Incremental Knowledge Extraction From IoT-Based System for Anomaly Detection in Vegetation Crops

Danilo Cavaliere and Sabrina Senatore, *Member, IEEE*

***Abstract*—Precision agriculture systems collect spectral images from satellites, from which vegetation indices (VIs) can be assessed to monitor vegetation and soil condition. It requires a near-daily data acquisition to perform robust crop monitoring and data analysis. Satellites provide a periodic data acquisition that need a further data integration using multiple satellite sources along with camera-equipped drones to achieve an accurate data collection on a selected area. Moreover, VIs are not enough for a proper vegetation evaluation of the monitored areas due to differences among cultivars, the phenological season in which the vegetation is evaluated, the latitude of the areas, etc. This article introduces a system model to detect anomalies regarding the vegetation and soil conditions according to the area phenology and the historical vegetation trends. The system collects spectral images of the regions of interest (ROIs) from satellites and drones, harmonized to calculate VIs and feeds a dataset of near-daily high-resolution integrated images. The harmonic analysis allows phenological data extraction about the ROIs, hence the territorial observation model (TOM) has been extended to represent phenological stages and build knowledge on the ROIs and their phenology that is stored on a triple store. The system selects the VI values, calculated during the learned growing seasons of the ROIs, and classifies them to detect vegetation anomalies affecting those ROIs. The collected knowledge can be used by end-users (e.g., agronomists, experts, etc.) to analyze the anomalies correlated to historical results and vegetation trends.**

***Index Terms*—Harmonic analysis, ontology, phenological context, precision agriculture (PA).**

I. Introduction

NOWADAYS, multidevice systems are widely employed in various contexts to detect and deal with anomalies and critical events. They can provide benefits to IoT-based applications in many fields, including public defense, smart homes, and smart cities, building management, fire fighting, and precision agriculture (PA). Multidevice systems have been employed with success in the agriculture domain, to collect data and support various activities, such as seeding, spraying, etc. PA employs specialized sensors to collect data from the cultivated fields to constantly monitor plant growth and crop quality and increase production and incomes and thus better meet food demand. Historically, PA adopted satellite multispectral images, from which, vegetation indices (VIs) can be assessed. VIs allow for assessment of vegetation vigor, coverage, water presence, and soil conditions in the crop areas monitored. However, satellites, including Modis and Landsat, return data within long periods of time mining a near-daily data acquisition, which is necessary for a proper and constant assessment of the vegetation state. Additionally, satellites may also acquire spectral images at low or different spatial resolutions, even though agriculture demands for images at medium to high resolutions for reliable vegetation state evaluations. To address these issues, solutions focus on processing imagery from multiple satellites to enable an integrated view of near-daily image acquisition at medium to high resolutions, aimed at vegetation assessment and anomaly detection.

Multiple devices of different types (ground vehicles, air vehicles, etc.) equipped with different types of sensors (e.g., infrared sensors, spectral cameras, thermometers, etc.) can effectively benefit environmental monitoring and vegetation assessment. Heterogeneous devices can acquire large amounts of data about the whole environment, providing different features of the environment, from weather to plant characteristics (e.g., temperature, transpiration, jointing structure, etc.). However, the use of heterogeneous devices and sensors may lead to data integration issues, i.e., the data come in different formats and require solutions to be integrated.

Additionally, in PA, data integration is not just a simple format issue, but also a domain issue. The same data, such as relative humidity, air temperature, plant transpiration, etc., can be read and interpreted in different ways depending on their application (e.g., irrigation management, seeding, bacteria infection prevention, etc.). Thus, data contextualization can be necessary to correctly interpret the data and thoroughly support the accomplishment of the application.

Multidevice PA systems must also consider differences in the type of environment; the interpretation of VIs and vegetation parameters for evaluating vegetation health may vary with latitude (i.e., the vegetation of a crop field in Jamaica have different features than a field in England) and context, for instance, expected vegetation in a crop field is different from that in a wood in the same period. Parameter evaluation may also change with the cultivar type, i.e., grapes have different growing seasons than oranges. To address all these differences and thus detect any anomalies in vegetation, the data must be interpreted based on the phenological data.

Manuscript received August 4, 2021; revised October 26, 2021 and December 13, 2021; accepted December 24, 2021. Date of publication December 29, 2021; date of current version January 12, 2022. *(Corresponding author: Danilo Cavaliere.)*

The authors are with the Department of Computer Engineering, Electrical Engineering, and Applied Mathematics (DIEM), University of Salerno, 84084 Fisciano, Italy (e-mail: dcavaliere@unisa.it; ssenatore@unisa.it).

Digital Object Identifier 10.1109/JSTARS.2021.3139155

This article introduces a multidevice system model, which acquires spectral images from different devices such as satellites and unmanned vehicles (UVs). Satellite data are integrated through harmonization techniques. Fourier analysis is applied to the VIs calculated from the satellite images to extract the phenological data of the area to monitor. The data are stored in ontology-based format. In particular, the territorial observation model (TOM) ontology is extended to represent phenological data in the selected areas and then stored as triples. Environment monitoring is performed on each region using if–then rules defined for anomaly detection. When an anomaly is detected, an alert is sent to humans with a report describing the anomaly, the geographical region, where the anomaly occurred and the severity level of the anomaly. This approach aims at demonstrating that low level data integration (data satellite harmonization), combined with incremental knowledge building on the scenario and VIs, can aid in comprehensive vegetation and soil status analysis. Through online status assessments indeed, a near-daily image dataset will be generated along with a posterior analysis, providing human operators with reusable knowledge about vegetation and soil conditions built incrementally with the dataset.

The rest of this article is organized as follows: in Section II, related work is analyzed and discussed, then in Section III, a background on data from precision agriculture field is given. Then, Section IV introduces and details the whole system model step-by-step. Section V presents a simulated real-time implementation of the system model, which is demonstrated through a case study accomplished using the Apache Kafka framework. Finally, Section VII concludes this article.

II. Related Work

PA involves swarms of IoT devices that collect various types of data from the environment to help to monitor vegetation with aim of improving crop growth. To this purpose, multidevice intelligent systems are demanded to keep fields under constant monitoring and check for anomalies.

One major issue coming with the employment of different sensors is the heterogeneity of data they return. The data are of different types (analog and digital) and are represented in different formats. To deal with this issue, researchers provide various solutions to interface and integrate data coming from different sources. In [1], the authors introduce Hydra as an IoT multilevel data fusion system that fuses data at different layers, including raw sensor data, events and decision-making, and decision fusion based on applications, to support water management applications for cultivated fields. In another work [2], the authors proposed an approach that exploits multivariate geostatistical data fusion techniques to fuse multitemporal data from a multiband radiometer and a geophysical sensor, to delineate homogeneous zones to be assigned to differential agricultural management. In [3], the authors developed the phenological vegetation index, that is, synergistically used with the well-known NDVI index to thoroughly extract Spartina Alterniflora from images.

Beyond data integration and analysis, other challenges affect other stages of the PA ecosystem. Among them, there is a reduction of energy consumption in vast IoT sensor networks. In [4], an approach based on the quantum-inspired gravitational search algorithm evaluates the operational mode of each sensor in the network and improves energy consumption accordingly.

To support data integration and contextualization, many works in literature employ semantics and ontologies. A new ontology network is introduced in [5] as composed of two ontologies, one representing concepts on agricultural production and the other on food processing, that are populated with data coming from different sources to support applications in viticulture and winemaking. In [6], the authors explore the knowledge building for smart farming applications, defining a service based on an ontology, modeling various domains (e.g., production resources, agricultural machinery, etc.), that can be easily accessed and used by farmers and experts. Another system based on semantics for agriculture has been defined as part of the CANDELA Horizon2020 European project aimed at bridging the gap between big data and earth observation data users [7]. CANDELA introduces a platform to access Earth Observation data, provided by the Copernicus satellite Sentinel-2, that are processed with data mining techniques for area classification and change detection purposes. Deep learning is also applied for change detection and data fusion is applied to merge data. The platform has also a semantics-based module composed of two submodules: a semantic indexing submodule devoted to knowledge building from a user-selected dataset that is then stored on Starbon triple store, and a semantic search submodule for accessing the semantic data through queries. Notwithstanding the benefits of such a platform, let us notice that CANDELA processes data coming only from Sentinel satellites, ignoring data coming from other satellites (e.g., MODIS, Landsat). Furthermore, in the ontologies, employed in this project, concepts explaining phenological data and anomalies are missing. Our approach, instead, extends the TOM ontology with new classes and properties to allow building a detailed knowledge on phenology and features of the anomalies that can serve the generation of more explanatory warnings for humans.

