applied sciences

MDPI

Article

A Proof of Concept for an Agrifood Data Space Based on Open Data and Interoperability

Cristina Martinez-Ruedas [1,*], Adela Pérez-Galvín [2] and Rafael Linares-Burgos [2]

[1] Department of Electronic and Computer Engineering, Campus de Rabanales, University of Cordoba, 14071 Córdoba, Spain
[2] Civil Construction and Engineering Projects, Department of Rural Engineering, University of Cordoba, 14071 Córdoba, Spain; g82pegaa@uco.es (A.P.-G.); g82libur@uco.es (R.L.-B.)
* Correspondence: z42maruc@uco.es

Abstract

The creation of unified, open, secure, reliable, and agile data spaces is essential for collecting, storing, and sharing data in a standardized and accessible manner, promoting data reuse and addressing current interoperability limitations. In this context, this research presents a proof of concept for a unified agronomic data space based on the structured integration of heterogeneous open data sources. The central hypothesis is that the automated acquisition, preprocessing, and harmonization of publicly available agronomic data can significantly improve accessibility, usability, and interoperability for agricultural decision support applications. To this end, a comprehensive analysis of relevant open data sources was conducted, followed by the design and implementation of configurable algorithms for automated data downloading, cleaning, validation, and integration. The proposed approach explicitly addresses key challenges such as heterogeneous data formats, inconsistent spatial and temporal resolutions, missing values, and outlier detection. As a result, a unified access point was developed, providing reliable agronomic information, including (i) preprocessed climatological time series, (ii) crop and phytosanitary data, (iii) high-resolution aerial orthophotography, (iv) remote-sensing imagery, (v) pest-related information, and (vi) time series of major vegetation indices. The proof of concept was implemented for olive groves in the Andalusian region of Spain; however, the methodology is fully transferable to other crops, regions, and institutional contexts where comparable open data sources are available. The results demonstrate the potential of shared agronomic data spaces to enhance data reuse, support scalable analytics, and facilitate interoperable, data-driven agricultural management beyond the specific regional case study.

Keywords: data space; interoperability; open data; APIs; algorithms; smart agriculture; digital agriculture

Check for updates

Academic Editors: Wei Ma and Xiu Wang

Received: 12 January 2026
Revised: 1 February 2026
Accepted: 7 February 2026
Published: 12 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Digital technologies are continuously transforming the economy and society, reshaping how organizations operate and make decisions. As a result, data have become a central asset for competitiveness and innovation across all economic sectors. Data lie at the core of digital transformation, as they represent qualitative or quantitative variables that enable the analysis of real-world phenomena and support evidence-based decision making. From an ownership perspective, data can be classified as personal or non-personal, as well as private or public.

According to their level of structure, data are commonly categorized as structured, semi-structured, or unstructured. Structured data, such as relational databases or spreadsheets, are organized according to predefined schemas and are relatively easy to manage. Semi-structured data exhibit a partial structure without a fixed schema, while unstructured data lack a predefined organization and represent the most complex category to process. Despite this complexity, unstructured and semi-structured data account for the majority of the information currently generated and offer significant potential for extracting valuable insights through data mining, machine learning, and advanced statistical techniques. The effective exploitation of these data types enhances knowledge generation and improves decision-making capabilities.

One of the main barriers to data exploitation is limited availability, often driven by data ownership and restricted access. To address this challenge, initiatives promoting data sharing have emerged, including Governments-to-Business (G2B), Business-to-Business (B2B), and inter-governmental exchanges (G2G) [1]. These approaches aim to facilitate broader access to data and ensure that the benefits of digital transformation are more evenly distributed. At present, however, a small number of large technology companies concentrate a substantial portion of global data assets, while small and medium-sized enterprises face greater obstacles in their digitalization processes due to limited resources and technical expertise.

In this context, the creation of data spaces based on unified norms and standards is essential to support digital transition, foster interoperability, and enhance competitiveness across sectors. Such spaces must be open, secure, reliable, and agile to effectively integrate data into production processes and decision-making workflows [2]. At the same time, the global volume of data is growing at an unprecedented rate, with forecasts estimating up to 175 zettabytes worldwide by 2025 [3]. This growth is accompanied by an increasing decentralization of data generation, driven by the expansion of edge computing capabilities. Together, these trends highlight the need for robust data management frameworks that enable the efficient handling, integration, and interpretation of heterogeneous data sources.

These changes in data management and handling are also necessary to drive artificial intelligence forward, as large volumes of data are required for its development. Currently, due to the lack of management and shared spaces, as well as the economic and energy costs involved, there are not enough data for the proper development of applications in this field. Data reuse for innovative purposes is crucial for artificial intelligence to fully take off without resulting in a consequent increase in environmental footprint. The quality, authenticity, and integrity of the data in this context are decisive, where interoperability problems have been identified so far, hindering the inclusion of different data sources [4]. Regarding the Sustainable Development Goals [5], the digital component will also be key to achieving the ambitions of the European Green Deal [6]. Digital solutions are key factors for advancing the circular economy, supporting the decarbonization of all sectors, and reducing the environmental footprint of all products. Therefore, data strategies tend to create a single data market to address the identified needs. Specifically, the European Data Strategy [7] aims to create a single European data space, open to data from around the world, where data are secure and companies also have easy access to a nearly infinite amount of high-quality industrial data, thus driving growth and creating value while minimizing the environmental and carbon footprint. For the European Commission, data spaces are key results of the European Data Strategy, and thus in the digital call for proposals on preparatory actions for common European data spaces, support for an open source middleware infrastructure is established “cloud to edge” benefiting all data spaces, and a data space support center that will ensure interoperability between data spaces by developing harmonization across the European data spaces technical framework [8]. In

this way, there are increasingly more initiatives to open controlled information, mainly from public administrations. Currently, a review of the Open Data Directive [1] has been carried out to ensure that the public sector facilitates the use of data. The European Open Data Portal [9] contains examples of open data availability. Additionally, the Horizon Europe Program has adopted investing in a large European data space project as a key measure to support advances in data technology [10]. Within this program, the creation of common spaces in strategic sectors is envisaged, defining nine common data spaces: (i) Manufacturing Industry; (ii) Mobility; (iii) Health; (iv) Financial Matter; (v) Energy; (vi) Agriculture; (vii) Public Administration; (viii) Competences; and (ix) the European Open Science Cloud.

Within the European common data space related to the agri-food sector, the aim is to improve the sustainability, performance, and competitiveness of this sector. This will contribute to the emergence of an innovative ecosystem that reinforces capacities for monitoring and implementing common policies, as well as new opportunities to monitor and optimize the use of natural resources. Another objective is to orient it towards the end customer and potential users, including farmers, where the digital divide is still high.

There are other ongoing European initiatives focusing, among other things, on addressing obstacles related to interoperability in the agri-food sector. Specifically, the Fostering precision agriculture and livestock farming through secure access to large-scale HPC-enabled virtual industrial experimentation environment empowering scalable big data analytics project (CYBELE) is a project funded by European funds from the H2020 program [11] framed in precision agriculture and livestock farming, presenting real industrial cases aiming to demonstrate that the convergence of high-performance computing technologies can revolutionize agriculture. It aims to ensure integrated and intermediary-free access to large volumes of data from different sources with the goal of extracting knowledge. CYBELE proposes a distributed Big Data architecture, a data management strategy, mathematical algorithms requiring high computing power, and visualization platforms. For this purpose, it proposes a layered architecture consisting of modular components that guarantee interoperability [12].

On the other hand, the Agricultural Interoperability and Analysis System project (ATLAS) [13], funded by European funds from the H2020 program, aims to develop an open platform and create a sustainable environment for innovative agriculture, encompassing hardware and software interoperability. The platform will allow the flexible combination of agricultural machinery, sensor systems, and data analysis tools to overcome the problem of interoperability [14]. Another objective is the definition of standards to enhance data-driven agriculture.

Furthermore, as evidenced in [15,16], new technologies allow massive data storage, such as cloud computing and Hadoop, as well as processing and analysis through big data and machine learning. More specifically, digital technologies enable the collection, analysis and utilization of data using sensors or images, advanced analytics or intelligent algorithms [17–19].

Finally, Building an Interoperable, Data-Driven, Innovative and Sustainable European Agri-Food Sector (DEMETER), also funded by the H2020 program, aims to carry out large-scale deployment of interoperable platforms focused on the farmer and based on IoT. To this end, 20 pilot projects have been defined in 18 countries. DEMETER aims to demonstrate the potential in interoperability by expanding advanced IoT standards to a global agricultural information model. Details can be found in [20], which are not included in this report as they are considered to deviate from the central theme.

