols (TLS/SSL)	
SYS_NF_INTEG_3	UR_C25	The server could support data redundancy and automated back-up processes.	Support of a data redundancy technology and/or of automated backups.	All data are automatically mirrored to I-BEC server as a backup. Automatic back-up images are also created on a weekly basis	Must
SYS_F_SOIL_1	UR_S32, UR_S33, UR_S34	The server must support the deployment and calling of a custom ML-based model which will infer the soil properties (including soil organic carbon and Clay content) from the collected spectral signature of the spectrometers.	Server should be able to run scripts in Python (version 3.5 or newer) and R (version 3.2 or newer) environment	Outlier/novelty detection mechanisms and ML-based point estimation models have been developed in Python and deployed as custom plugins producing real-time estimations and classifications	Must
SYS_F_SOIL_2	UR_S32, UR_S33, UR_S34	The server should be able to effortlessly update the model on-the-fly (e.g. without needing to restart any services or restarting the machine).	Access to the deployed algorithms directory	Updated models can be developed by direct substitution through the file directory	Should
SYS_F_SOIL_3	UR_S35, UR_C25	The server must be able to ingest (i.e. receive as input and store) as well as serve soil maps to external actors. These maps will be generated in an offline manner through a ML process.	The server must store and serve soil maps.	The Developed ODC framework receives as input geospatial layers from sources supporting API	Must
SYS_F_SOIL_4	UR_C24, UR_S36, UR_S37	The ML models that generate the soil maps should be able to handle heterogeneous inputs.	Models will use different inputs.	The developed models are trained over observation with different spatial and temporal resolution, retrieved from	Should

				various sources including point estimations from the in situ component	
SYS_F_GEO_1	Joao is a farmer and wants to check if it is required to provide additional geotagged images for his declared parcels. He opens the application and he is requested to fill in the form with his personal ID in order to log in to the app.	All app users will be linked to their respective farmer unique id.	Achieved - All related authentication and authorization processes are supported; Integration with Toolbox API has been achieved	Included in the alpha version	Must
SYS_F_GEO_2	Joao is a farmer and wants to check if his device is able to benefit from all EGNSS differentiators. He opens the main screen of the application where indicator lights (green, red) display the enabled differentiators.	All geotagged images contain high accuracy positional metadata.	Achieved - EGNSS systems are exploited for enhanced positional accuracy and location integrity; OSNMA implementation available	Included in the final version	Must
SYS_F_GEO_3	Joao is a farmer and wants to check real time data extracted from raw GNSS data in order to see the accuracy of his position.	All geotagged images contain positional metadata.	Achieved – Positional metadata are extracted and made available	Included in the alpha version	Must
SYS_F_GEO_4	Joao is a farmer and wants to identify his position and his related parcels on the map. He taps the map menu option and the application open a map where his position is	All app users must be able to immediately identify their position and related parcels on map.	Achieved – Current position on the map is displayed	Included in the alpha version	Must

	identified with a marker and his related parcels are drawn as a different layer on the map.				
SYS_F_GEO_5	Joao is a farmer and wants to navigate to a specific spot of a reference parcel in order to take requested images. He selects the reference parcel on the map and presses the navigated button which enables a route to the specific spot from his current position.	All users should easily identify on map where they need to go in order to take the requested images.	Achieved – Map based navigation is supported	Included in the alpha version	Should
SYS_F_GEO_6	Joao is a farmer and wants to take an image but he does not know how to do it. So, he opens the manual/help provided from the app to watch the tutorial provided. The tutorial provides him with step by step information of how to capture the image.	All application users must be able to take correct images.	Achieved – In app user tutorial is implemente d	Included in the final version	Must
SYS_F_GEO_7	Joao is a farmer and reaches the location where he is requested to take images. AR content is enabled dictating the exact spot and direction he needs to place his mobile device. After he does this the image capture procedure is	All geotagged images must fit acceptance criteria.	Achieved - Geotagged photos capturing restrictions & guidance is provided though the AR functionality	Included in the alpha version	Must

	enabled requesting him to superimpose parcel borders and have an acceptable device tilt.				
SYS_F_GEO_8	Joao is a farmer has captured the requested images. The send button is enabled. On press all images and their respective metadata are uploaded to the database.	All geotagged images must be saved.	Achieved - Geotagged photos are sent and saved to a database	Included in the alpha version	Must
SYS_F_GEO_9	Joao is a farmer opens the application and navigates to a reference parcel where no internet connection is available and captures the requested images. When an internet connection is available an upload button is activated.	Application should be usable either online or offline.	Achieved - AR mode is operational in offline mode	Included in the alpha version	Must
SYS_F_GEO_10	Joao is a farmer opens the application and wants to change the interface language to Greek. He opens the settings menu and selects from the language dropdown menu GREEK.	All farmers must be able to use and understand the application menus.	Achieved - Application menus are available to both Greek and Lithuanian languages	Included in the final version	Must
SYS_F_GEO_11	Joao is a farmer and wants to see detailed information for his related parcels. He taps the parcel on the	User friendly visualisation of detailed parcel information through the app.	Achieved – Different UI/UX improvemen ts based on	Included in the final version	Nice to have

	map and a pop up with detailed parcel information is activated.		user feedback		
SYS_NF_GEO_1	Mobile devices must be adequately charged. Using raw GNSS data gives application higher positional accuracy but the battery performance is affected.	Mobile device must be active.	Achieved		Must
SYS_NF_GEO_2	GNSS raw data must be enabled while using the application to capture the requested images. Additionally, an amount of time must lapse from the time of the application activation until the capture of images.	All image metadata have accurate positional values.	Achieved – Use of native Android location in place of Unity; dual frequency; EGNOS services	Included in the final version	Must
SYS_F_SEC_1	When a user tries to send data (images with metadata) from the GEO application to the INTEG component to be saved in the database a secure connection must be established. This is achieved by using HTTPS protocol enabling TLS/SSL encryption. Additionally, a certificate pinning must be implemented.	All data transmitted are secure.	Achieved – Integrity component in place	Included in the alpha version	Must