For what concerns anomaly detection, a reference point is the anomaly hot spots of agricultural production (ASAP) system introduced in [8]. ASAP allows detecting anomalies in crop production in two steps. First, the subnational administrative areas are classified into four categories based on vegetation and rainfall anomalies, and updated every 10 days. Then, as the second step, all the information acquired, including the warnings, are analyzed by experts who support the labeling of countries in potentially critical conditions. Despite the interesting warning generation mechanism introduced, ASAP mainly processes satellite data ignoring the potential of using UVs, and they do not use any semantics-based model to support management and dissemination of data, such as those ontology-based mentioned before. Additionally, ASAP employs just a VI, namely NDVI; and assesses time series on NDVI from medium–low resolution satellite images (10-day SPOT-VEGETATION NDVI time series at 1 km spatial resolution). In our approach, instead, a dense dataset of high-resolution images is built to robustly

assess VIs and phenological data that are encoded as ontological assertions to support the end-users with the anomaly detection and analysis.

III. PA Data and Devices

The main data processed by PA systems are spectral images collected by satellites and UVs equipped with spectral cameras, which allow the evaluation of the vegetation status by comparing images of the areas at different spectral bands. This section goes into detail about the role of these devices in PA applications.

A. Satellites

In this study, three types of satellites were considered: Sentinel 2, Landsat 7, and Landsat 8.

Copernicus Sentinel-2 mission includes two polar-orbiting satellites placed in the same sun-synchronous orbit, phased at 180° to each other. This constellation of satellites allows monitoring changes in Earth's surface due to the wide swath width (290 km) and high revisit time. The satellites are equipped with a multispectral optoelectronic sensor for surveying with a sentinel-2 resolution of 10–60 m in the visible, near-infrared (NIR), and short-wave infrared (SWIR) spectral zones, including 13 spectral channels, which ensures the capture of differences in vegetation state, including temporal changes, and also minimizes impact on the quality of atmospheric photography. These two satellites acquire data every 5 days at the equator and every 2–3 days at middle latitudes. The multispectral sensor generates 12 images achieved by capturing the light spectrum into 12 distinct bands, including coastal aerosol, blue, green, red, three different ranges of vegetation red edges, NIR, Narrow NIR, water vapor, and three types of SWIR. The revisit frequency of each Sentinel-2 satellite is 10 days, and the revisit frequency of the combined constellation is about 5 days.

The Landsat 7 satellite is equipped with an observing tool instrument called enhanced thematic mapper plus (ETM+). The ETM+ allows global change studies, land cover monitoring and assessment, and large area mapping. The main features of this satellite are a panchromatic band with 15 m spatial resolution, on-board, full aperture, 5% absolute radiometric calibration, a thermal IR channel with 60 m spatial resolution, and an on-board data recorder. This satellite provides images on 8 bands, including three different bands in the visible spectrum, a NIR, a SWIR, a thermal infrared, a mid-infrared, and a panchromatic band.

Landsat 8 (formerly the Landsat data continuity mission, LDCM) has been launched in orbit more recently than the other Landsat satellites. It is currently used to collect valuable data and imagery for assisting applications in agriculture, education, business, science, and government. The multispectral sensors of this satellite allow the acquisition of high-resolution multispectral images of the Earth's surface on a global basis. The data acquired by Landsat spacecraft constitute the longest record of the Earth's continental surfaces as seen from space. It is a record unmatched in quality, detail, coverage, and value. Landsat 8 acquires images on 11 bands, including deep blue and violet, blue, green, red, NIR, two different slices of SWIR, panchromatic, a band with a thin slice of wavelengths (1370 ± 10) and two slices of thermal infrared (TIR). Landsat satellites have a revisit frequency of around 16 days.

The main advantages of Sentinel-2 are the high number of different bands and the high frequency of data acquisition. Landsat satellites return data with a lower frequency and have lesser bands, but they have the thermal NIR band, which is missing in Sentinel satellites, useful for evaluating land surface temperature, and a bigger historical image archive (i.e., Landsat 7 was launched in 1999, Landsat 8 in 2013, Sentinel-2 in 2015).

B. UVs

UVs have been successfully employed to address various tasks, including public surveillance, breeding, filmmaking, and more. In PA, they have been used for various sensing and actuation tasks, including plant temperature acquisition, irrigation, pesticide spraying, etc. These devices have also been used in combination with other IoT sensors in a dedicated environmental monitoring swarm for PA and forestry management applications. Being mobile, such devices can acquire measurements from the area of interest and communicate with other fixed sensors, automated ground controllers, and humans to perform articulated tasks.

UVs can be equipped with hyperspectral imaging cameras and used alongside or as substitutes for satellites. They can certainly better avoid cloud covering by acquiring data from lower heights than satellites, however, they can sometimes suffer from calibration and stability issues that can compromise the quality of the collected measurements (e.g., blurred images). Additionally, satellite services may be cheaper solutions as UVs need to be equipped with expensive spectral image cameras. UVs have been extensively investigated to collect and process the information over large areas [9]; therefore, they might be used alongside satellites to collect more images and satisfy the need of acquiring a near-daily dataset. Although UVs can be leveraged as a further source of spectral images, due to imagery collection costs, the method and experimentation presented in this article have been focused exclusively on satellite-taken spectral images.

IV. Framework for Anomaly Detection in Vegetation

The system model and the interactions among the main modules are sketched in Fig. 1.

The input data are the spectral images, that can be acquired from satellites and camera-equipped drones. Let us notice that in this study, only satellites (Landsat 7, 8, and Sentinel-2) have been used, due to a matter of hardware availability. The collected satellite images are processed by the *data preparation and integration* module, that merges them through data harmonization techniques, to generate fused images of the environment. Specifically, the harmonic method generates phenological data on each region of interest (ROI), under monitoring. The *VI calculation* module reads the harmonized dataset and calculates two VIs: standardized vegetation index (SVI) and normalized burn ratio (NBR). Meantime, the phenological data are sent to

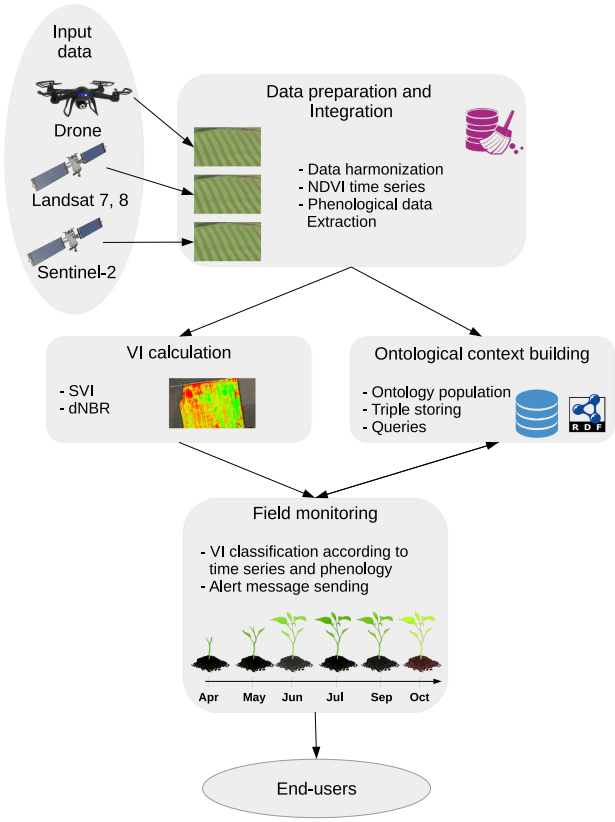

Fig. 1. System overview.

the *ontological context building* module, that builds a knowledge base (KB) by using integrated state-of-the-art ontologies, and then storing the KB on an Apache Fuseki triple store. This module also provides queries to extract knowledge from the KB and support environment monitoring. Finally, the *Area monitoring* module acquires the VIs and retrieves phenological data for each processed ROI by running queries on the KB provided by the *ontological context building*. This way, the VIs on the ROI are contextualized according to the type of environment and seasonal growth stages to finally achieve the anomaly description through rules. According to the results, the system alerts human operators on potential detected vegetation anomalies and the possible worsening of vegetation over time. Then, the architecture includes satellites and camera-equipped drones to collect, stream, and process spectral images about fields. Apart from dynamic collection and sensing capabilities, the IoT-based infrastructure also enables the involved devices to share and fuse the information they collect individually in a KB. Thanks to the inference, the system knowledge gathers sensing and environmental data that are decisive to support humans in decisions, thanks to accurate evaluations on vegetation and soil statuses. The remainder of the section details the modules of the architecture.

A. Data Preparation and Integration

PA requires near-daily imagery at medium to high spatial resolution (10–30 m) to clearly detect anomalies in vegetation.