On the other hand, Ref. [21] comprises a literature review and conceptualization of design principles of Agriculture Data Space implementation. Also, it includes an overview of

the most significant challenges and solutions for agricultural data sharing and exchange in agro-food value chains, and it provides a description of basic notions and principles behind data ecosystems and data spaces as potential novel solutions to address those challenges.

In this sense, the Agriculture International Chair (ENIA) has been created to address the challenges facing the agroforestry sector in terms of productivity, environmental impact, food security, and sustainability, where the transformation of current production systems and processes through digitization and data-driven intelligence interconnected with all its actors is essential [22]. On the other hand, the hub developed by the Association for Research in the Agri-Food Industry (AINIA) aims to interconnect and process the various existing sources of information in this area into a single system capable of reasoning, with a volume of interconnected data. Through the GlobalSafeFood II project, it is expected to generate an advanced datalake that combines external sources and internal sources of companies to feed food safety control prototypes and fraud risk analysis applications and combine them with the new intelligence capabilities of the system [23].

In Spain, under the framework of the Recovery, Transformation, and Resilience Plan funded by the European Union through the Next Generation EU, a strategic call has been launched to drive the digital transformation of key productive sectors. This initiative focuses on the development of demonstrators and use cases for secure data-sharing ecosystems. To this end, the Ministry of Digital Transformation and Public Administration has unveiled an advanced Sectoral Data Spaces Promotion Plan. The primary objective is to accelerate innovation, enhance competitiveness, and maximize value creation across economic sectors by enabling the establishment of secure, interoperable data-sharing infrastructures. These spaces are designed to unlock the full potential of the European single data market, fostering data-driven collaboration and economic growth. The initiative will mobilize a €500 million budget from the Recovery, Transformation, and Resilience Plan. It is structured around six strategic pillars and eleven targeted initiatives, with implementation scheduled through 2026 [24].

The approaches of the different proposals are still to be validated, but the commitment of the H2020 program to improving interoperability in the agricultural sector highlights the benefits that compatible and integrated data can provide to precision agriculture.

All these studies and initiatives aimed at solving the problems of interoperability and semantic information extraction from data highlight that data availability is fundamental for decision-making. Therefore, the creation of integrated data spaces from various coherent and interoperable sources, the development of architectures that promote their access and reuse, as well as the establishment of fair access standards with an open approach will allow all companies, large or small, to benefit from the data.

In this regard, there are numerous publications related to field data acquisition mechanisms and isolated platforms for their management and visualization [25–30]. However, the true value of data lies in sharing and reusing it [1]. Therefore, significant efforts are being made to create common data spaces [7,13,14,20,31,32] that overcome existing barriers in global data management. The main obstacles encountered are the complexity of data management [33], lack of interoperability [13,14,31], insufficient storage centers and processing platforms [34,35], and the scarcity of reference architectures [33,36–40]. Overcoming these limitations would fully harness the potential of data analysis and management, strengthening the capabilities of decision support systems.

For these reasons, this study aims to develop a proof of concept for a unified data space that enhances agricultural management and decision-making processes. The central hypothesis is that integrating diverse open data sources through standardized methods can improve data accessibility, interoperability, and usability in the agricultural sector. To achieve this, open data sources were analyzed and characterized, and mechanisms for

data extraction, cleaning, and integration were implemented via Application Programming Interfaces (APIs). Additionally, specific developments were proposed to prototype a shared data space that enables automatic data integration, promotes reuse and scalability, and facilitates access through user-friendly interfaces.

2. Materials and Methods

The study area has focused on Spain (Southern Europe), as the main consulted sources of open data come from the national territory. However, some of the developments are transferable to other territories.

2.1. Data Sources

In the present study, the following data sources have been evaluated (Figure 1).

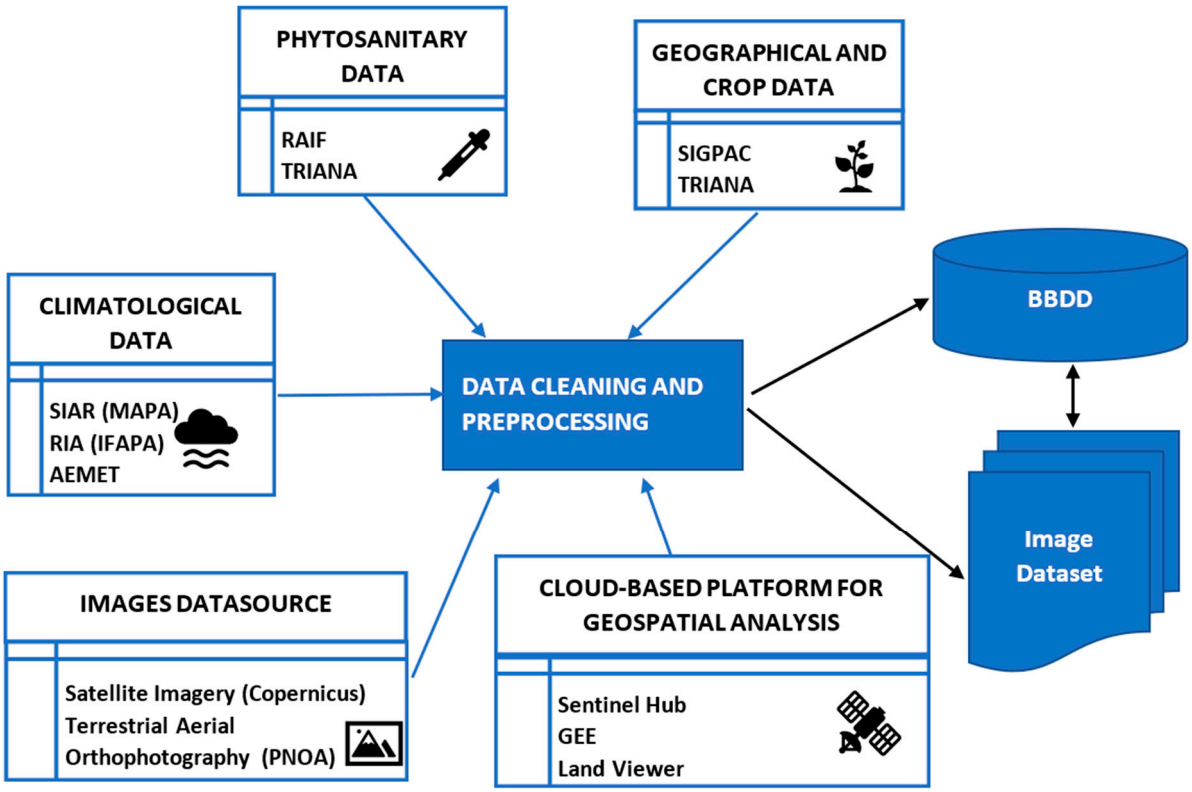


Figure 1. Integration and cleaning of different data sources (climatological, image datasets, remote sensing, phytosanitary, geographic and crop data) to generate a unified data space.

1. Climatological Data: (i) Agricultural Information System for Irrigation (SIAR, by its acronym in Spanish) [41]; (ii) Andalusian Agroclimatic Information Network (RIA, by its acronym in Spanish) [42]; (iii) State Meteorological Agency (AEMET, by its acronym in Spanish) [43].
2. Image Datasets: (i) Copernicus program [44]; (ii) National Plan for Aerial Orthophotography (PNOA, by its acronym in Spanish) [45].
3. Cloud-based Platforms for Geospatial Analysis of Satellite Images: (i) Sentinel Hub [46]; (ii) Google Earth Engine (GEE) [47]; (iii) EOS Land Viewer [48].
4. Phytosanitary Information: (i) Andalusian Phytosanitary Alert Network (RAIF, by its acronym in Spanish) [49]; (ii) Integrated Treatments in Andalusia in Agriculture (TRIANA Program, by its acronym in Spanish) [50].
5. Geographical and Crop Information: (i) Agricultural Parcel Geographic Information System (SIGPAC, by its acronym in Spanish) [51]; (ii) TRIANA Program [50].

2.2. Algorithm Implementation and Parameter Configuration

Figure 2 provides an overview of the complete processing workflow implemented in this study and serves as a conceptual guide to the methodological steps described in detail below. As illustrated in the figure, the proposed pipeline follows a sequential and modular structure, starting from automated data acquisition and storage, and progressing through exploratory analysis, automated validation, and data integration. Each stage of the workflow has been designed to support robustness, automation, and reproducibility, ensuring that heterogeneous agronomic data sources can be consistently processed and harmonized within a unified data space.