SYS_F_SEC_2	The geotagged photos application is connected to a NTP server to get the date and time to use.	All metadata must be secured.	Achieved – Time integrity algorithm in place	Included in the alpha version	Must
SYS_NF_INTER_1	The geotagged photos application sends images and their corresponding metadata to the INTEG component to be saved in the database. This is accomplished through API services provided by INTEG server.	All accumulated data from GEO application are saved.	Achieved – Communication between data storage system and DIONE Toolbox PI supported	Included in the alpha version	Must
SYS_F_INTEG_7	INTEG receives data from GEO application. It authenticates data by identifying the user, the device and uses spoofing methods to trap fake GNSS.	All data are authenticated before being saved.	Achieved – User authentication is supported; Algorithm to detect fake GNSS Use of OSNMA to enhance the location integrity verification	Included in the alpha version	Must
SYS_F_INTEG_8	INTEG receives images from GEO application. Images go through the image forensics middleware in order to be validated.	All data and metadata are validated before further use.	Achieved – Image forensics middleware implemented	Included in the alpha version	Must
SYS_NF_LEG_1	Images and metadata stored in the DIONE system should comply with relevant EU and national directives regarding personal data.	Proper retention and anonymization practises should take place for the collected photos and metadata during the processing cycle (i.e. removal of car plates and people faces from photos).	Achieved – Anonymization component has been deployed and tested.	Included in the alpha version	Must

References

Ballabio, C., Panagos, P., & Monatanarella, L. (2016). Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma*. https://doi.org/10.1016/j.geoderma.2015.07.006

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, *43*(5), 772–777. https://doi.org/10.1366/0003702894202201

Ben Dor, E., Ong, C., & Lau, I. C. (2015). Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma*. https://doi.org/10.1016/j.geoderma.2015.01.002

Bouyoucos, G. J. (1962). Hydrometer method improved for making particle size analyses of soils. *Agron. J.*, *54*, 464–465.

Castaldi, F., Chabrillat, S., Don, A., & van Wesemael, B. (2019). Soil organic carbon mapping using LUCAS topsoil database and Sentinel-2 data: An approach to reduce soil moisture and crop residue effects. *Remote Sensing*. https://doi.org/10.3390/rs11182121

Cho, Y., & Sudduth, K. A. (2015). Estimation of soil profile physical and chemical properties using a VIS-NIR-EC-force probe. *American Society of Agricultural and Biological Engineers Annual International Meeting 2015*, *4*, 2843–2854. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84951972399&partnerID=40&md5=3c1435eb5b9cfa58101bfaae948f17ea

Demattê, J. A., Huete, A., Jr, L., Nanni, M., Alves, M., & Fiorio, P. (2009). Methodology for Bare Soil Detection and Discrimination by Landsat TM Image. *The Open Remote Sensing Journal*, *2*. https://doi.org/10.2174/1875413900902010024

Guo, L., Wu, S., Zhao, D., Yin, Y., Leng, G., & Zhang, Q. (2014). NDVI-Based Vegetation Change in Inner Mongolia from 1982 to 2006 and Its Relationship to Climate at the Biome Scale. *Advances in Meteorology*, *2014*. https://doi.org/10.1155/2014/692068

Karyotis, K., Angelopoulou, T., Tziolas, N., Palaiologou, E., Samarinas, N., & Zalidis, G. (2021). Evaluation of a Micro-Electro Mechanical Systems Spectral Sensor for Soil Properties Estimation. *Land*, *10*(1). https://doi.org/10.3390/land10010063

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, *32*(9), 1378–1388. https://doi.org/https://doi.org/10.1016/j.cageo.2005.12.009

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthes, B., Dor, E. B., Brown, D. J., Clairotte, M., Csorba, A., Dardenne, P., DemattÃa, J. A., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., & Wetterlind, J. (2015). Chapter Four - Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. *Advances in Agronomy*.

Rani, M., Kumar, P., Pandey, P., Srivastava, P., Chaudhary, B. S., Tomar, V., & Mandal, V. (2018). Multi-Temporal NDVI and Surface Temperature Analysis for Urban Heat Island inbuilt surrounding of Sub-humid Region: A Case Study of two Geographical Regions. *Remote Sensing Applications: Society and Environment*, *10*. https://doi.org/10.1016/j.rsase.2018.03.007

Rinnan, Å., Nørgaard, L., Berg, F., Thygesen, J., Bro, R., & Engelsen, S. (2009). Data Pre-Processing. In *Infrared Spectroscopy for Food Quality Analysis and Control* (pp. 29–50). https://doi.org/10.1016/B978-0-12-374136-3.00002-X

Roderick, G. L. (1962). *A History of Particle-Size Limits* (I. Iowa State University: Ames (ed.); 1st ed.).

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Tóth, G., Jones, A., & Montanarella, L. (2013). The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. *Environmental Monitoring and Assessment*. https://doi.org/10.1007/s10661-013-3109-3

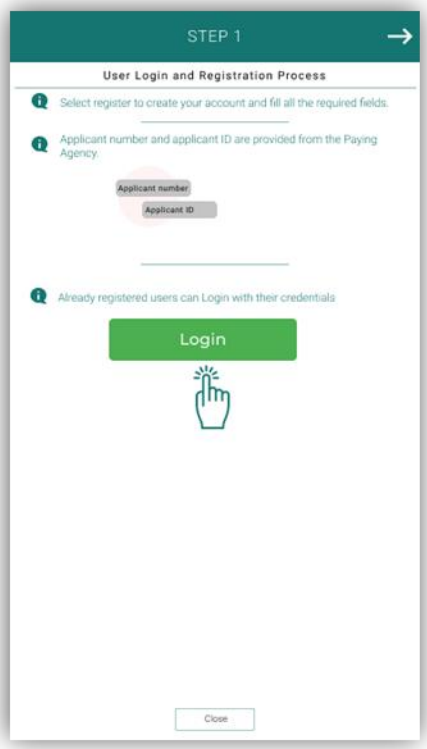
WALKLEY, A., & BLACK, I. A. (1934). AN EXAMINATION OF THE DEGTJAREFF METHOD FOR DETERMINING SOIL ORGANIC MATTER, AND A PROPOSED MODIFICATION OF THE CHROMIC ACID TITRATION METHOD. *Soil Science*, *37*(1). https://journals.lww.com/soilsci/Fulltext/1934/01000/AN_EXAMINATION_OF_THE_DEGTJAREFF_METHOD_FOR.3.aspx