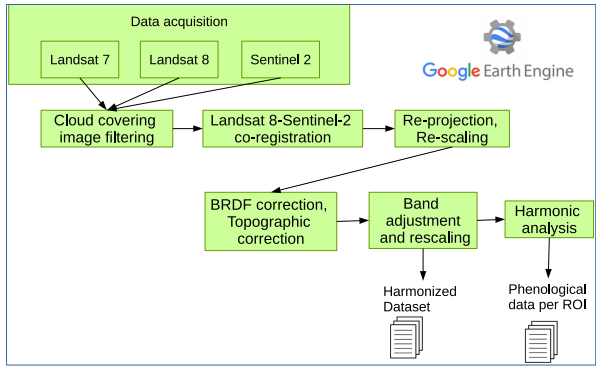

Fig. 2. Data preparation and integration module.

Satellite types such as Sentinel-2 and Landsat fail to support these requirements fully (see Section III). First of all, the two satellite types have different spatial resolutions, in fact, Landsat satellites have medium-high spatial resolution (30 m), while Sentinel-2 acquires data with a spatial resolution that changes according to the specific band from 10 to 60 m.

The two types of satellites also return data with different frequencies over time: Sentinel-2 returns spectral images of the Earth's surface within 5–10 days, while Landsat satellites have a longer review time (16 days). This long interval between two successive data can lead to the loss of important observations relating to the earth's surface, especially in critical growth phases. Solutions are needed to overcome this tradeoff between spatial resolution and temporal acquisition between these satellites. To address these issues, data fusion techniques enable image integration coming from these different satellites, by converging toward near-daily high-resolution integrated images for better vegetation evaluation. Inspired by [10], the *data preparation and integration* module implements a harmonization process to get an integrated dataset of images from Sentinel and Landsat satellites. The process consists in several steps, as reported in Fig. 2, that are carried out by using Google Earth Engine (GEE) API.[1] As a first step, spectral images are acquired from Landsat 7, 8, and Sentinel-2, generally, satellite measurements are acquired at top-of-the-atmosphere reflectance, which is a mix of light reflected off the surface of the Earth and off the atmosphere, therefore, spectral images require some atmospheric corrections [11] to get rid of the effect of the atmosphere on the reflectance values and get images showing only the actual reflectance of the areas on the surface of the Earth (bottom-of-atmosphere reflectance). GEE provides land surface reflectance datasets, where atmospheric corrections have already been done so that images are reliable for vegetation analysis. The second step is called *cloud covering image filtering* that consists of removing cloud covering, to prevent these phenomena, only images with cloud covering at 0% of cloud covering have been acquired. The *Landsat 8-Sentinel-2 coregistration* step processes Landsat 8 and Sentinel-2 images to revise misalignment (precisely, 38 m) between them [12]. This misalignment depends on the residual geolocation error in Landsat 8 framework. To solve this problem,

[1][Online]. Available: https://developers.google.com/earth-engine

the displacement between two overlapping images from the two satellites is measured thanks to GEE functionalities. The *reprojection, rescaling* step is in charge of reprojecting bands of Sentinel 2 (WGS84) and rescaling them (to 30 m) by using the cubic interpolation method [13]. After the image rescaling, the *BRDF correction, Topographic correction* step applies the bidirectional reflectance distribution functions (BRDF) approach, introduced in [14], to reduce the directional effects depending on differences in solar and view angles between the satellites. This method sets the view angle to the NADIR (i.e., the lowest point of the celestial sphere vertically downward to the observation point) and the illumination according to the center latitude of the tile. Another correction involves topography: the Sun-Canopy-Sensor Topographic Correction method [13] is used to account for reflectance variations generated by slope, aspect, and elevation. It handles mountain shadowing on the fields. The *band adjustment and rescaling* step consists of adjusting bands across satellites to make them compatible. Six bands, including blue, green, red, are adjusted by cross-sensor transformation coefficients provided by [15], where the coefficients have been determined by running absolute difference metrics and major axis linear regression analysis on 10 000 image pairs of the USA territory. The collection of harmonized band images generated by the abovementioned steps from the images of Landsat 7, 8, and Sentinel 2 form the so called harmonized dataset.

With the scope of extracting phenological data, the harmonic or Fourier analysis is applied to the generated data (*Harmonic analysis* step) due to its proven effectiveness in characterizing seasonal cycles and variation in cover types [16]. Thus, the normalized difference vegetation index (NDVI, see Definition 1) is calculated over the band images from the harmonized dataset, by the harmonic analysis over time series of NDVI values to recognize the growing seasons of each area. The harmonic regression model to determine fitted values on the NDVI time series is described by the following equation:

$$\begin{aligned} p_t &= \beta_0 + \beta_1 \cdot t + \alpha \cdot \cos(2 \cdot \pi \cdot \omega \cdot t - \phi) + \epsilon_t \\ &= \beta_0 + \beta_1 \cdot t + \beta_2 \cdot \cos(2 \cdot \pi \cdot \omega \cdot t) + \beta_3 \\ &\quad \cdot \sin(2 \cdot \pi \cdot \omega \cdot t) + \epsilon_t \end{aligned}$$

where p_t is the scalar pixel at the time t, ϵ_t is a random error, α is the curve amplitude, ω is the curve frequency, ϕ is the phase. Note that $\beta_2 = \alpha \cdot \cos(\phi)$ and $\beta_3 = \alpha \cdot \sin(\phi)$, which implies that the amplitude $A = (\beta_2^2 + \beta_3^2)^{1/2}$.

To fit the harmonic model to the NDVI time-series, ϕ is set to the number of cycles per unit time, hence the β_0, β_1, β_2, β_3 parameters are learned by applying the least squares method to solve regression and determine the curve on fitted values. The resulting curve allows studying the vegetation trends of the ROIs over the years, from which growing seasons on the cropland can be extracted. The seasonal variations in vegetation are recognized according to curve amplitude and phase angle: the higher the amplitude and phase angle, the higher the seasonal variations. By analyzing the curve amplitude, the phenology is extracted using three parameters: the start of the season (SOS), the senescence (SEN), and the end of the season (EOS). These parameters help to individuate the growing seasons characterizing various plant growth stages (seeding, flowering, fruiting, etc.). These data are acquired for each ROI and stored in the *ontological context building* module.

B. VI Calculation

The *VI calculation* module processes the images in the harmonized dataset, generated by the *data preparation and integration* module to assess VIs. The VIs allow the evaluation of various vegetative features, including the vegetation coverage, vigor, nitrogen presence, moisture, and water presence in the soil. A VI is assessed through formulas combining different bands; for instance, the most famous VI called NDVI is assessed by relating the red band image with the NIR band image.

Let us consider Red and NIR be, respectively, the red band value and the NIR band value in the pixel p, the NDVI is assessed as follows:

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}. \tag{1}$$

The red value and the NIR value in the same pixel p of the image are used to assess the NDVI value in that pixel. This way, by processing all the pixels of the image, a new raster image with the NDVI values is generated. This image evidences high coverage and vigor in vegetation with higher pixel values, and low coverage and vegetation with lower pixel values. NDVI has been chosen to evaluate the vegetation in terms of variations of greenness, since it has been proven to have a very sensitive response to green vegetation, even for areas with low vegetation coverage. Additionally, it is the most used index for regional and global vegetation assessments and, it is related not only to canopy structure and Leaf Area Index (LAI), but also to canopy photosynthesis [17].

NDVI describes vegetation in a specific moment in time, other indices, instead, such as SVI, are used to evaluate vegetation parameters over time. The SVI can provide an evaluation of vegetation and plant health that takes into account the comparison of current values with others acquired in past years. This index is also considered a drought index and helpful in preventing natural hazards. As introduced in [18], the SVI is mainly based on the NDVI and describes the probability of variation from the normal NDVI over multiple years of data, on a weekly time step. The index is defined as the z-score deviation from the mean in units of the standard deviation, calculated from the NDVI values assessed for each pixel location of a composite period for each year during a given reference period. Therefore, let us consider the pixel i in the image I, the week j in the year k over a period of n years, the SVI is defined as follows:

$$z_{i,j,k} = \frac{\text{NDVI}_{i,j,k} - \mu_{i,j}}{\sigma_{i,j}} \tag{2}$$

where $z_{i,j,k}$ is the z-score for the pixel i, in the week j during year k; $\text{NDVI}_{i,j,k}$ is the NDVI value on the same pixel (i) in the same period (week j, year k); $\mu_{i,j}$ is the mean for pixel i in the week j over the n years considered and $\sigma_{i,j}$ is the standard deviation for the pixel i in the week j over the n years. Another important index is the normalized burn ratio (NBR) used to detect the

presence of recently burnt areas. It relates the reflectance of the land surface in the NIR and SWIR bands to detect burnt areas. The rationale behind this is that the vegetation presents high reflectance in the NIR band and a low reflectance in the SWIR band, while it is the exact opposite in burnt areas (low reflectance in the NIR band, high reflectance in the SWIR band). NBR has been considered to depict changes in soil moisture due to its proven features to detect scars in the soil, perform vegetation monitoring years after a fire, and achieve satisfactory results in supporting vegetation recovery monitoring [19]. Let NIR be the value of NIR band in a pixel p and SWIR be the SWIR value in p, the NBR is defined in function of these two bands as follows:

$$\text{NBR} = \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}}. \tag{3}$$

Low values of NBR identify recently burnt areas, while high values denote vegetation. Values around the 0 denote a nonburnt area.