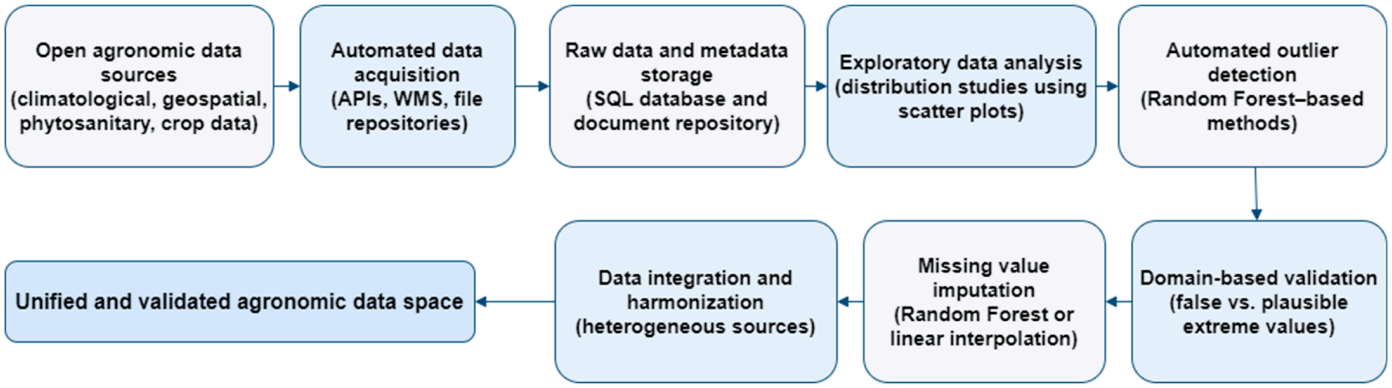


Figure 2. Workflow of the agronomic data space, showing automated data acquisition, validation, cleaning, and integration processes leading to a unified and updated agronomic dataset.

In this context, the following paragraphs describe the specific algorithms and techniques applied at each stage of the workflow, including exploratory distribution analysis, outlier detection, missing value imputation, and data integration. The rationale behind the selection of algorithms, parameter configurations, and software libraries is detailed to enhance transparency and facilitate reproducibility.

The algorithms developed in this study were designed with a focus on robustness, automation, and reproducibility, rather than on predictive optimization. For this reason, widely adopted and well-established libraries were used, and parameter configurations were selected based on standard practices reported in the literature.

For outlier detection and data validation in climatological datasets, a combination of exploratory and automated techniques was employed. First, distribution studies of the stored variables were conducted through scatter plots [52] in order to identify inconsistent values, anomalous patterns, and potential discrepancies between data sources. This visual analysis provided an initial validation step and supported the subsequent automated detection process.

Outlier detection was then performed using a Random Forest-based approach implemented through the outForest library [53]. This method was selected due to its ability to capture nonlinear relationships and interactions among variables, making it suitable for identifying anomalous observations in multivariate datasets. Since the objective of the Random Forest model was anomaly detection and data validation rather than predictive optimization, no exhaustive hyperparameter tuning was performed. Instead, the default parameter values recommended by the library were adopted to ensure robustness and reproducibility across heterogeneous datasets. Outliers were identified based on the ranking returned by the algorithm, which measures the deviation between observed and predicted values. Only those instances identified as false or inconsistent values after domain-based validation were corrected using the model-predicted values.

Missing value imputation in time-series data was carried out using the imputeTS [54] and missForest [53] libraries. Linear interpolation was applied for short temporal gaps in vegetation index time series, while Random Forest-based imputation was used for multivariate climatological datasets. These methods were selected due to their robustness in handling nonlinear relationships and mixed data types.

All data acquisition, preprocessing, and analysis pipelines were implemented using Python and R programming languages. Table 1 summarizes the main algorithms, libraries, and software versions used in this study to support reproducibility.

Table 1. Summary of the algorithms, software libraries, programming environments, and key parameter settings employed across the different stages of the agronomic data space processing workflow.

Task	Purpose	Algorithm/ Method	Library	Programming Language	Version
Data acquisition	Automated download of open data	REST API calls	requests	Python	≥2.28
Exploratory data analysis	Validation of variable distributions and detection of inconsistencies	Scatter plots	plot3D	R	1.4
Outlier detection	Identification of anomalous values in multivariate datasets	Random Forest	outForest	R	0.1.1
Missing value imputation (climatological data)	Imputation of missing values in multivariate time series	Random Forest	missForest	R	1.5
Missing value imputation (vegetation indices)	Interpolation of short temporal gaps	Linear interpolation	imputeTS	R	3.2
Geospatial data processing	Handling of vector and raster data	Geospatial analysis tools	geopandas	Python	≥0.12
Data storage and management	Metadata storage	Relational database	SQL	-	-

3. Implementation and Results

This section has been divided into four sections. The first section, Section 3.1, studies the different open data sources. After that, Section 3.2 details the automation process of the data sources selected from the previous section. Finally, in Section 3.3, the process carried out for data cleaning is explained.

3.1. Survey of Open Agronomic Data Sources

The below details the main spaces where climatological, geospatial, phytosanitary, geographical, and crop data from the agri-food sector have been collected for this study.

3.1.1. Climatological

- Agroclimatic Information System for Irrigation (SIAR)

SIAR is a system operated by the Ministry of Agriculture, Fisheries, and Food of Spain (MAPA) [41]. Its primary objective is to provide quality agroclimatic data for calculating the water demand of soil zones. It comprises 468 agroclimatic stations distributed throughout the national territory, with 361 stations belonging to the Ministry and 107 owned by various autonomous communities (see Figure 3). SIAR has 12 zonal data concentrator centers associated with different autonomous communities and one national coordinating center for these local centers.

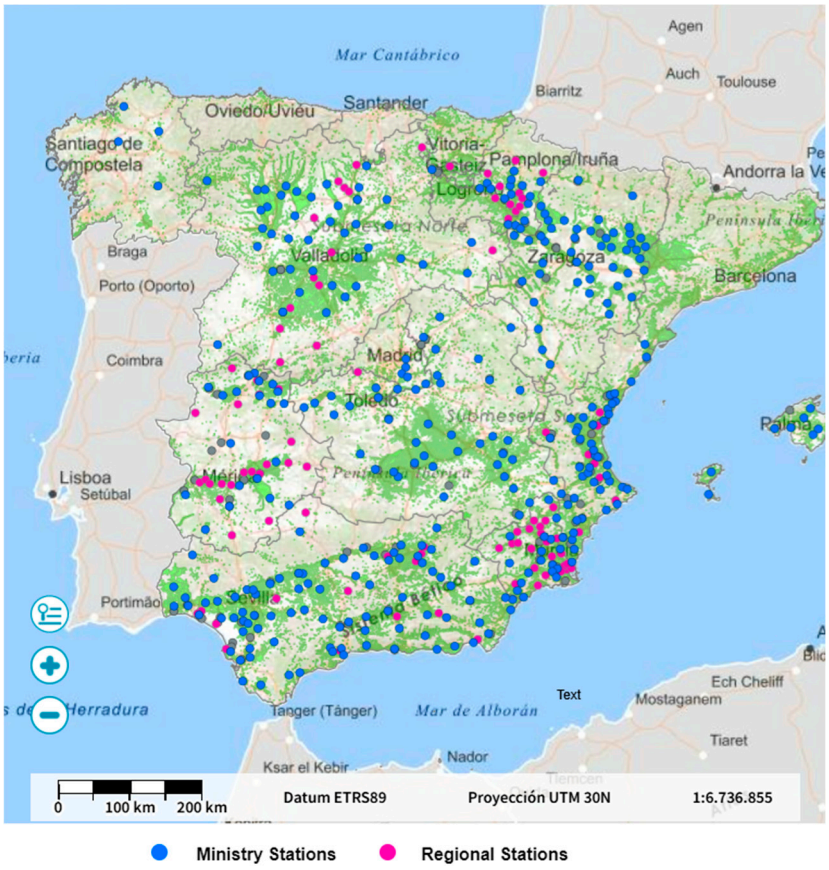


Figure 3. Distribution of meteorological stations of SIAR. Source SIAR (https://servicio.mapa.gob.es/siarweb/consultaDatos/inicio, accessed on 28 January 2026).

The data provided includes the date, day of the year, maximum and minimum temperatures (°C), times of maximum and minimum temperatures, average temperature (°C), maximum and minimum humidity levels (%), times of maximum and minimum humidity levels, average humidity (%), maximum wind speed (m/s), time of maximum wind speed, average wind speed (m/s), wind direction (°), direction of the strongest wind (°), precipitation (mm), effective precipitation (mm) (MAPA only), and Et0 (actual evapotranspiration) (mm).