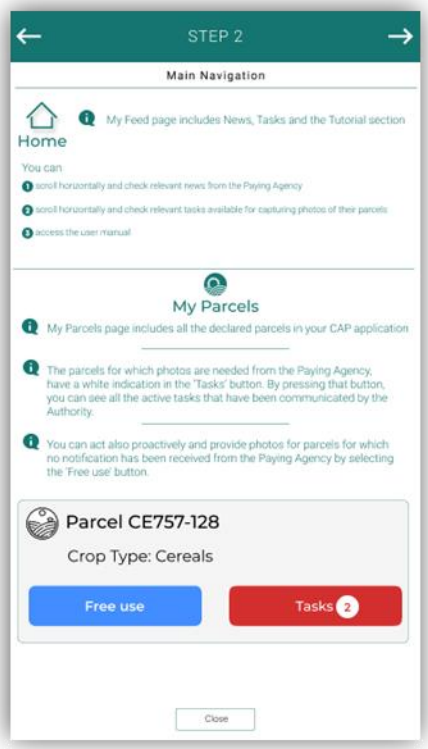
Wang, W., Anderson, B., Phillips, N., & Kaufmann, R. (2006). Feedbacks of Vegetation on Summertime Climate Variability over the North American Grasslands. Part I: Statistical Analysis. *Earth Interactions*, *10*. https://doi.org/10.1175/EI196.1

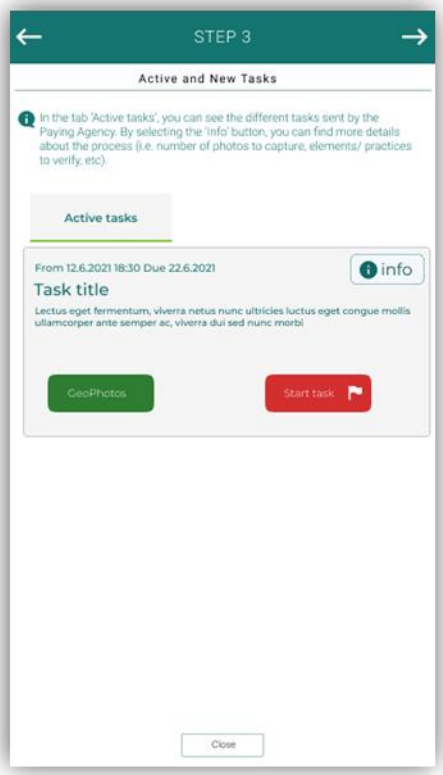
Wickham, H., & François, R. (2014). *dplyr: A Grammar of Data Manipulation*.

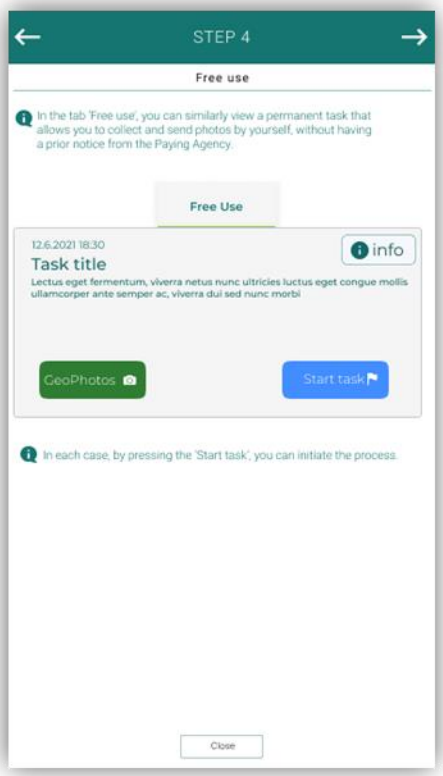
Appendix A

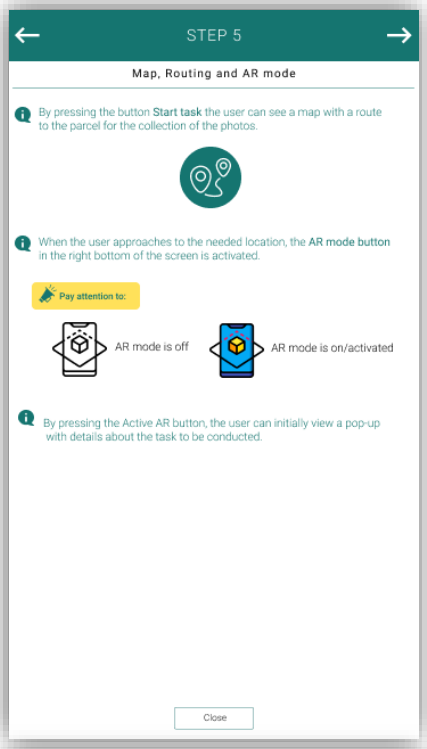
Main steps of tutorial included in the DIONE geotagged photos mobile application.