To assess the burn severity of an area over time, another index is delta NBR (dNBR), which considers several NBR values calculated over time to detect ROIs that have been affected by the fire. High, greater-than-0, dNBR values denote severe damages caused by the fire, while negative values indicate regrowth situations following a fire episode. Let NBR^t and NBR^{t+1} denote the NBR assessed on a pixel p at the time t and the NBR assessed on p at a later time $(t+1)$, respectively, the dNBR is calculated as follows:

$$\text{dNBR} = \text{NBR}^t - \text{NBR}^{t+1}. \tag{4}$$

The formula is applied to pixels of consecutive images to evaluate the severity of fire between two consecutive moments in time on ROIs. USGS provides a classification table to evaluate the dNBR value and reports a corresponding burn severity status.

The SVI and dNBR values calculated on pixels of ROIs are input to the *area monitoring* module to handle anomaly detection.

C. Ontological Context Building

The *ontological context building* module stores as semantic assertions the collected information on the ROIs and the acquired phenological data and vegetation parameters. To achieve this task, a well-known ontology, composed of various subontologies, has been employed to model knowledge on the agriculture domain and support environmental monitoring. This ontology, taken from Candela project,[2] has then been extended to include knowledge related to the harmonized spectral images and area phenology. The core ontology model is shown in Fig. 3, where the "Feature Of Interest" class represents all the territorial elements to monitor, while the *GeoFeature* class includes the specific territorial units, such as the ROIs in the environment identified by an ID. Other classes identify the feature of the ROI to observe (e.g., vegetation coverage) and the parameters, such as VIs, to evaluate the feature (i.e., instances of the "GeoFeature Observable Property" class). The classes in red contour, including *GeoFeatureCollection*, "Growing Season" and *Anomaly*, are

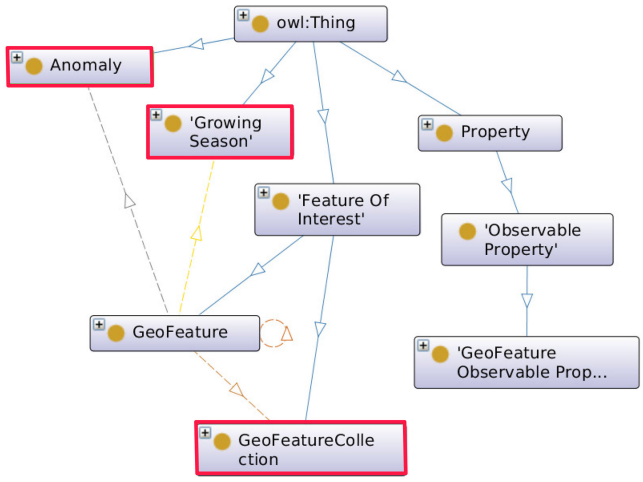

Fig. 3. Extended ontology model: new classes are evidenced with a thick red contour.

imported by TOM to represent knowledge on the areas, their phenology and anomalies. The *GeoFeatureCollection* represents an AOI as a set of all the ROIs lying within the same AOI. The "Growing Season" class allows knowledge modeling on the extracted growing seasons of a ROI. The *Anomaly* class represents the kind of anomaly (depending on the VI, e.g., SVI anomaly, dNBR anomaly) that affects a ROI (*GeoFeature* instance).

When the *data preparation and integration* and *VI calculation* modules accomplish their tasks, they pass their data to the *ontological context building* module. At this time, the incoming data are used to populate the integrated and extended TOM ontology with facts about the ROIs and their phenology, and vegetation parameters assessed on the ROIs. In detail, each ROI, encoded as a *GeoFeature* instance and equivalent to a *Feature* instance in GeoSPARQL, is stored along with its GPS data (area and position). The growing seasons, which have been detected for each ROI by the *data preparation and integration* module, are added as a "Growing Season" instances to the ontology. Along with each "Growing Season" instance, the SOS, EOS, and SEN parameters describing the growing season are also added to the ontology and related to the growing season through data properties. This way, the ROI (*GeoFeature* instance) is related to its detected growing seasons "Growing Season" instances).

The KB can be queried to extract ROIs and phenological data when it is required. Specifically, when the *area monitoring* module needs to interpret the SVI and dNBR, provided by the *VI calculation* module (see Section IV-D), sends a request to the *ontological context building* module that queries the KB to get the SVI values on a ROI according to its growing seasons. The query results are returned to the *area monitoring* module that evaluates the SVI for checking the presence of an anomaly. Examples of queries will be presented as part of the case study discussed in Section V.

Finally, when the *area monitoring* detects an anomalous SVI or dNBR value, the anomaly is detected and added by the *ontological context building* module as an *Anomaly* instance to the ontology. The anomalous SVI value, a string expressing

[2][Online]. Available: https://candela-h2020.eu/

TABLE I
ANOMALY INTENSITY ACCORDING TO SVI

SVI range	Anomaly intensity
$[-1, 0]$	Low
$[-2, -1]$	Medium
> -2	High

the level of the anomaly and the day of the year when the anomaly happened are also added to the ontology and related to the anomaly through ad-hoc data properties. This way, the KB is filled with facts about ROIs (*GeoFeature* instances) and their anomalies (*Anomaly* class instances) to support vegetation analyses in retrospect on the area.

D. Area Monitoring

The *area monitoring* module is in charge of analysing the acquired data, find eventual anomalies on vegetation and notify about it to humans. It gets the SVI and dNBR values from the *VI calculation* module; then it retrieves contextual data about the ROI querying the *ontological context building* module. The latter, in turn, returns the phenological data to the *area monitoring* module. Depending on the growing season, the SVI values can be considered for the evaluation of the anomaly on vegetation.

In this case, the *area monitoring* module assesses the intensity of anomaly by applying if–then rules to SVI values. In this study, Table I has been defined to classify anomaly intensity based on the SVI values. The table includes several ranges that are selected in accordance with the z-score definition at basis of SVI [see (2)]. In detail, in a normal standard distribution, the z-score z can be determined from every random variable X by the following formula:

$$z = \frac{(X - \mu)}{\sigma} \tag{5}$$

where μ and σ are the mean and standard deviation of the distribution.

The z-score represents how many standard deviations the value X is away from the mean. From the sigma rule [20], it is known that in a large dataset, the 68% of the values lie within the first standard deviation ($[\mu - \sigma, \mu + \sigma]$), the 95% of the data within two standard deviations ($[\mu - 2\sigma, \mu + 2\sigma]$) and 99.7% within three standard deviations of the mean ($[\mu - 3\sigma, \mu + 3\sigma]$). Since in normal distribution $\mu = 0$ and $\sigma = 1$, the resulting ranges ($[-1, 0]$, $[0, 1]$, etc.) on deviations have been associated with labels expressing different types of anomaly intensity and used to define the classification as shown in Table I. The table reports three classes of anomaly intensity based on three ranges of negative values. Positive values are not given, since they represent nonanomalous situations. By using this table, the *area monitoring* checks the presence of the anomalies occurred in the scene. In a nutshell, if the *area monitoring* module detects positive SVI values, no anomaly on vegetation occurs, and the module keeps monitoring the environment by analyzing the incoming data. Otherwise, if the SVI value is negative, the corresponding anomaly intensity is returned.

TABLE II
BURN SEVERITY CLASSIFICATION TABLE BASED ON dNBR VALUES PROVIDED BY USGS

dNBR range	Anomaly severity
$[-500, -251]$	Enhanced regrowth, (high) post fire
$[-250, -101]$	Enhanced regrowth, (low) post fire
$[-100, 99]$	Unburned
$[100, 269]$	Low severity
$[270, 439]$	Moderate-low severity
$[440, 659]$	Midrate-high severity
$[660, 1300]$	High severity

For what concern burn severity evaluation, the *area monitoring* module interprets the dNBR values provided by the *VI calculation* module by using a classification table developed by USGS[3] and reported in Table II. It reports various ranges of the dNBR values and the corresponding class of burn severity; the positive values evidence increasing levels of severity, while values around 0 and negative values denote unburned and postfire regrowth scenarios.