The data is available through a web form openly. After querying, the data is displayed in a table and can be downloaded in CSV format. It is not necessary to register on the website to access the data. The records start from 2000, and daily updates are made. The data sampling frequency is daily, weekly, or monthly. The accessible date range is limited, requiring multiple queries for downloading large volumes of data. In addition to this access, SIAR offers advanced queries for which prior registration is required.

- Andalusian Agroclimatic Information Network (RIA)

The RIA has been operational since 1999 and comprises 121 meteorological stations distributed across Andalusia (see Figure 4). The Andalusian Institute for Agricultural, Fisheries, Food and Organic Production Research and Training (IFAPA by its acronym in Spanish) has made these data available through a web service API REST [42], participating in the open data initiative. These data are accessible with a daily or monthly frequency and are published raw. The exposed data include the battery, day, wind direction, maximum wind direction, Et0, date, last modification date, minimum humidity time, minimum/maximum temperature time, minimum/maximum temperature time, minimum/maximum speed time, maximum humidity, average humidity, minimum hu-

midity, precipitation, radiation, maximum temperature, average temperature, minimum temperature, wind speed, and maximum wind speed.

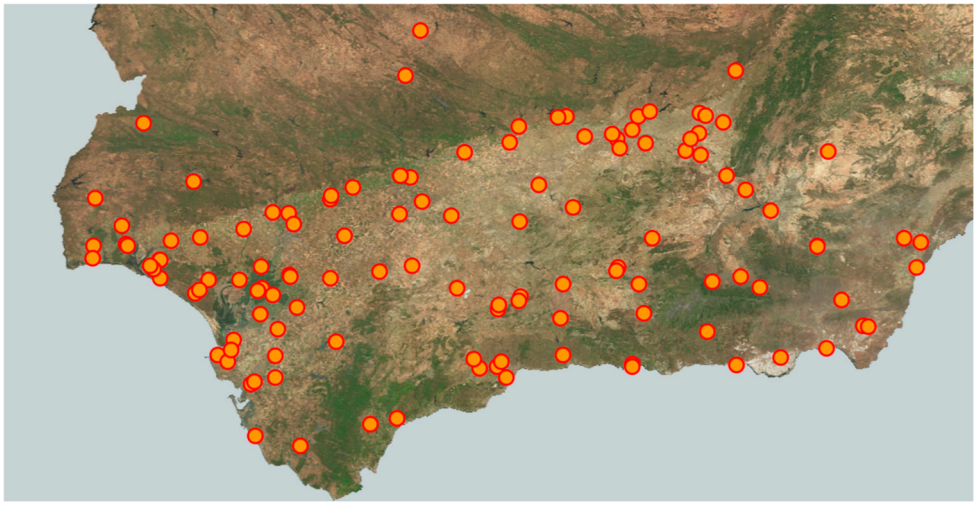


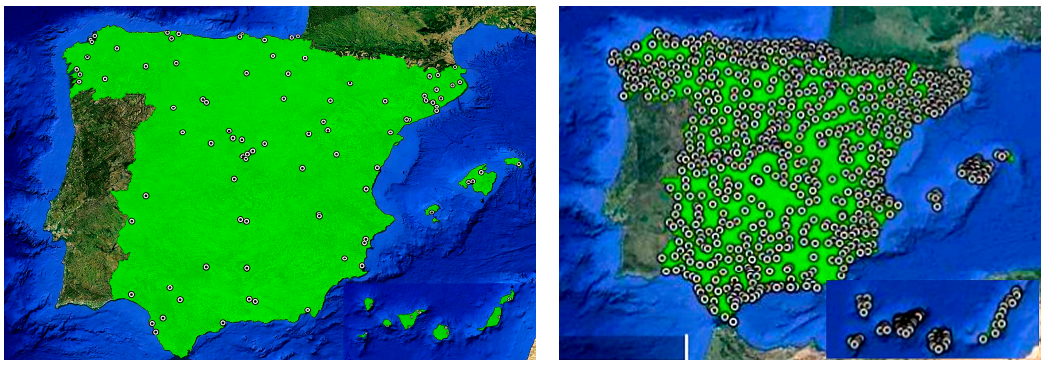
Figure 4. Distribution of meteorological stations of RIA in Andalusia. Source RIA (https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaweb/web/, accessed on 28 January 2026).

- State Meteorological Agency (AEMET)

The AEMET offers a set of climate services, among which AEMET OpenData [43] stands out. This service is a REST API through which exposed data can be downloaded for free. It has two types of access. One of them is graphical, aimed at the general public, where data is accessible on a website. The only requirement is to request the API Key for free. On the other hand, there is the AEMET OpenData API, which is more developer-oriented, providing clients in different programming languages (CURL, Java; Python, PHP, Ruby, HTTP, C, C#, Go, JS, NodeJS) to expedite access.

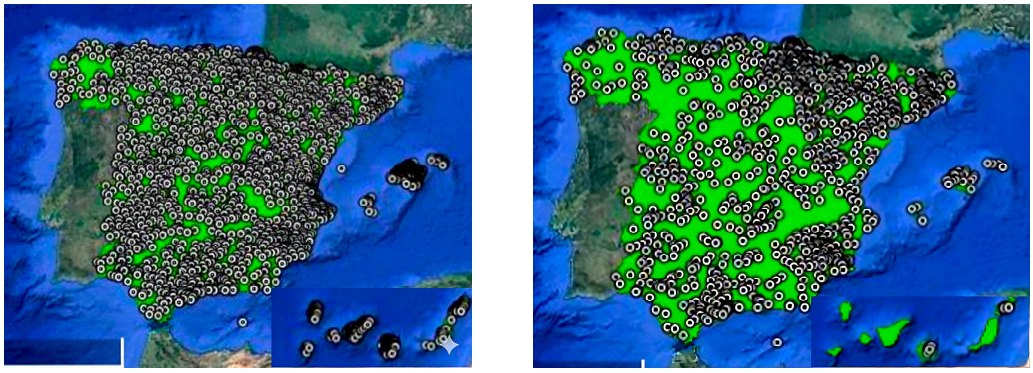
Regarding the data, AEMET has a large historical database, making available to users the following information: (i) normal climatological values from 1981 to 2010 from a set of observers, in CSV format, last updated in 2015; (ii) absolute extreme values since 1920, available in CSV format and updated daily; (iii) climatological summaries in PDF format updated monthly; (iv) hourly predictions per municipality (2 days); (v) municipality predictions (7 days).

Figure 5 shows the different types of meteorological stations or observatories distributed throughout the national territory that provides AEMET: (i) complete stations; (ii) thermometric stations; (iii) pluviometry stations; (iv) automatic stations.



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

Figure 5. *Cont.*



(**c**) (**d**)

Figure 5. Distribution of the different AEMET stations: (**a**) Complete, (**b**) Automatic, (**c**) Pluviometry, (**d**) Thermometric. Source (https://www.miteco.gob.es/es/cartografia-y-sig/ide/descargas/otros.html, accessed on 28 January 2026 © Ministerio para la Transición Ecológica (MITECO)).

The data provided by the stations included the date, identifier, name, province, altitude, average temperature, precipitation, minimum temperature, time of minimum temperature, maximum temperature, time of maximum temperature, wind direction, average wind speed, gust, time of gust, sunshine, maximum pressure, time of maximum pressure, minimum pressure, and time of minimum pressure.

3.1.2. Image Dataset

- The Copernicus Program

The Copernicus Program [44] is the European Union's Earth observation program, which provides its citizens with information services based on satellite Earth observation and in situ (non-spatial) data. The program partners include the European Space Agency (ESA), the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT), the European Centre for Medium-Range Weather Forecasts (ECMWF), EU agencies, and Mercator Océan. The services are free and open to users.

The Earth observation satellites are divided into two groups of missions: (i) The Sentinels, developed for the specific needs of the Copernicus program. Sentinel-1, -2, -3, and -6 are dedicated satellites, while Sentinel-4 and -5 are instruments onboard EUMETSAT's meteorological satellites. It is worth noting that Sentinel-5P, the precursor to Sentinel-5, is also a dedicated satellite. (ii) Contributing Missions, operated by national, European, or international organizations, provide a large amount of data for Copernicus services.

Regarding access to information, Copernicus offers various web-based services. These services allow for visualizations without registration, but registration is required to download the data. Figure 6 provides a brief description of the main access methods and services offered.

The data and information provided by the Copernicus program can be of two types: direct satellite images or services. Satellite images are accessible via the Access Hub, from platforms such as the European Space Agency (ESA) or the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT). Services can be accessed through DIAS (Data and Information Access Services) or through Copernicus's own service catalog.