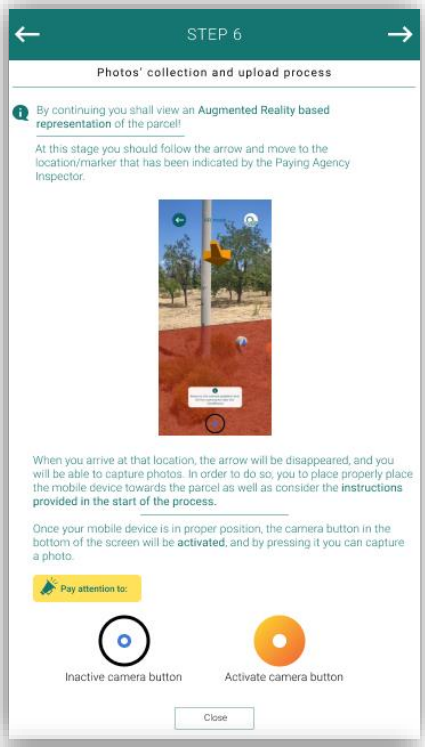


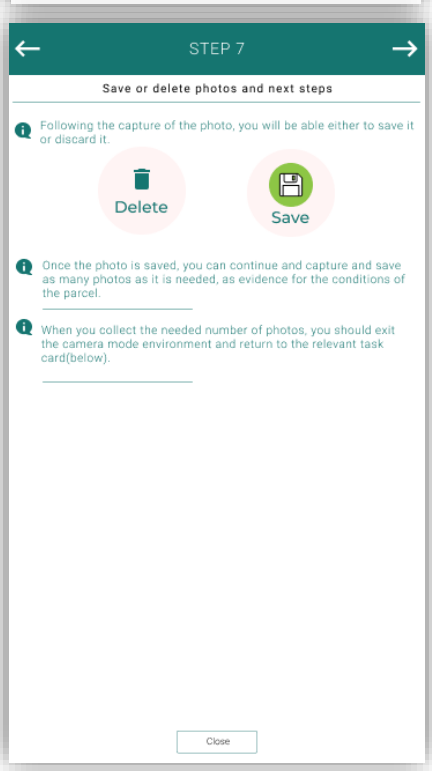


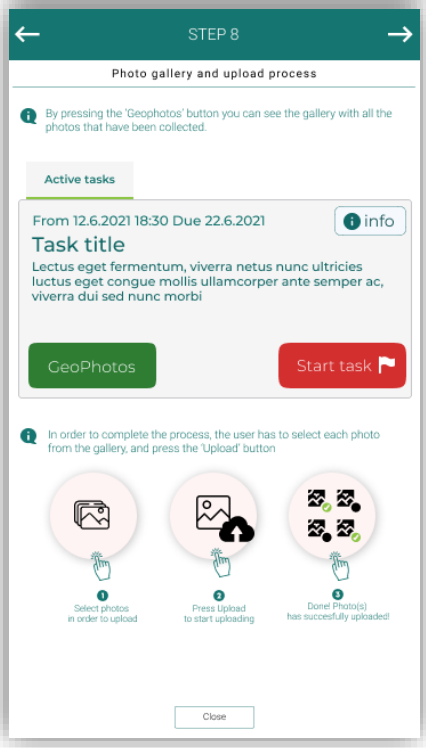


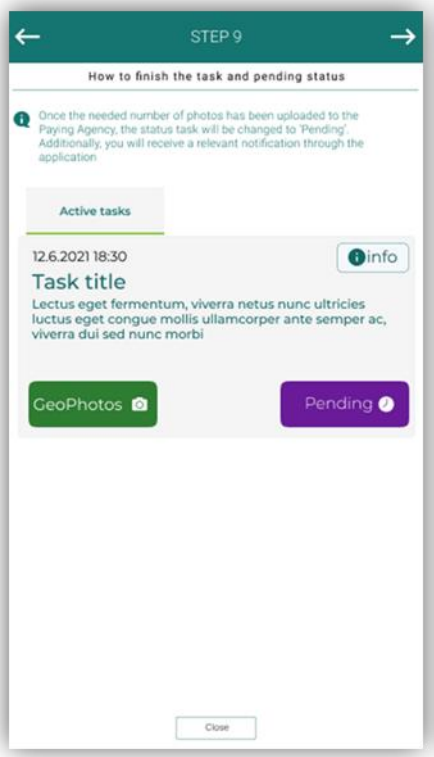



STEP 5
Map, Routing and AR mode
By pressing the button Start task the user can see a map with a route to the parcel for the collection of the photos.
When the user approaches to the needed location, the AR mode button in the right bottom of the screen is activated.
Pay attention to:
AR mode is off
AR mode is on/activated
By pressing the Active AR button, the user can initially view a pop-up with details about the task to be conducted.
Close


STEP 6
Photos' collection and upload process
By continuing you shall view an Augmented Reality based representation of the parcel!
At this stage you should follow the arrow and move to the location/marker that has been indicated by the Paying Agency Inspector.
When you arrive at that location, the arrow will be disappeared, and you will be able to capture photos. In order to do so, you to place properly place the mobile device towards the parcel as well as consider the instructions provided in the start of the process.
Once your mobile device is in proper position, the camera button in the bottom of the screen will be activated, and by pressing it you can capture a photo.
Pay attention to:
Inactive camera button
Activate camera button
Close


STEP 7
Save or delete photos and next steps
Following the capture of the photo, you will be able either to save it or discard it.
Delete
Save
Once the photo is saved, you can continue and capture and save as many photos as it is needed, as evidence for the conditions of the parcel.
When you collect the needed number of photos, you should exit the camera mode environment and return to the relevant task card(below).
Close


STEP 8
Photo gallery and upload process
By pressing the 'Geophotos' button you can see the gallery with all the photos that have been collected.
Active tasks
From 12.6.2021 18:30 Due 22.6.2021
info
Task title
Lectus eget fermentum, viverra netus nunc ultricies luctus eget congue mollis ullamcorper ante semper ac, viverra dui sed nunc morbi
GeoPhotos
Start task
In order to complete the process, the user has to select each photo from the gallery, and press the 'Upload' button
Select photos in order to upload
Press Upload to start uploading
Done! Photo(s) has succesfully uploaded!
Close