When an anomaly occurs, the *Area monitoring* module generates an alert message to humans containing the description of the detected anomalies, the anomaly intensity and the burn severity.

V. SYSTEM AT WORK

To show the potential of the proposed system, real-time environment monitoring has been simulated by using Apache Kafka Stream. The technology is used to implement the modules of our architecture and support the communication between the system components and human operators. The potential and efficiency of this implementation are proved through practical case studies.

A. Kafka-Based Implementation

The system has been implemented as an online tool capable of acquiring data from the environment, detecting and reporting anomalies in vegetation to humans in simulated real-time. To this purpose, a stream processing application based on Apache Kafka Stream has been developed to allow the framework to process data and interact with humans in real-time. In a Kafka network, two main entities, namely producers and consumers, exchange messages over the network through reading and writing operations on a Kafka topic. A topic is a category or feed name to which data is published by producers. In detail, the communication can be synchronous or asynchronous, therefore, the consumers wait for new incoming events on the topic, when the producers write messages on topic, the consumers will read the new events. The Kafka-based stream processing application is built on a processor topology defining its computational logic. A topology is composed of stream processors (nodes) that are connected by streams (edges) or shared state stores.

The architecture of the framework implemented as Kafka-based streaming application is presented in Fig. 4, along with the processor topology. At basis of this architecture, there are the producers in charge of feeding new data into the network. When new data are acquired by UVs and satellites, the producers

[3][Online]. Available: https://un-spider.org/advisory-support/recommended-practices/recommended-practice-burn-severity/in-detail/normalized-burn-ratio

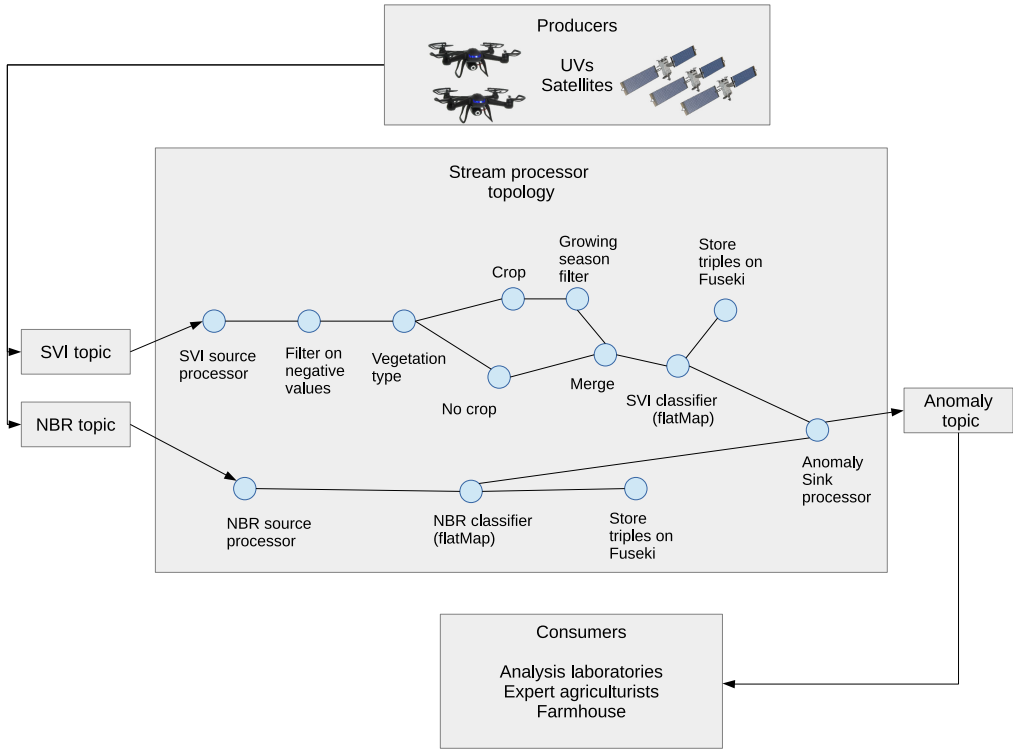

Fig. 4. Data stream processing for anomaly detection.

write these data on the topics. Data of the same type are written on the same topic. Two topics have been defined: one for SVI data and one for dNBR data. Then, the new data are processed and streamed out by the stream processors, that are organized according to the topology of Fig. 4. Each topic has its own stream processor (node) into two different subgraphs rooted at the *NBR source processor* and the *SVI source processor* that, respectively, process dNBR and SVI data.

The *NBR source processor* streams out dNBR data to the *NBR classifier* that performs the dNBR classification as explained in Section IV-D. According to the classification result, if anomalies are detected, the *store triples on Fuseki* node stores the data on the detected anomaly as semantic triples on the Fuseki triple store.

The *SVI source processor* node streams out SVI data coming from satellites and UVs to the *filter on negative values* node, that filters out all the positive values representing nonanomalous scenarios. The subsequent *vegetation type* processor splits ROIs in cultivated areas and noncultivated areas, which are then handled by two distinct processors (*no crop* and *crop* nodes). SVI values in cultivated areas are then processed by the *growing season filter* processor, that filters out data collected in nongrowing seasons. Data on crop and noncrop areas are merged on the ROIs by the *merge* node, to be finally processed by the *SVI classifier* processor that checks the presence of the anomaly and its intensity, as detailed in Section IV-D. In presence of anomalies, the *store triples on Fuseki* processor stores the data on the detected anomaly as semantic triples on the Fuseki triple store.

Anomalies detected on both the topics (SVI, dNBR) are finally processed by the *anomaly sink processor* that writes the detected anomalies on the *anomaly topic*. New anomalies written on the *anomaly topic* are collected by the consumers. Consumer outputs are read by experts on the agriculture domain working in analysis laboratories or from agriculturists in farmhouses.

B. Case Study in Simulated Real-Time

The case study is developed on an AOI in the province of Benevento, Italy, precisely, on three ROIs on the selected AOI. A map of the AOI to monitor is shown in Fig. 5, with its three ROIs coloured according to their type of environment: the light green-colored ROI_1 and the green-colored ROI_2 represent two distinct vineyards, while the ROI_3 in cyan is a wooded region.

By focusing on ROI_1, 327 spectral images have been acquired: 117 of them are taken from Landsat 7, 125 from Landsat 8 and 85 from Sentinel-2. Band images of interest for VI calculation are the red, NIR, and SWIR bands. They may have been acquired at a different wavelength and spatial resolution depending on the satellite, as shown in Table III comparing the three satellites involved. Sentinel-2 has different spatial resolutions for the three bands (10 m, 20 m) involved that are way smaller than the spatial resolution of Landsat satellites (30 m). Examples of processed images of ROI_1 from the Landsat 8 and Sentinel-2 satellites are shown in Fig. 6 along with map of the area.

According to the system overview shown in Fig. 1 data preparation is carried out by the *preprocessing and data integration* module, including satellite data harmonization. Two of the 85 images taken from Sentinel-2 are processed to remove excessive cloudiness, then the coregistration, rescaling and reprojection, BRDF correction and band adjustment tasks

Fig. 5. Area to monitor: an AOI showing Solopaca vineyards in Benevento, Italy, including three ROIs: ROI1 in light green (vineyard), ROI2 in green (vineyard), ROI3 in cyan (woodland). The image scale is 1:100000.

TABLE III
ACQUIRED SATELLITE IMAGE DETAILS

Sat	Band	Wavelength (μm)	Bandwidth (μm)	Spatial resolution (m)
Landsat 7	red	0.63 - 0.69	0.036	30
	NIR	0.77 - 0.90	0.030	30
	SWIR	1.55 - 1.75	0.080	30
Landsat 8	red	0.636 - 0.673	0.037	30
	NIR	0.851 - 0.879	0.028	30
	SWIR	1.567 - 1.651	0.085	30
Sentinel-2	red	0.646 - 0.685	0.039	10
	NIR	0.848 - 0.881	0.033	10
	SWIR	1.539 - 1.681	0.142	20

(a) (b) (c)

Fig. 6. Comparison between images captured from the editor map before the Band Adjustment phase: (a) is the ROI1 on the map; (b) is the L8 image (spatial resolution: 30 m); (c) is S2 image (spatial resolution: 10 m).

are accomplished on the collected images. Preprocessing helps improving the alignment between merged images coming from different satellites, by reducing differences caused by different spatial resolutions (see Section V-C for details on evaluations of the preprocessing task).