The ESA platform allows access to Sentinel satellite data through two types of interfaces: an interactive graphical interface or Application Programming Interfaces (APIs). In both cases, user configuration is allowed. EUMETSAT offers more than 380 different product collections, including EUMETSAT's own satellite data, marine and atmospheric data from Copernicus, and a wide range of third-party products.

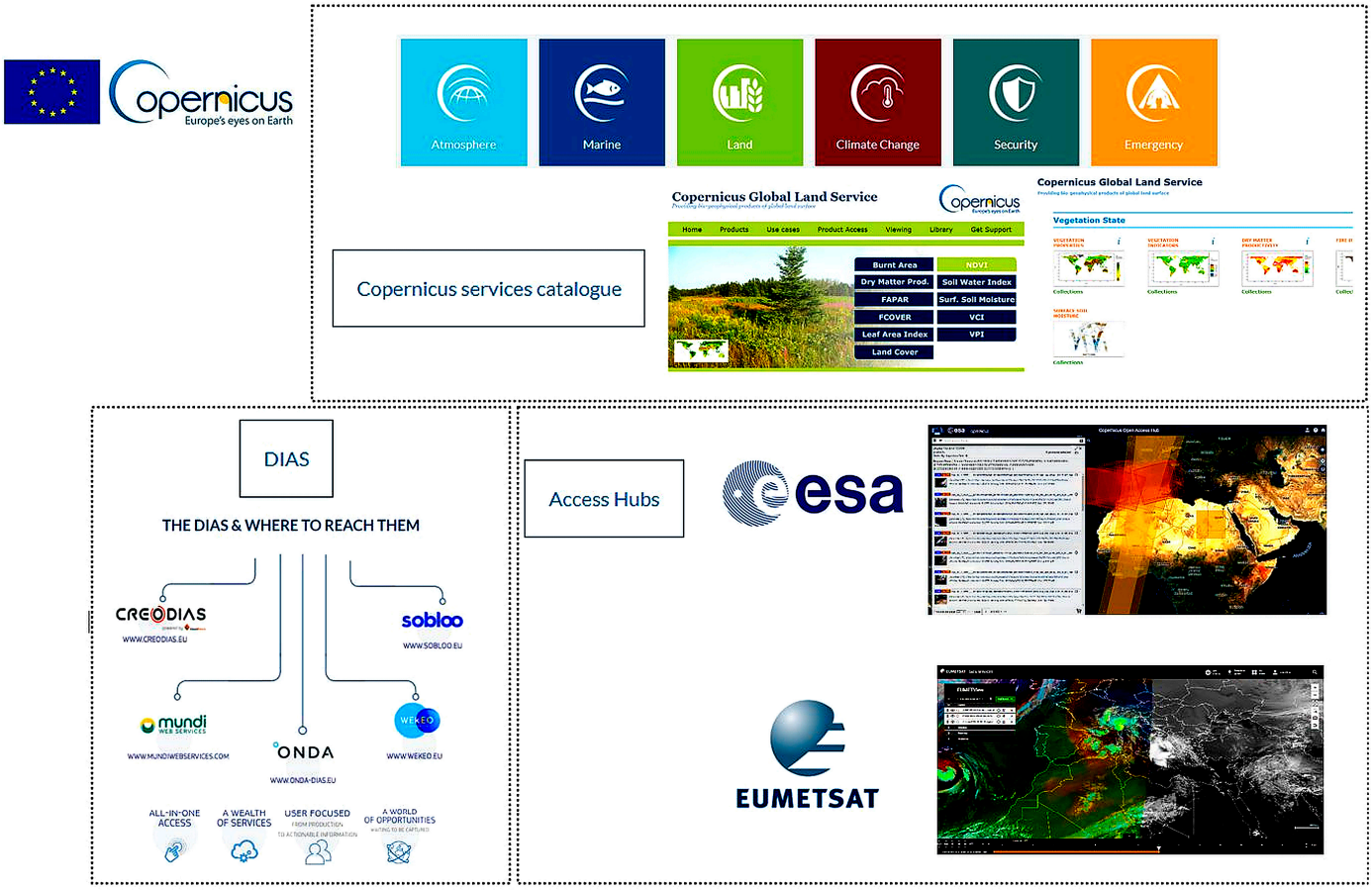


Figure 6. Access to Copernicus program data.

The platforms for accessing services are called Data and Information Access Services (DIAS). These facilitate access to data through five cloud platforms, offering a single access point to the data. This enables third parties to develop and host applications in the cloud without the need for downloads. Here, both free and pay-per-use developments can be found. Additionally, the Copernicus platform provides information through six thematic services: Atmospheric, Marine, Land, Climate Change, Security, and Emergency. All of these are freely accessible.

In this work, the terrestrial service provided by the Copernicus catalog has been delved into. The Copernicus Global Land Service (CGLS) systematically develops geophysical products on the state and evolution of the land, working in different domains (vegetation, energy, water, cryosphere, hotspots, land data). For monitoring the Earth's ecosystem, the land service provides a set of biophysical variables that describe changes in vegetation, such as burnt area, dry matter production, Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Fraction of Green Vegetation Cover (FCOVER), Leaf Area Index, Land Cover, Normalized Difference Vegetation Index (NDVI), Soil Water Index, Surf. Soil Moisture, Vegetation Condition Index (VCI), and the vegetation productivity indicator (VPI).

These variables were accessible in different ways: (i) direct downloads of images with 300 m (from sensors OLCI or PROBA) and 1 km spatial resolutions, and with a 3-day temporal resolution via the product portal (Copernicus Product Portal, 2023); (ii) regular FTP access: Allowed setting up periodic downloads. (iii) web services, including OGC CSW web services, WMT services, WMS services, and Time Series accessible via API; (iv) interactive notebooks from which APIs could be called using R or Python programming languages; (v) manifest files containing web links (URLs) to the main data files, in their native format (netCDF4 or Cloud-Optimized GeoTIFF). All these access methods

could be used together to synchronize a collection of products or time series and were freely accessible.

- National Aerial Orthophotography Plan (PNOA)

The PNOA is a cooperative project involving the Central Government and the Autonomous Communities [45]. It is part of the National Land Observation Plan (PNOT), coordinated by the National Geographic Institute (IGN) and the National Center for Geographic Information (CNIG). Its objective is to obtain digital aerial orthophotography of the entire Spanish territory with a refresh period of 3 years. The spatial resolution of the photos is available at 25 cm and 50 cm, with some areas at 15 cm. The spectral resolution consists of four sensors (blue, green, red, and near-infrared), and the radiometric resolution is 12 bits per band.

It has several viewers: (i) Iberpix [55], a cartographic viewer published by the National Center for Geographic Information (CNIG) [56]; (ii) map comparators, through which photos from 2004 onwards can be compared [57]; and (iii) the Spanish Agricultural Guarantee Fund [58]: viewer of the SIGPAC where PNOA photos are integrated. In addition to online viewers, there are web services based on WMS [59] and WMTS [60] standards.

The images are also available for download through download centers [61], where access is provided to both historical and most recent photos in various formats, such as Enhanced Compression Wavelet (ECW) or Cloud Optimized GeoTIFF (COG). These services are open, and registration in the system is not required.

3.1.3. Cloud-Based Platforms for Geospatial Analysis of Satellite Images

- Sentinel Hub

Sentinel Hub is a cloud service for satellite imagery [62], awarded by Copernicus Masters. This platform provides easy access to satellite information, handling storage, synchronization, processing, and metadata. The main indices (NDVI, False Color, SWIR, NDIS, and True Color, among others) are available, allowing the creation of new ad hoc products. All information is accessible via web services: (i) download of processed images, (ii) statistical information through time series, (iii) WMS services, and (iv) API for integration of advanced functions. To access these services, registration is required, as well as having a paid account. It offers a 1-month trial account.

- Google Earth Engine (GEE)

GEE [47] is a freely available tool that harnesses the computational power of Google's servers to perform operations on satellite imagery. The strength of this tool lies in its ability to execute geospatial analysis in near-real time. It is accessible through a web-based Integrated Development Environment (IDE) where scripts can be written and executed. Additionally, it provides libraries developed in Python and JavaScript to access the GEE API. These libraries are used by developers to integrate GEE services into local applications. Access is limited and requires registration in the system.

- Land Viewer

Land Viewer [48] is a satellite imagery platform that allows users to search, process, and obtain satellite data. It is a paid tool that offers global coverage with high-resolution images of up to 0.4 m per pixel. Despite being a paid service, it allows the free download of 10 images per day and provides time series of vegetation indices. Additionally, the tool also provides maximum, minimum, and mean values of indices for a specified time interval through time series provided via API.