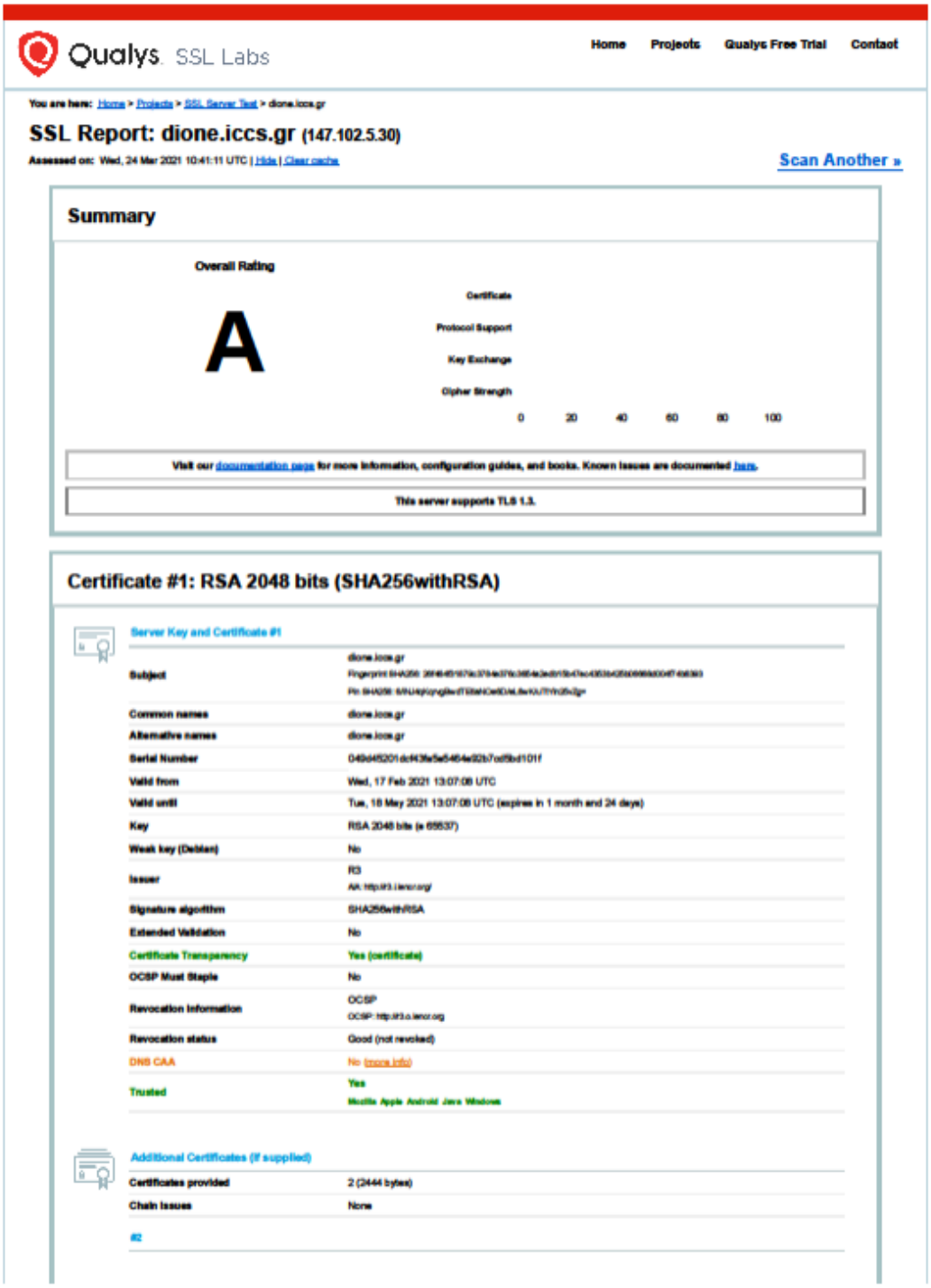

STEP 9
How to finish the task and pending status
Once the needed number of photos has been uploaded to the Paying Agency, the status task will be changed to 'Pending'. Additionally, you will receive a relevant notification through the application
Active tasks
12.6.2021 18:30
info
Task title
Lectus eget fermentum, viverra netus nunc ultricies luctus eget congue mollis ullamcorper ante semper ac, viverra dui sed nunc morbi
GeoPhotos
Pending
Close


STEP 10
Start
Notifications
Notifications received by the Paying Agencies can be viewed also in a dedicated page available in the upper right corner. Once new notifications are available, a dedicated red circle will appear in the icon.
By pressing the 'Show task' button in each notification, you will be directed to the corresponding task, whilst you can also dismiss the notification if the task is already completed.
Close

Appendix B

SSL/TLS Report by SSL Labs for dione.iccs.gr, which demonstrates the support of TLS v1.3.

SSL Server Test: dione.iccs.gr (Powered by Qualys SSL Labs) https://www.ssllabs.com/ssltest/analyze.html?d=dione.iccs.gr



Qualys SSL Labs

Home Projects Qualys Free Trial Contact

You are here: Home > Projects > SSL Server Test > dione.iccs.gr

SSL Report: dione.iccs.gr (147.102.5.30)

Assessed on: Wed, 24 Mar 2021 10:41:11 UTC | Hide | Clear cache

Scan Another »

Summary

Overall Rating

A

Certificate
Protocol Support
Key Exchange
Cipher Strength
0 20 40 60 80 100

Visit our documentation page for more information, configuration guides, and books. Known issues are documented here.

This server supports TLS 1.3.

Certificate #1: RSA 2048 bits (SHA256withRSA)

Server Key and Certificate #1

Subject	dione.iccs.gr
Common names	dione.iccs.gr
Alternative names	dione.iccs.gr
Serial Number	049d45201dcf43fe5e5464e82b7ad5bd101f
Valid from	Wed, 17 Feb 2021 13:07:08 UTC
Valid until	Tue, 18 May 2021 13:07:08 UTC (expires in 1 month and 24 days)
Key	RSA 2048 bits (e 65537)
Weak key (Debian)	No
Issuer	R3
Signature algorithm	SHA256withRSA
Extended Validation	No
Certificate Transparency	Yes (certificate)
OCSP Must Staple	No
Revocation information	OCSP
Revocation status	Good (not revoked)
DNS CAA	No
Trusted	Yes

Additional Certificates (if supplied)

Certificates provided	2 (2444 bytes)
Chain issues	None

#2

1 of 5 3/24/2021, 12:43 PM

Appendix C

Accuracy metrics of point estimations for each property per region, model, extend and pre-processing technique.

Pilot areas of Lithuania

Lithuania Western pilot region – Point estimations accuracy of Sand fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-1, SNV	12.19	0.41	1.66
National	RF	Abs, SG-2, SNV	11.68	0.40	1.54
Global	CUBIST	Abs, SG-2	11.35	0.61	2.33

Lithuania Western pilot region – Point estimations accuracy of Silt fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Abs, SG-1, SNV	8.67	0.38	1.24
National	SVR	Abs, SG-1, SNV	8.32	0.34	1.20
Global	SVR	Abs, SG-2, SNV	8.31	0.29	1.24

Lithuania Western pilot region – Point estimations accuracy of Clay fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-2, SNV	5.72	0.44	1.35
National	CUBIST	ABS, SG-2	5.70	0.61	2.46
Global	CUBIST	ABS, SG-2	6.45	0.69	2.58

Lithuania Western pilot region – Point estimations accuracy of SOC content (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-1, SNV	0.53	0.41	1.23
National	RF	Ref, SG-1	0.54	0.55	1.64
Global	RF	Ref, SNV	0.56	0.52	1.34

Lithuania Western pilot region – Point estimations accuracy of pH					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Ref, SG-1	0.36	0.78	3.07

National	CUBIST	Abs, SG-2	0.41	0.72	2.59
Global	CUBIST	Abs, SG-1, SNV	0.46	0.68	1.99

Lithuania Western pilot region – Point estimations accuracy of CaCO3 (%)

Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SNV	15.93	0.50	0.71
National	CUBIST	Ref, SNV	11.17	0.50	0.50
Global	RF	Abs, SG-2	11.78	0.51	0.65

Lithuania Central pilot region – Point estimations accuracy of Sand fraction (%)

Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Abs, SG-2, SNV	12.43	0.42	1.57
National	CUBIST	Ref, SNV	11.73	0.40	1.53
Global	CUBIST	Abs, SG-2	11.57	0.60	2.29

Lithuania Central pilot region – Point estimations accuracy of Silt fraction (%)

Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	Cubist	Abs, SNV	8.33	0.45	1.38
National	SVR	Abs, SG-1, SNV	8.32	0.34	1.20
Global	SVR	Abs, SG-2, SNV	8.28	0.30	1.24

Lithuania Central pilot region – Point estimations accuracy of Clay fraction (%)

Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-1	5.02	0.55	1.39
National	SVR	Ref, SNV	5.15	0.64	2.55
Global	CUBIST	Abs, SG-2	7.55	0.61	2.52

Lithuania Central pilot region – Point estimations accuracy of SOC content (%)

Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	CUBIST	Abs, SG-1, SNV	0.49	0.63	1.42
National	RF	Ref, SG-1	0.53	0.52	1.05
Global	CUBIST	Ref	0.57	0.49	1.05

Lithuania Central pilot region – Point estimations accuracy of pH

Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ

Regional	SVR	Ref, SNV	0.37	0.77	2.96
National	CUBIST	Abs, SG-2	0.35	0.76	2.95
Global	RF	Ref, SNV	0.46	0.68	2.01

Lithuania Central pilot region – Point estimations accuracy of CaCO3 (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	Cubist	Abs, SNV	12.56	0.66	1.28
National	Cubist	Abs, SNV	11.70	0.75	1.48
Global	Random Forest	Abs, SNV	12.21	0.65	1.45

Lithuania Eastern pilot region – Point estimations accuracy of Sand fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	CUBIST	Abs, SG-2, SNV	11.72	0.44	1.60
National	RF	Abs, SG-2, SNV	11.61	0.41	1.55
Global	CUBIST	Abs, SG-2, SNV	11.35	0.61	2.33

Lithuania Eastern pilot region – Point estimations accuracy of Silt fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	Cubist	Abs, SNV	8.58	0.38	1.19
National	SVR	Abs, SG-1, SNV	8.27	0.35	1.21
Global	RF	Abs, SG-1	8.39	0.30	1.22

Lithuania Eastern pilot region – Point estimations accuracy of Clay fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Abs, SNV	4.65	0.66	1.77
National	RF	Abs, SG-1	5.50	0.64	2.55
Global	CUBIST	ABS, SG-1	6.1	0.68	2.45

Lithuania Eastern pilot region – Point estimations accuracy of SOC content (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	Cubist	Abs, SG-1, SNV	0.43	0.72	1.32
National	RF	Abs, SG-1	0.62	0.66	1.19
Global	SVR	Ref, SNV	0.59	0.54	1.05

Lithuania Eastern pilot region – Point estimations accuracy of pH

Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Ref, SG1	0.35	0.79	3.1
National	Cubist	Abs, SG-2	0.39	0.76	2.79
Global	RF	Abs, SG-2, SNV	0.46	0.70	2.02
Lithuania Eastern pilot region – Point estimations accuracy of CaCO3 (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	Cubist	Abs, SNV	11.75	0.50	1.11
National	Cubist	Abs, SG-1, SNV	11.70	0.45	0.48
Global	Cubist	Abs, SG-2	11.81	0.58	1.25

Pilot areas of Cyprus

Cyprus Leukara region – Point estimations accuracy of Sand fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-2, SNV	12.54	0.61	2.54
National	RF	Abs, SG-2, SNV	11.39	0.71	3.34
Global	CUBIST	Abs, SG-1	11.67	0.70	3.26
Cyprus Leukara pilot region – Point estimations accuracy of Silt fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	PLS	Abs, SG-2, SNV	7.50	0.43	1.74
National	PLS	Abs, SG-2, SNV	7.92	0.37	1.71
Global	SVR	Abs, SG-2, SNV	7.60	0.32	1.51
Cyprus Leukara pilot region – Point estimations accuracy of Clay fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	CUBIST	Ref, SG-1	7.76	0.64	2.43
National	RF	Abs, SG-2, SNV	7.90	0.62	2.41
Global	CUBIST	Abs, SG-2	7.92	0.61	2.37
Cyprus Leukara pilot region – Point estimations accuracy of SOC content (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Ref, SNV	0.63	0.52	1.65

National	RF	Abs, SG-1	0.45	0.65	2.22
Global	RF	Abs, SG-2	0.69	0.40	1.88
Cyprus Leukara pilot region – Point estimations accuracy of pH					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Ref, SNV	0.56	0.17	0.82
National	RF	Abs, SG-1	0.50	0.15	0.71
Global	RF	Abs, SG-1	0.51	0.12	0.69
Cyprus Leukara pilot region – Point estimations accuracy of CaCO3 (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Abs, SG-1, SNV	6.21	0.61	1.98
National	SVR	Abs, SG-1, SNV	6.11	0.59	1.91
Global	RF	Abs, SG-1	6.82	0.65	1.72
Cyprus Agia Varvara pilot region – Point estimations accuracy of Sand fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-2	12.53	0.61	2.35
National	RF	Abs, SG-2	11.54	0.71	3.30
Global	CUBIST	Abs, SG-1, SNV	11.60	0.70	3.28
Cyprus Agia Varvara pilot region – Point estimations accuracy of Silt fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	PLS	Abs, SG-1, SNV	7.50	0.43	1.74
National	SVR	Abs, SG-2, SNV	7.52	0.41	1.63
Global	RF	Abs, SG-2, SNV	7.67	0.32	1.43
Cyprus Agia Varvara pilot region – Point estimations accuracy of Clay fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-2, SNV	7.72	0.62	2.44
National	RF	Abs, SG-1, SNV	7.91	0.60	2.42
Global	CUBIST	Abs, SG-1	7.68	0.64	3.12
Cyprus Agia Varvara pilot region – Point estimations accuracy of SOC content (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ

Regional	RF	Abs, SG-1	0.65	0.55	1.71
National	RF	Abs, SG-1	0.49	0.67	1.81
Global	RF	Abs, SG-1	0.68	0.51	1.65
Cyprus Agia Varvara pilot region – Point estimations accuracy of pH					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Ref, SNV	0.57	0.14	0.80
National	RF	Abs, SG-1, SNV	0.50	0.12	0.71
Global	RF	Abs, SG-1	0.50	0.15	0.70
Cyprus Agia Varvara pilot region – Point estimations accuracy of CaCO3					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Abs, SG-1, SNV	7.17	0.67	2.57
National	SVR	Abs, SG-1	7.78	0.62	2.11
Global	RF	Abs, SG-2	7.85	0.61	2.04
Cyprus Episkopi pilot region – Point estimations accuracy of Sand fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	CUBIST	Abs, SG-1	9.90	0.64	2.86
National	RF	Abs, SG-2	9.80	0.72	2.85
Global	CUBIST	Abs, SG-2	9.85	0.70	3.28
Cyprus Episkopi pilot region – Point estimations accuracy of Silt fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	PLS	Ref, SNV	8.06	0.32	1.59
National	PLS	Ref, SG-1	7.70	0.31	1.49
Global	RF	Abs, SG-2, SNV	7.69	0.31	1.50
Cyprus Episkopi pilot region – Point estimations accuracy of Clay fraction (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-2, SNV	8.47	0.51	2.39
National	CUBIST	Abs, SG-2	7.95	0.68	2.49
Global	CUBIST	Abs, SG-2	8.15	0.61	2.40
Cyprus Episkopi pilot region – Point estimations accuracy of SOC content (%)					

Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-1, SNV	0.59	0.66	2.15
National	RF	Abs, SG-1, SNV	0.52	0.71	2.12
Global	RF	Abs, SG-1	0.65	0.59	1.97
Cyprus Episkopi pilot region – Point estimations accuracy of pH					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	RF	Abs, SG-1	0.53	0.17	0.78
National	RF	Abs, SG-1	0.55	0.15	0.70
Global	RF	Abs, SG-2, SNV	0.50	0.18	0.71
Cyprus Episkopi pilot region – Point estimations accuracy of CaCO3 (%)					
Extend	Selected model	Pre-processing technique	RMSE	R^2	RPIQ
Regional	SVR	Abs, SG-1, SNV	10.17	0.55	1.89
National	SVR	Abs, SG-2, SNV	9.84	0.58	2.02
Global	RF	Abs, SG-2	8.84	0.66	1.97

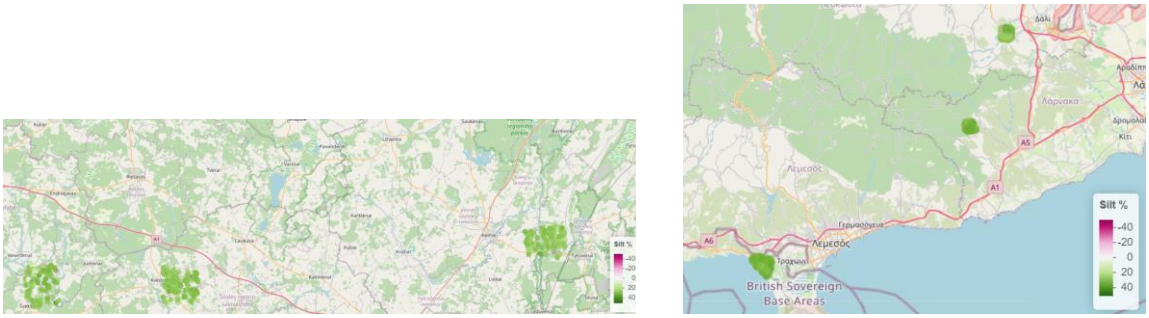
Appendix D

Visualizations of point estimations

(a) (b)

(c) (d)

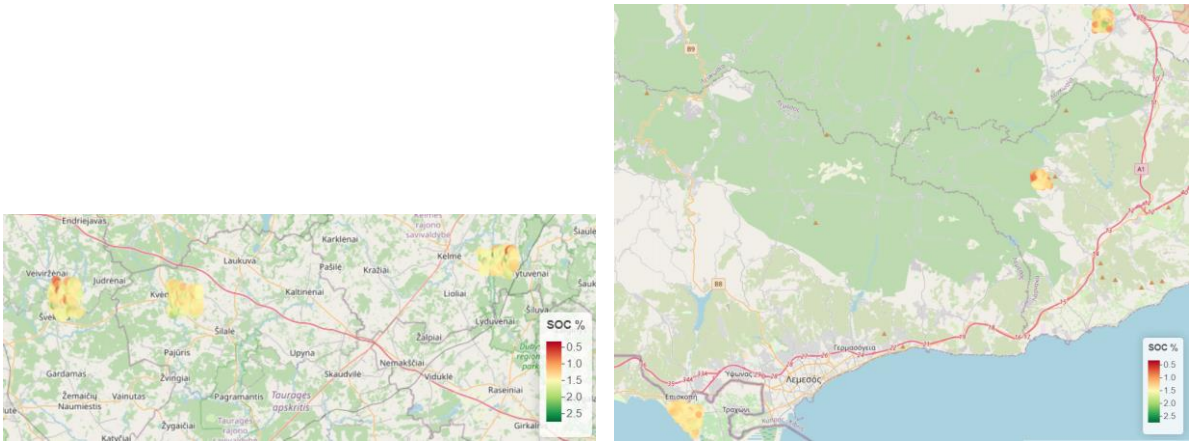
Figure 61: Sand % estimations for Lithuanian pilot areas (a), Cypriot pilot areas (b), Lithuania Eastern region (c) and Agia Varvara region of Cyprus