For each preprocessed NDVI image in the series, the mean NDVI is calculated as the average of the NDVI values in each pixel of the image. Then, to extract the phenological data about the ROI, the harmonic model described in Section IV-A is run over the NDVI time series. As shown in Fig. 7, the harmonic regression applied to ROI1 allows fitting the model to the NDVI time series as composed of the mean NDVI of each image captured in a given date (blue points) providing a curve on the fitted values (red values) depicting the main trend of the

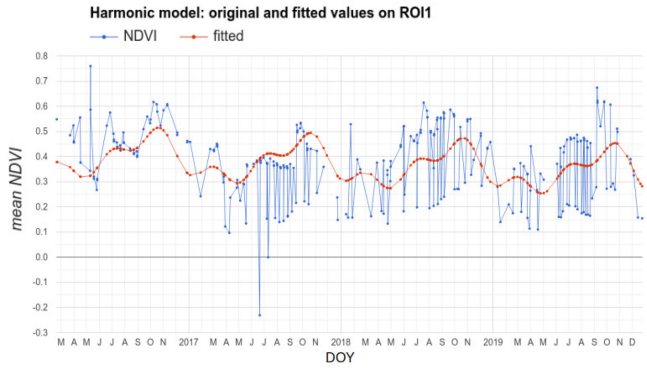

Fig. 7. Harmonic model over four years (from 2016 to 2020): harmonic regression generates a fitted model (red curve) on the NDVI time series (blue points) describing the phenological trend of the area.

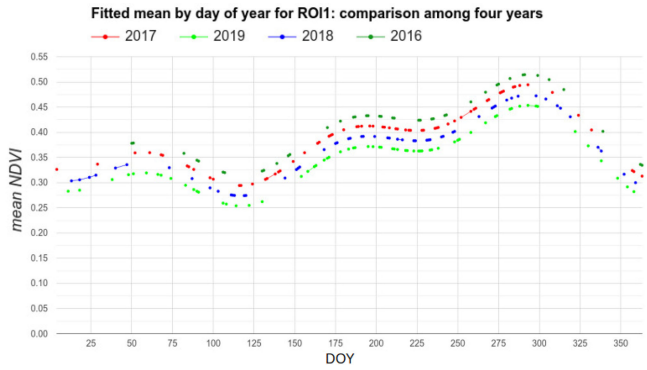

Fig. 8. Harmonic patterns: comparison of fitted curves for four years (2016, 2017, 2018, 2019) over one year time (365 days).

vegetation over the analyzed years. The amplitude and phase of this curve helps depicting changes in vegetation (i.e., curve peaks evidence the periods of the year where vegetation grows more). From a first glance, let us notice that the greenness has a little dip through the years, in fact, the red curve tends to go down a little as the years go by. A comparison of the fitted NDVI time series from year to year is displayed in Fig. 8 that shows some trends, such as the year with the lowest mean NDVI values (2016) or the one with the highest values (2019), as well as the very close NDVI time series throughout the year between 2017 and 2018. Let us notice from figure that the harmonic patterns of the four years are similar throughout the year timeline, this phenomenon may depend on the fact that the area was not subject to climatic changes, did not suffer from diseases (e.g., fungal attack or fire episodes) and the weather (temperatures, humidity, etc.) have registered little changes from year to year. By zooming in on the curve of Fig. 8, it is possible to see curve amplitudes and phases, as shown in Fig. 9, which allow the system to detect phenological data on the ROI (see Section IV-A). In this case, two growing seasons, identified by SOS, SEN, and EOS parameters expressed in day of the year (doy), are detected for ROI_1; they are: $[\text{SOS} = 124, \text{SEN} = 203, \text{EOS} = 231]$, $[\text{SOS} = 234, \text{SEN} = 299, \text{EOS} = 363]$.

Once the growing seasons have been detected, the ROIs and their growing seasons, expressed by SOS, EOS and SEN, are

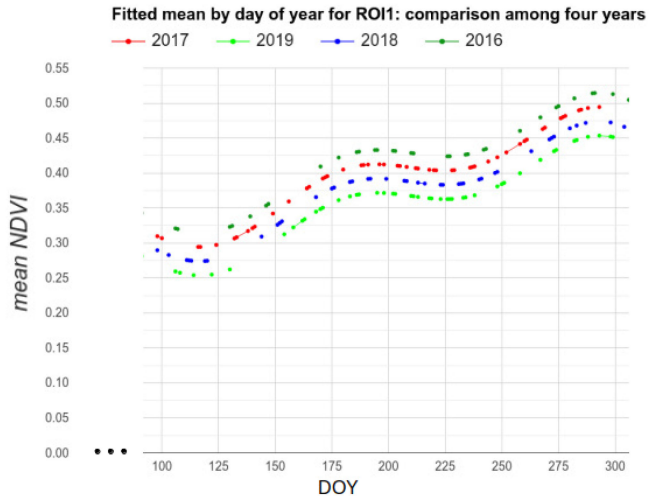

Fig. 9. Phenological data extraction: zoom in on the fitted curve shows how the growing seasons are selected.

```
1  PREFIX http: <http://www.w3.org/2011/http#>
2  PREFIX tom: <http://melodi.irit.fr/ontologies/tom.owl#>
3  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#
       >
4  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
5  PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
6
7
8  SELECT ?ROI ?anomaly ?value ?level ?doy
9  WHERE {
10   ?ROI rdf:type tom:GeoFeature.
11   ?ROI tom:hasGrowingSeason ?growingSeason.
12   ?growingSeason tom:SOS "124"^^xsd:integer.
13   ?growingSeason tom:EOS "231"^^xsd:integer.
14   ?ROI tom:hasAnomaly ?anomaly.
15   ?anomaly rdfs:label "SVI".
16   ?anomaly tom:anomaly_level ?level.
17   ?anomaly tom:anomaly_value ?value.
18   ?anomaly tom:day_of_year ?doy.
19   FILTER (?doy >= "124"^^xsd:integer && ?doy <= "231"^^
         xsd:integer)
20 }
21 ORDER BY ?ROI ?growingSeason ?doy
```

Listing 1: Query to get vegetation anomalies (SVI anomalies) on ROI 1.

encoded into triples by the *ontological context building* module (see Section IV-C), and stored on Fuseki triple store. Then, the system simulation starts with reproducing the SVI and dNBR values stream from 01/01/2020 to 31/12/2020, for each of the three ROIs considered.

Each satellite and UV are simulated by Kafka producers that generate every 4 s SVI and dNBR values data from the collected VI data, aimed at simulating the time passage (specifically, each day lasts around 4 s). In this case study, three producers have been defined: P_1, P_2, P_3. The P_1 thread represents a satellite that generates SVI values, the P_2 thread simulates a UV that provides SVI values by using a random generator and the P_3 thread is a satellite providing NBR values. These threads randomly select data entries from the collected VI data to be sent in a day to consumers, in accordance with the time passage simulated by the time thread. Therefore, each time a producer generates new data, these are written on a specific topic, according to the type of data detected (e.g., SVI, dNBR). The streamed data are processed by the processor topology, as detailed in Section V-A, and then, the detected anomalies are written in output to the consumers. In this case study, the consumer threads represent an analysis laboratory and a farmhouse, respectively. The consumers wait for new eventual upcoming data, and whenever anomalies are detected, they read the ROI affected by the anomaly, the type and intensity of the detected anomaly from the messages sent by the producers on the SVI and dNBR topics.

The detected anomalies are notified to end-users by the consumer threads. Then, the *environment monitoring* module allows the end-users (e.g., experts, agronomists) to query the knowledge base to have details on the occurred anomalies. For instance, if the end-users want details on vegetation anomalies affecting the vineyard ROI_1 in the first growing season ($[\text{SOS} = 124, \text{SEN} = 203, \text{EOS} = 231]$), the system runs an ad-hoc query. This query is shown in Listing 1, it retrieves the ordered list of detected anomalies related to the ROI (line 14) that lie in the reference growing season (line 11) by comparing the doy (line 18), when the anomaly happened to the SOS and EOS of the season (lines 12, 13, 19). The returned results, reported in Table IV, show the anomalies on vegetation according to the SVI value in the year period between the doy 127 and 133. On average, the anomaly intensity is low or medium on ROI_1 according to the SVI values. During the first growing season, no severe soil conditions are found, as shown in Table V where the dNBR-based anomalies detected on ROI_1 are listed, by using a query similar to the previous one.