3.1.4. Phytosanitary

- Phytosanitary Alert Network of Andalusia (RAIF)

RAIF is an open dataset obtained from the monitoring of pests and diseases in biological control stations [49]. The data are presented in Excel files categorized by the crop type from 2006 to 2021. Additionally, it has a viewer [63] that allows users to visually observe how crops are being affected by pests or diseases geographically. The information displayed in this viewer is updated weekly.

- Integrated Treatments in Agriculture in Andalusia (TRIANA Program)

The TRIANA Program is a software tool designed for crop management from a technical perspective. It incorporates various types of data, including climatic information, analytical data, crop operations, and pest-related information. Access to this database was obtained by submitting a request to the Andalusian Regional Government, which provided the requested information. The data covers the time interval from 2006 to 2013.

Among the provided information, the plots table includes the RAIF code, from which pests or phytosanitary information can be identified at the parcel level.

3.1.5. Geographic and Crop Information

- Geographic Information System of Agricultural Parcels (SIGPAC)

SIGPAC [64] allows for the geographical identification of parcels declared by farmers and livestock breeders, under any aid scheme related to cultivated land or land used by livestock. It is accessible through a viewer [58] or via WMS services [59].

The terrain is segmented, with the smallest unit to be dealt with being the plot (province/municipality/polygon/parcel/plot). Selecting such a sector provides relevant information about it. This information is also available through ShapeFile files provided openly by various regional authorities [31].

- Integrated Treatments in Agriculture in Andalusia (TRIANA Program)

The TRIANA Program [50], in addition to providing phytosanitary information, provides crop information. Among the relevant cultivation information, it is possible to find the associated crop, cultivated area, irrigation, nearest weather station, planting frame, main variety, main variety planting date, hectares, secondary variety, secondary variety planting date, hectares, density/row, fraction of canopy covered, and tree estimation.

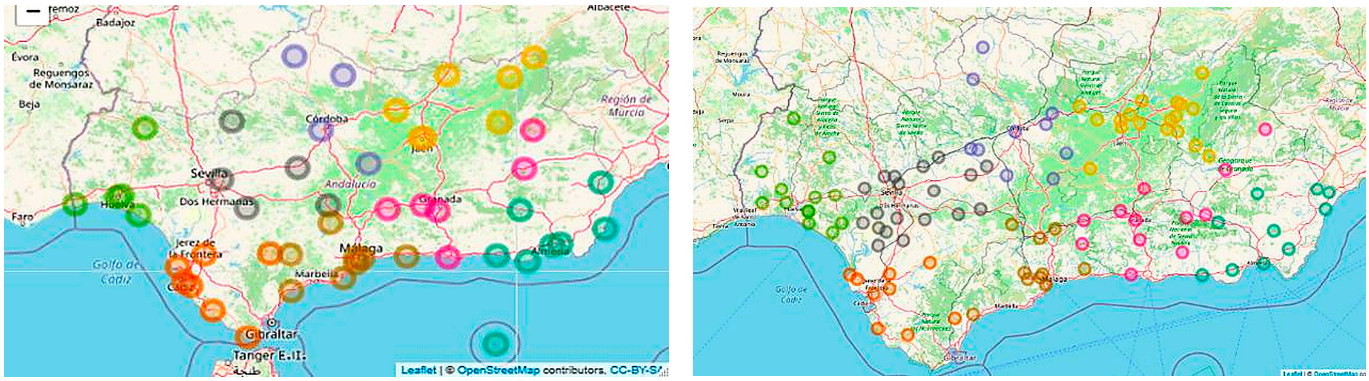
3.2. Automation of Different Processes

The following part shows the automation of the different processes associated with climatological, geospatial, phytosanitary, geographical and crop data from the agri-food sector for this study.

3.2.1. Climatological

The download and data automation have focused on the Andalusia region. The process has been divided into two stages, one where data from 1970 to 2001 from AEMET [43] was downloaded, and another with data from 2001 to 2022 obtained from RIA [42]. Integrating both data sources allows for meteorological data across the territory for the years between 1970 and 2022.

The selection criteria for the data have been based on RIA having a greater number of meteorological stations deployed in Andalusia, 121 compared to the 50 stations of AEMET (Figure 7), allowing for greater sectorization of the terrain. On the other hand, complete data from RIA is only available from 2001 onwards, so data integration from AEMET for previous years has been employed.



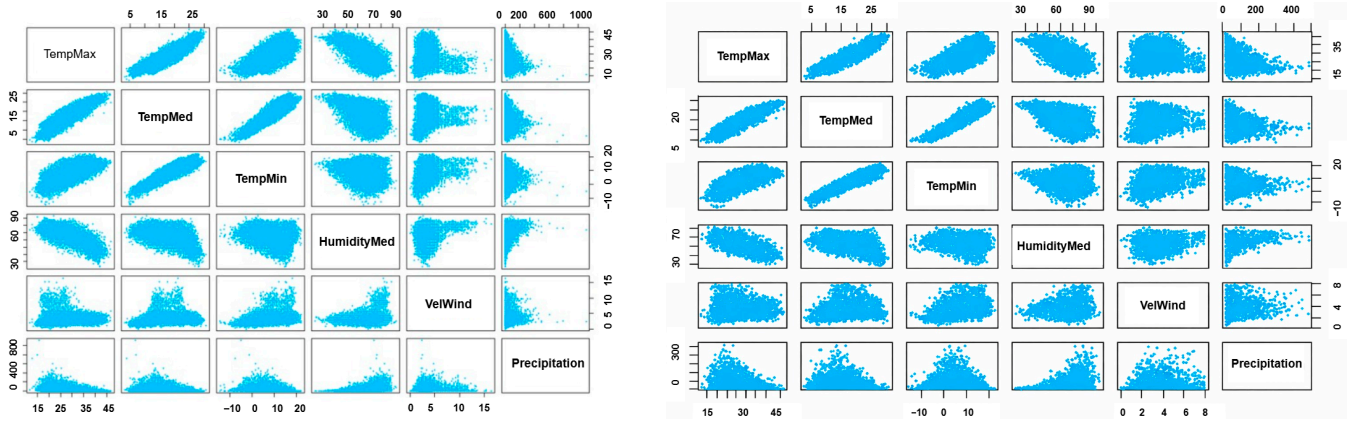
(**a**) (**b**)

Figure 7. Meteorological stations in Andalusia: (**a**) AEMET; (**b**) RIA. Different colors indicate the provinces of Andalusia.

The data obtained from AEMET includes the Average Temperature; Maximum Temperature; Minimum Temperature; Average Relative Humidity; Wind Speed; and Total Precipitation. AEMET provides much more information [65], but these are the common data with RIA. Access to this data is through an API provided by the open data source and the API Key obtained during registration. This process was only necessary to perform once, and the data was stored in the database.

The data from RIA was downloaded through the API provided by IFAPA [42]. Monthly data is downloaded and stored for subsequent preprocessing. This process is executed monthly through a scheduled task, ensuring that the climatological data remains updated within the data space.

For the data combination, a study of the distributions of the variables to be integrated was conducted. As shown in Figure 8, the distributions are similar, with some variations in the ranges.



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

Figure 8. Distributions of meteorological variables: (**a**) AEMET; (**b**) RIA.

3.2.2. Image Dataset

For the download of PNOA images, the WMS service provided by the National Geographic Institute was used [59]. This service allows the download of PNOA photos in the maximum resolution given their coordinates. These coordinates were taken from the geospatial information provided by SIGPAC through ShapeFile files [51]. In addition to geospatial information, SIGPAC provides crop information. This information has been used to generate the AOI within the image.

In this regard, the image download procedure is as follows: (i) geospatial and crop information of the area of interest is obtained from the shapefiles of SIGPAC and stored in the database; (ii) geospatial and crop information is stored at the polygon level in the database, as it is the selected territorial unit for download; (iii) the Automatic Image Acquisition module, developed for this purpose in this study, will take the geospatial information from the database and call the IGN WMS services to download the orthophotograph; and (iv) the Automatic Image Acquisition module generates two outputs: the orthophotograph of the polygon with the AOI of the crop and a shapefile with metadata (see Figure 9).