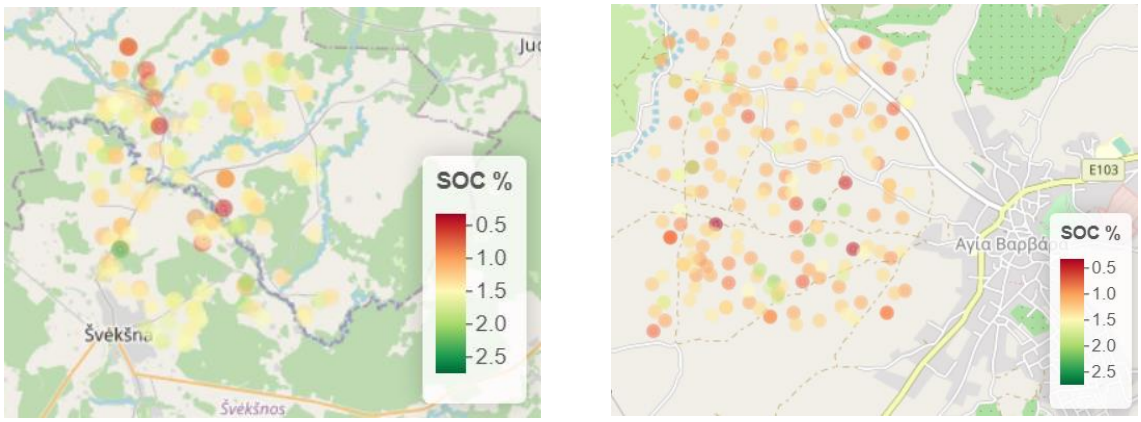
(a) (b)

(c) (d)

Figure 62: Silt % estimations for Lithuanian pilot areas (a), Cypriot pilot areas (b), Lithuania Western region (c) and Leukara region of Cyprus



(a) (b)



(c) (d)

Figure 63: SOC % estimations for Lithuanian pilot areas (a), Cypriot pilot areas (b), Lithuania Western region (c) and Agia Varvara of Cyprus

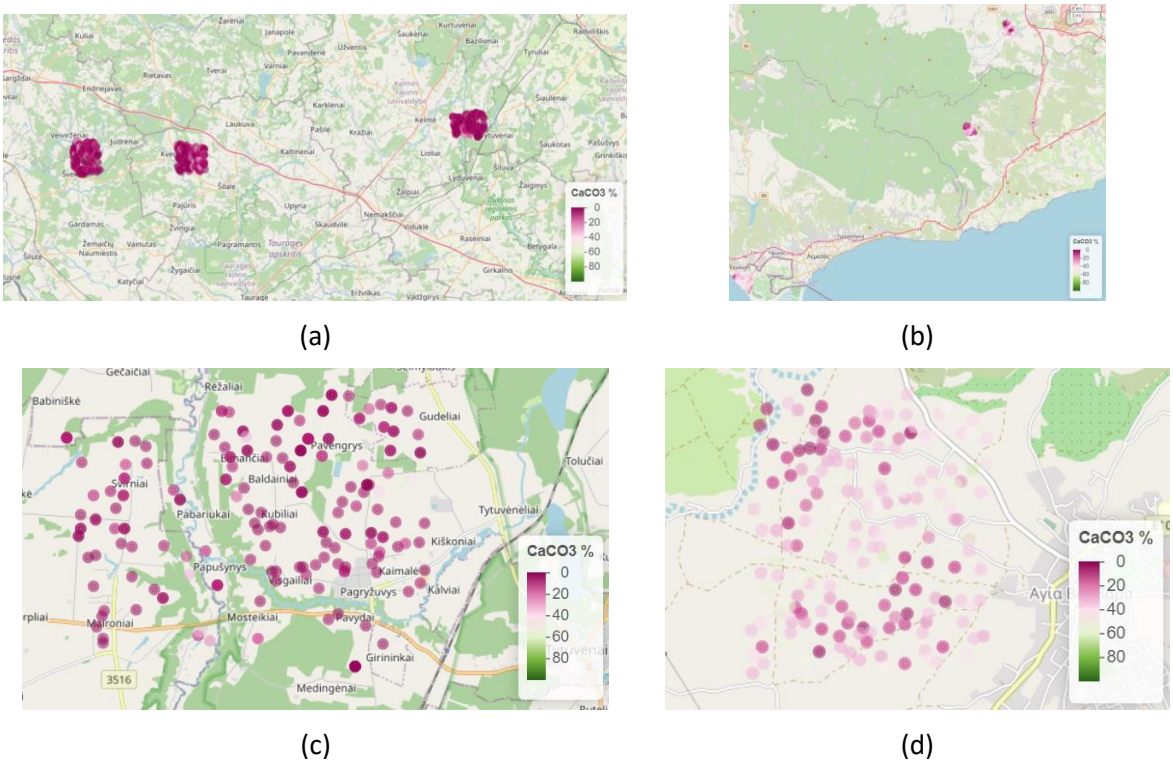


Figure 64: CaCO3 % estimations for Lithuanian pilot areas (a), Cypriot pilot areas (b), Lithuania Eastern region (c) and Leukara of Cyprus

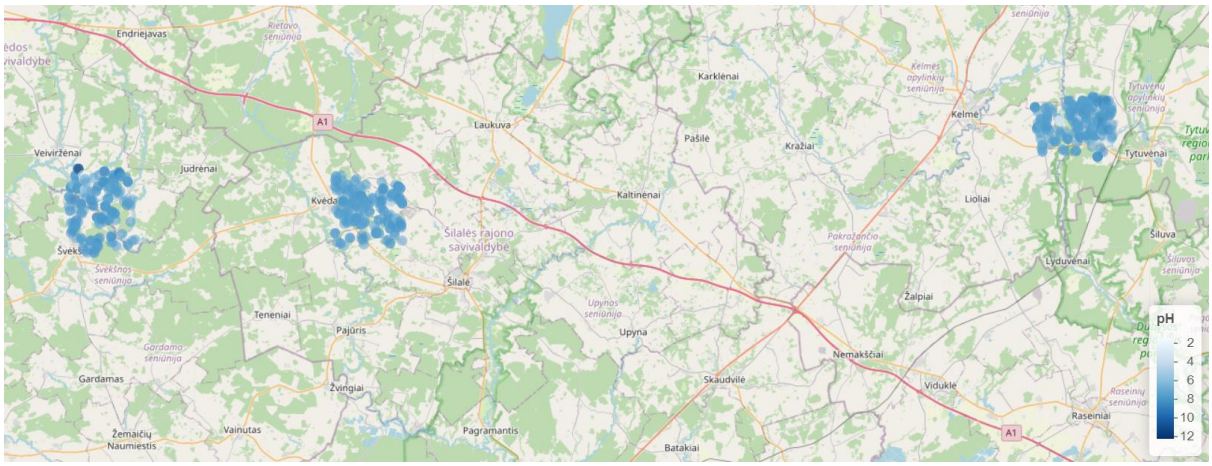


Figure 65: pH % estimations for Lithuanian pilot areas