TABLE IV
ANOMALIES ON VEGETATION OF ROI_1 IN THE FIRST GROWING SEASON

ROI	anomaly	value	intensity	doy
tom:ROI1	tom:SVI1_ROI1	-0.56033295	Low	127
tom:ROI1	tom:SVI1_ROI1	-0.999617	Low	127
tom:ROI1	tom:SVI1_ROI1	-1.4176002	Low	127
tom:ROI1	tom:SVI1_ROI1	-0.56033295	Medium	127
tom:ROI1	tom:SVI1_ROI1	-0.999617	Medium	127
tom:ROI1	tom:SVI1_ROI1	-1.4176002	Medium	127
tom:ROI1	tom:SVI2_ROI1	-0.46167532	Low	130
tom:ROI1	tom:SVI2_ROI1	-0.5535184	Low	130
tom:ROI1	tom:SVI2_ROI1	-0.7194249	Low	130
tom:ROI1	tom:SVI1_ROI1	-0.56033295	Low	133
tom:ROI1	tom:SVI1_ROI1	-0.999617	Low	133
tom:ROI1	tom:SVI1_ROI1	-1.4176002	Low	133
tom:ROI1	tom:SVI1_ROI1	-0.56033295	Medium	133
tom:ROI1	tom:SVI1_ROI1	-0.999617	Medium	133
tom:ROI1	tom:SVI1_ROI1	-1.4176002	Medium	133

TABLE V
ANOMALIES ON POST-FIRE CONDITIONS OF ROI_1 IN THE FIRST GROWING SEASON

ROI	anomaly	value	severity	doy
tom:ROI1	tom:dNBR8_ROI1	0.268	Unburned	143
tom:ROI1	tom:dNBR9_ROI1	0.217	Unburned	183
tom:ROI1	tom:dNBR10_ROI1	0.205	Unburned	203
tom:ROI1	tom:dNBR11_ROI1	0.187	Unburned	208
tom:ROI1	tom:dNBR12_ROI1	0.193	Unburned	213
tom:ROI1	tom:dNBR13_ROI1	0.174	Unburned	223

C. Tests on Case Study Results

According to [10], the efficiency of the preprocessing stage has been evaluated by assessing the Pearson correlation between two overlapped images of the ROI1 with the same date and hour, taken from Landsat and Sentinel satellites on the red, NIR, and SWIR bands used for calculating the VIs considered.

TABLE VI
PEARSON CORRELATION (R) ON A HARMONIZED IMAGE OF THE ROI1, COMING OUT FROM LANDSAT AND SENTINEL INTEGRATION

Preprocessing step	r (red)	r (NIR)	r (swir)
Original data	0.46	0.34	0.58
cloud covering	0.46	0.34	0.58
co-registration	0.57	0.47	0.88
rescaling	0.81	0.80	0.88
BRDF correction	0.92	0.89	0.90
band adjustment	0.92	0.98	0.99

For instance, the achieved results on a couple of images captured on June are reported in Table VI, the correlation has been evaluated at each preprocessing step, from the original data acquisition to the band adjustment phases. Let us notice that the correlation value increases from values between 34% and 58% on the starting data to values greater than 92% after the last preprocessing subtasks for all the bands involved. This result demonstrates that the preprocessing perfectly accounts for the spatial difference between the original images taken from the satellites and guarantees the quality of the built dataset. Depending on the Pearson value between images from Sentinel-2 and Landsat satellites, images can be further filtered to improve the quality of the harmonized dataset (i.e., images with low correlation can be filtered out).

To evaluate the accuracy of our system at detecting the growing seasons of the ROIs from the NDVI time series, the coefficient of determination has been calculated on the harmonic regression results. The coefficient of determination (R^2) is defined in terms of the residual sums of squares ($\mathrm{SS_{res}}$) and the total sums of squares ($\mathrm{SS_{tot}}$) as follows:

$$\mathrm{SS_{res}} = \sum_i (y_i - f_i)^2 \tag{6}$$

$$\mathrm{SS_{tot}} = \sum_i (y_i - \mu)^2 \tag{7}$$

where $y_1, y_2, \ldots, y_n$ are the values of the NDVI time series, $f_1, f_2, \ldots, f_n$ are the predictions generated by the harmonic regression (fitted values) and $\mu = 1/n \sum_{i=1}^{n} y_i$. Therefore, R^2 is calculated as follows:

$$R^2 = 1 - \frac{\mathrm{SS_{res}}}{\mathrm{SS_{tot}}} \tag{8}$$

An R^2 very close to 1 implies a good fitting curve, otherwise the fitted curve fails in describing the vegetative growing trend. The coefficient has been calculated on the NDVI time series determined from the harmonized dataset (blue points in Fig. 7) and the predicted values (red points in Fig. 7). The calculated R^2 is equal to 0.94, which means that the assessed fitted curve (red curve in Fig. 7) perfectly describes the vegetative growing trend of the selected area.

Since no field measurements on the selected geographical area are available, the accuracy has been assessed by building a "Ground Truth (GT)" dataset composed of Landsat 8 high-quality images (with cloudiness coverage at 0%), manually checked to remove any anomalous values and/or missing pixels (due to cloud or sensor damage). The NDVI and NBR values

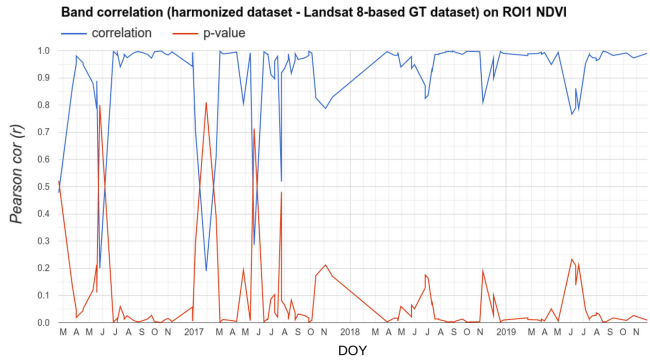

Fig. 10. Pearson correlation (ρ) between NDVI images of our harmonized dataset and Landsat 8.

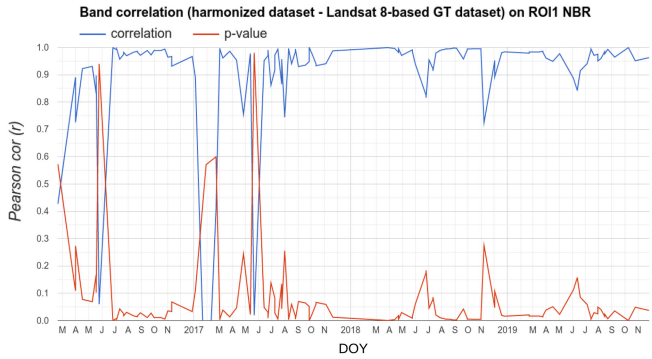

Fig. 11. Pearson correlation (ρ) between NBR images of our harmonized dataset and Landsat 8.

calculated on the harmonized dataset, achieved by fusing data from Landsat 8, Landsat 7, and Sentinel-2, have been compared with the same indices assessed on the GT dataset, over the same period. The use of Landsat to build the GT dataset provides high-resolution images and accurate estimation parameters and resolution comparable to our harmonized dataset. The similarity between two images, one belonging to the harmonized dataset and the other to the ground truth dataset, has been evaluated by the Pearson coefficient (ρ). A Pearson correlation value ($|\rho|$) greater than 0.7% represents a strong correlation between two images. Therefore, the stronger the correlation, the more accurate the NDVI-based vegetation status evaluation or the NBR-based soil status evaluation. Pearson correlation on the NDVI images and NBR images between the harmonized dataset and Landsat-based ground truth dataset are reported in Figs. 10 and 11. The correlation between the harmonized dataset and the ground truth dataset on the NDVI is constantly very strong (ρ closer to 1) over the four years considered, except for few values occurring in May–June 2016, early months of 2017 (January, February) and May 2017. Similarly, the correlation on the NBR images is very strong throughout the four year period, except in May 2016, January, February and between May and June 2017. The achieved results demonstrate the accuracy of the VIs assessed on the harmonized datasets, and consequentially the reliability of our approach in the anomaly detection.

VI. Discussion

The approach introduced several benefits to multidevice field smart monitoring with respect to existing solutions. Exiting semantics-based methods in literature introduce new ontologies aligned to existing ones for knowledge representation of remote sensing images [21], define rules to classify images exploiting expert-knowledge [22], build knowledge on data coming from just one satellite [7], [21], or use the ontology constructs to help interoperability on high-level information coming from different satellites [5]. Our approach, instead, not only interfaces heterogeneous data through the knowledge but provides a way to achieve a near-daily dataset of reliable images for vegetation monitoring and a knowledge base to help anomalous vegetation status analysis in two integration steps, one at low level and one at the high level. In fact, first, the approach fuses, at the row level, images coming from different satellites (Landsat 7, Landsat 8, Sentinel 2) through data harmonization, then it builds high-level knowledge on the indices calculated from those images to contextually relate VIs, depicts soil and vegetation conditions, and supports anomalous vegetation status analysis by querying the knowledge acquired. This approach provides the chance to assess more reliable VI values calculated by fusing data from different satellites (low-level integration) and exploiting ontological knowledge to relate the indices with contextual data (high-level integration) to reach a high-level explanation of soil and vegetation status. Additionally, data harmonization helps to satisfy the need for near-daily image acquisition, while the semantic annotation provides reusable knowledge that can be reused for a-posteriori analysis of the images, vegetation status, and occurred anomalies.