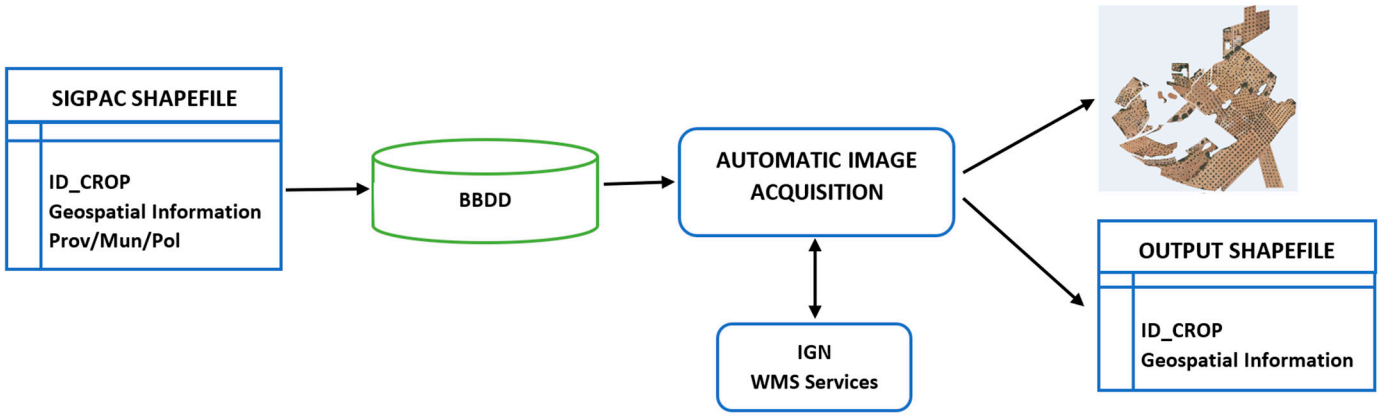
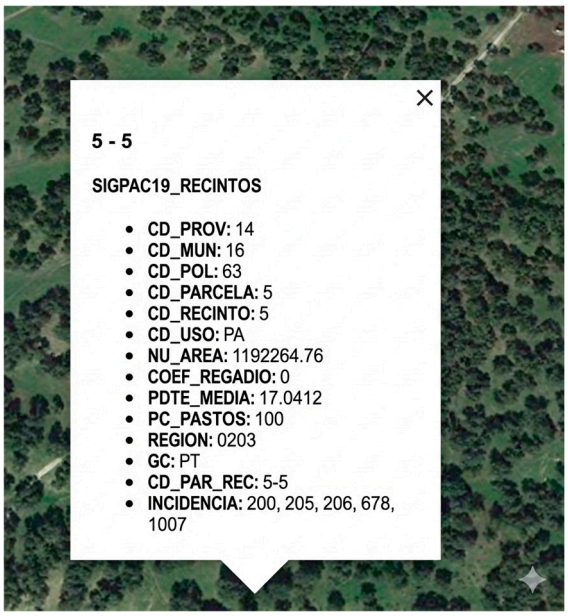


Figure 9. Automated PNOA image download process.

In the procedure, both the crop type and the geographical area of interest can be selected. The crops available for selection are those provided by SIGPAC [58], with most being perennial crops, as shown in Figure 10. On the other hand, the geographical area of this study can be selected in different ways: (i) indicating the polygons; (ii) indicating the municipality or province, in which case all the polygons of the selected municipality or province will be stored in the database.



(**a**)

FF	Asòciaciön Frutales-Frutales
CI	CITRICOS
CS	CITRICOS
CV	Asòclaciön Cïtricos-Frutales
ED	EDIFICACIONES
EP	ELEMENTO DEL PAISAJE
FI	FRUTALEES
IM	FRUTOS SECOS
FV	FRUTOS SECOS Y OLIVAR
OA	FORESTAL
OR	IMPRODUCTIVOS
IN	INVERNADEROS Y CULTIVOS BAJO PLASTICO
OA	OLIVAR -FRUTAL
OR	OLIVAR
PA	PASTO CON ARBOLADO
PB	PASTO ARBUSTIVO
PS	PASTIZAL
TA	TIERRAS ARABIES
VF	HUERTA
NC	VINEDO - OLIVAR
VI	VINEDO
ZC	VINEDO -OLIVAR
ZC	ZOMA CONCENTRADA NO INCLUIDA

(**b**)

Figure 10. (**a**) Metadata from SIGPAC ShapeFile files; (**b**) crop codes.

3.2.3. Cloud-Based Platforms for Geospatial Analysis of Satellite Images

The platform selected for automating the information from satellite images was Google Earth Engine (see Figure 11), due to its easy and free access. Land Viewer and Sentinel Hub limit free access to the number of images and/or days. Copernicus also exposes its information through an API [50], which is freely accessible. From it, time series of different indices such as NDVI, FCOVER, Solid Water index, and Surt Soil Moisture can be obtained. Its API is also accessible from Swagger UI and notebook [66], which simplifies its handling. However, the spatial resolutions with which the API operates are between 300 m and 1 km, which is why its automation has been discarded.

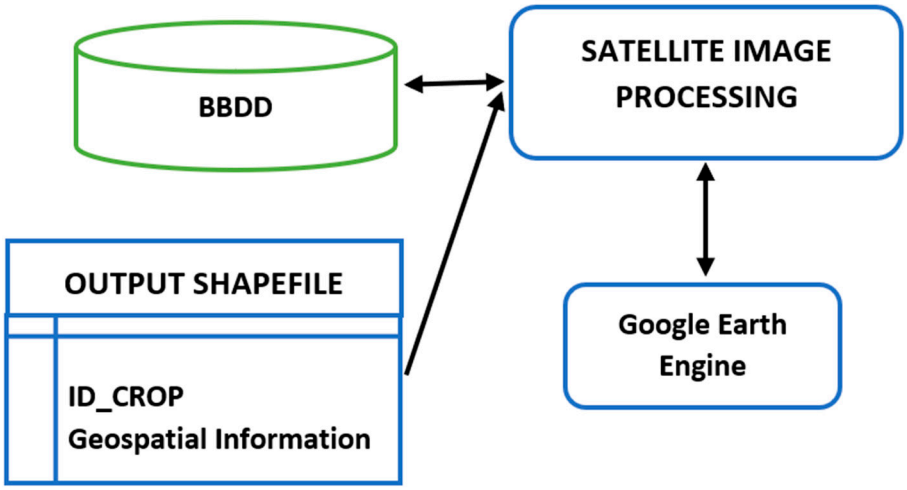


Figure 11. Process of obtaining time series of vegetation indices through Google Earth Engine.

To address the automation process of GEE, a script was developed locally. It retrieves information from the area of interest from the database and ShapeFile files generated by the Automatic Image Acquisition module. With this information, it makes a call to GEE through the Python API, obtaining time series of different indices (NDVI; Stem Area Index (SAI); Leaf Area Index (LAI)) for the Area of Interest (AOI) indicated in the ShapeFile. The obtained information is stored in the database.

Multiple indices can be calculated, as well as the geographical area of operation through these platforms. In order to narrow down the development and provide an example of use, the NDVI was obtained for olive groves in Andalusia. However, this solution can be extrapolated to any index and geographical area. The procedure is repeated periodically, through a scheduled task, with the aim of keeping the Data Space updated.

3.2.4. Phytosanitary

Phytosanitary information was obtained from RAIF [49] and the TRIANA program's database [50]. RAIF provides two types of files: one with geospatial information and another with sampling data of pests detected in the field. The files contain information from 2006 to 2021, are provided by type of crop, and are updated quarterly. To automate this process, a preliminary study is required to select the pests to be incorporated into the Data Space. Afterward, the polygon is identified using geospatial data, and the pests from the sampling data are associated with it. All this information is stored in the database and updated quarterly.

The TRIANA program's database provides additional sampling data collected in the field. To unify them, the RAIF code provided by the TRIANA database is used, as can be seen in Figure 12.

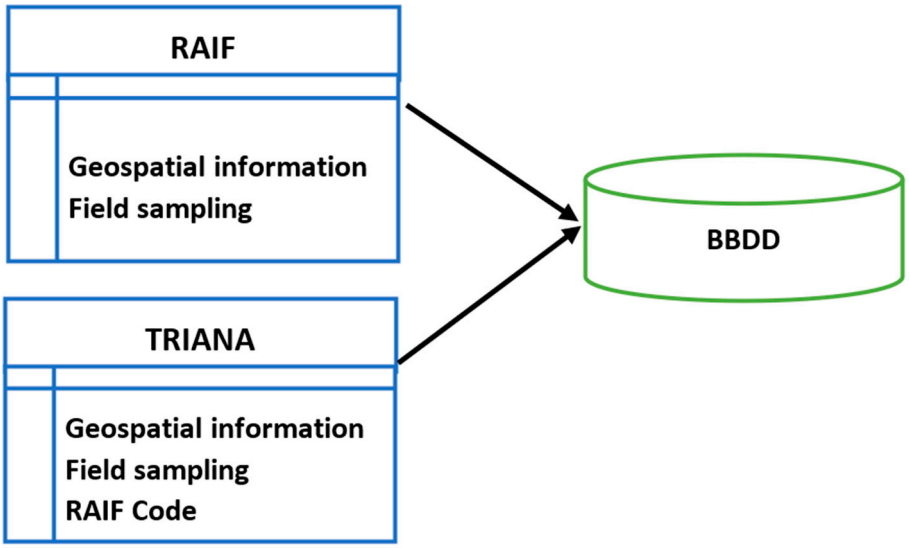


Figure 12. Phytosanitary data acquisition process.