VII. Conclusion

This article presents a system model for IoT-based PAto detect anomalies in vegetation using phenological and environmental data. The system acquires spectral images from satellites and drones, that allow building an integrated dataset of near-daily-acquired high-resolution images. Harmonic regression is then applied to the images to extract phenological data about the monitored ROIs, which are then stored on a knowledge base. The acquired images are processed to assess VIs and field status over time by considering historical VI series acquired over the years. Finally, the model allows environment monitoring by classifying the VIs that lie in the detected growing seasons. The system introduces many benefits that support end-users (experts, agronomists, etc.) in vegetation monitoring for anomaly detection, summarized as follows:

1) The harmonization of the spectral images allows a consolidated store of near-daily high-resolution images necessary for achieving robust evaluation and continuing monitoring of crop fields.
2) The harmonic regression provides phenological parameters for assessing the growing seasons and vegetation trends of the specific ROIs.
3) The combination of the phenological context with the historical VI time series allows a more complete evaluation of the vegetation and soil anomalies on several areas (e.g., crop fields, wooded areas, meadows, etc.).
4) The knowledge base describing the ROIs and their phenological and environmental context help end-users to get appropriate data to analyze the occurred anomalies on vegetation and soil conditions.

Future research directions will be focused on an extension of the proposed framework to include and process data acquired from UVs and further data sensors to achieve vegetation status analysis through an enhanced contextual knowledge acquisition of the scenario.

References

[1] A. B. Torres, A. R. da Rocha, T. L. C. da Silva, J. N. de Souza, and R. S. Gondim, "Multilevel data fusion for the Internet of Things in smart agriculture," *Comput. Electron. Agriculture*, vol. 171, 2020, Art. no. 105309.

[2] E. Anastasiou, A. Castrignanò, K. Arvanitis, and S. Fountas, "A multisource data fusion approach to assess spatial-temporal variability and delineate homogeneous zones: A use case in a table grape vineyard in Greece," *Sci. Total Environ.*, vol. 684, pp. 155–163, 2019.

[3] R. Xu, S. Zhao, and Y. Ke, "A simple phenology-based vegetation index for mapping invasive spartina alterniflora using google earth engine," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 190–201, Jan. 2021.

[4] M. Mirhosseini, F. Barani, and H. Nezamabadi-pour, "Design optimization of wireless sensor networks in precision agriculture using improved BQIGSA," *Sustain. Comput., Informat. Syst.*, vol. 16, pp. 38–47, 2017.

[5] A. R. Muljarto, J.-M. Salmon, B. Charnomordic, P. Buche, A. Tireau, and P. Neveu, "A generic ontological network for agri-food experiment integration–Application to viticulture and winemaking," *Comput. Electron. Agriculture*, vol. 140, pp. 433–442, 2017.

[6] P. Skobelev, E. Simonova, S. Smirnov, D. Budaev, G. Voshchuk, and A. Morokov, "Development of a knowledge base in the 'smart farming' system for agricultural enterprise management," *Procedia Comput. Sci.*, vol. 150, pp. 154–161, 2019.

[7] J.-F. O. Rolland *et al.*, "Candela: A cloud platform for copernicus earth observation data analytics," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 3104–3107.

[8] F. Rembold *et al.*, "ASAP: A new global early warning system to detect anomaly hot spots of agricultural production for food security analysis," *Agricultural Syst.*, vol. 168, pp. 247–257, 2019.

[9] D. Cavaliere, V. Loia, A. Saggese, S. Senatore, and M. Vento, "A human-like description of scene events for a proper UAV-based video content analysis," *Knowl.-Based Syst.*, vol. 178, pp. 163–175, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119301996

[10] M. D. Nguyen, O. M. Baez-Villanueva, D. D. Bui, P. T. Nguyen, and L. Ribbe, "Harmonization of landsat and sentinel 2 for crop monitoring in drought prone areas: Case studies of Ninh Thuan (Vietnam) and Bekaa (Lebanon)," *Remote Sens.*, vol. 12, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/2

[11] M. Nazeer *et al.*, "Evaluation of atmospheric correction methods for low to high resolutions satellite remote sensing data," *Atmospheric Res.*, vol. 249, 2021, Art. no. 105308. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016980952031245X

[12] D. Roy, "A note on the temporary misregistration of landsat-8 operational land imager (OLI) and sentinel-2 multi spectral instrument (MSI) imagery," *Remote Sens. Environ.*, vol. 186, pp. 121–122, 2016.

[13] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981.

[14] A. Poortinga *et al.*, "Mapping plantations in myanmar by fusing landsat-8, sentinel-2 and sentinel-1 data along with systematic error quantification," *Remote Sens.*, vol. 11, no. 7, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/7

[15] R. Chastain, I. Housman, J. Goldstein, M. Finco, and K. Tenneson, "Empirical cross sensor comparison of sentinel-2a and 2b MSI, landsat-8 oli, and landsat-7 etm top of atmosphere spectral characteristics over the conterminous United States," *Remote Sens. Environ.*, vol. 221, pp. 274–285, 2019.

[16] M. E. Jakubauskas, D. R. Legates, and J. H. Kastens, "Crop identification using harmonic analysis of time-series AVHRR NDVI data," *Comput. Electron. Agriculture*, vol. 37, no. 1, pp. 127–139, 2002.
[17] X. Jinru and B. Su, "Significant remote sensing vegetation indices: A review of developments and applications," *J. Sensors*, vol. 2017, pp. 1–17, 2017.
[18] A. Peters, E. Walter-Shea, L. Ji, A. Viña, M. Hayes, and M. Svoboda, "Drought monitoring with NDVI-based standardized vegetation index," *Photogrammetric Eng. Remote Sens.*, vol. 68, pp. 71–75, 2002.
[19] D. Fornacca, G. Ren, and W. Xiao, "Evaluating the best spectral indices for the detection of burn scars at several post-fire dates in a mountainous region of Northwest Yunnan, China," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1196. [Online]. Available: https://www.mdpi.com/2072-4292/10/8/1196
[20] F. Pukelsheim, "The three sigma rule," *Amer. Statistician*, vol. 48, no. 2, pp. 88–91, 1994.
[21] J. F. Aldana-Martín, J. García-Nieto, M. del Mar Roldán-García, and J. F. Aldana-Montes, "Semantic modelling of earth observation remote sensing," *Expert Syst. Appl.*, vol. 187, 2022, Art. no. 115838. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421012008
[22] S. Andrés, D. Arvor, I. Mougenot, T. Libourel, and L. Durieux, "Ontology-based classification of remote sensing images using spectral rules," *Comput. Geosci.*, vol. 102, pp. 158–166, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098300417302078

Danilo Cavaliere received the master's degree in computer science from the University of Salerno, Italy, in 2014, and the Ph.D. degree computer science and computer engineering from the University of Salerno, Italy, in 2020.

He reviewed papers for various journals, including *Information Sciences, Expert Systems with Application*, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING and IEEE ACCESS, he was in the TPC for IEEE Symposium on Intelligent Agents in 2019 (IEEE SSCI 2019), 2020 (IEEE SSCI 2020), 2021 (IEEE SSCI 2021) and in the International Conference on Emerging Techniques in Computational Intelligence (ICETCI 2021). His research interests are in the areas of artificial and computational intelligence, data science, data mining and knowledge discovery, areas in which he has published many papers.

Sabrina Senatore (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from the University of Salerno, Italy, in 1999 and 2004, respectively.

From 2005, she is currently a Faculty Member with the University of Salerno. Her current position is Associate Professor of computer science with the University of Salerno. Her current research interests include the development and application of intelligent systems based on the combination of techniques from soft computing, computational intelligence, text mining, information retrieval, semantic web and software agents, areas in which she has published extensively.

Prof. Senatore is the Chair of IEEE CIS Task Force on Intelligent Agents (TFIA), an Associate Editor of Information Science, Elsevier, and she is also an Editorial Board Member of *Applied Intelligence Journal* and *International Journal of Computational Intelligence Theory and Practice*.