3.2.5. Geographic and Crop Information

Geographic and crop information was obtained from the ShapeFile files of the SIGPAC [50] and the TRIANA program [50]. The data provided by SIGPAC are more related to geospatial information, providing some crop data, but at a high level. These files are updated every year and have been available on the web since 2017. TRIANA provides more detailed information at the crop level. To unify both sources of information, the UTMX and UTMY coordinates of the centroid of the parcel provided by TRIANA were used, identifying with this point the parcel of SIGPAC to which it refers, since the parcel codes from different data sources are different (see Figure 13).

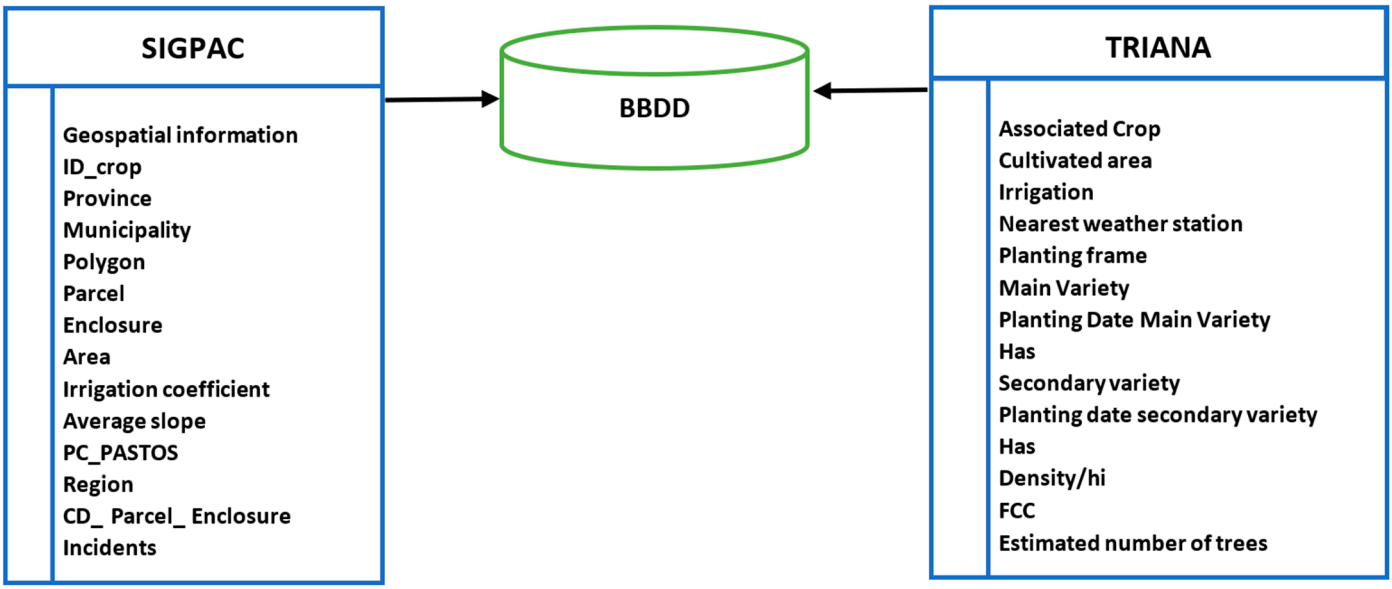


Figure 13. Geographic and crop information included in the data space.

3.3. Data Cleaning

Data preprocessing and cleaning have been performed on meteorological data and the time-series data obtained from satellites. The remaining data come from preprocessed databases.

3.3.1. Climatological

Data from Andalusia has been obtained for the period between 1970 and 2024 from different data sources: AEMET [43] for the period from 1970 to 2001, and RIA [42] for the period between 2001 and 2024. Although RIA indicates that it has data available since 1999, the number of operational stations for the year 2000 is considerably lower than for the rest of the years.

In Figure 14, it can be observed that in the first half of 2000, there is no data available for all provinces, with a presence in all provinces by the end of the year. The graph for the year 2001 reflects a possible failure in the data retrieval from the API, as it does not make sense for so many stations to become inoperative. To verify this, queries were made during that period of time to the supposedly non-operational stations, obtaining data from them. This demonstrated the failure in the API call that provides the number of stations for that period.

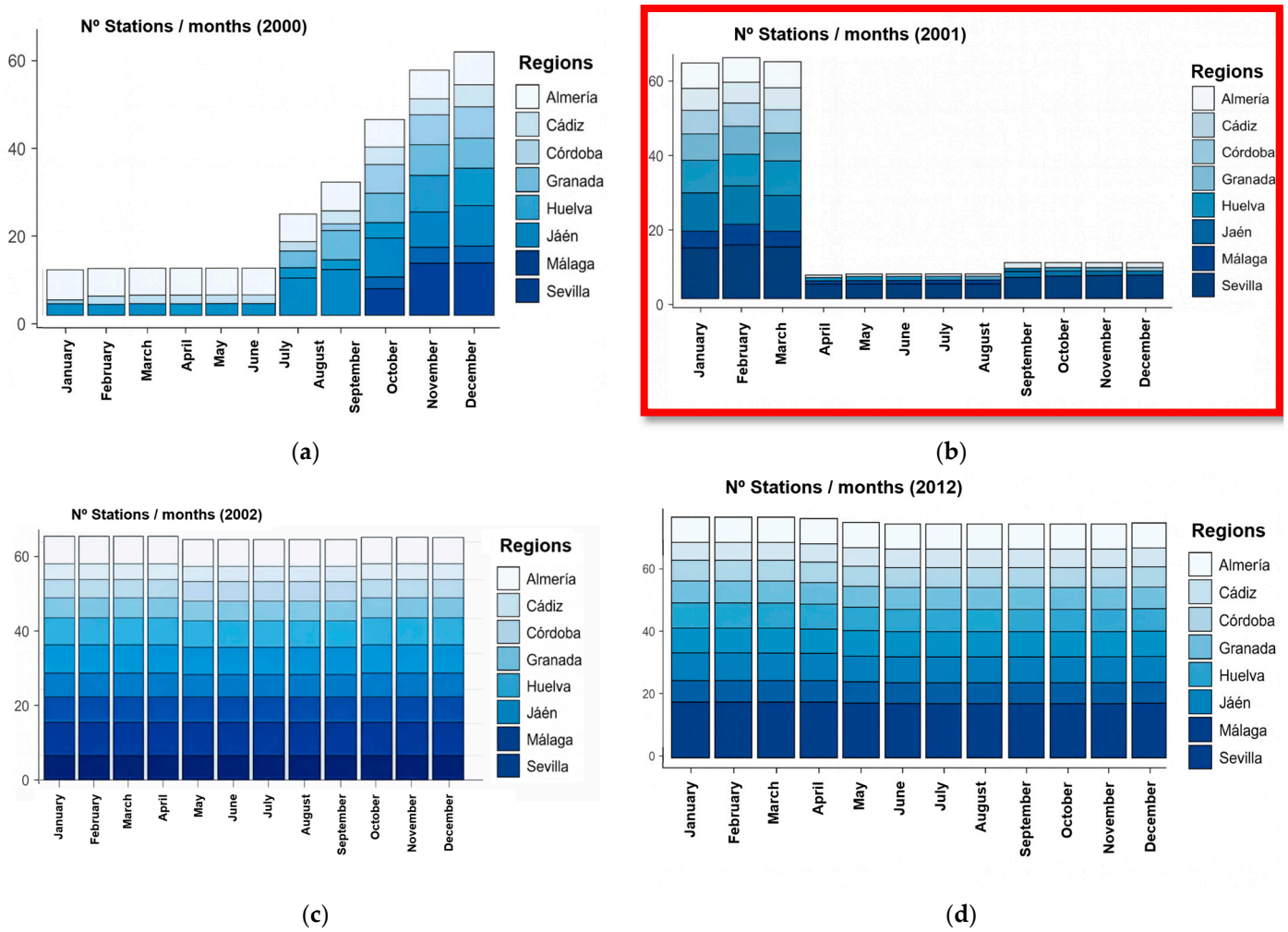


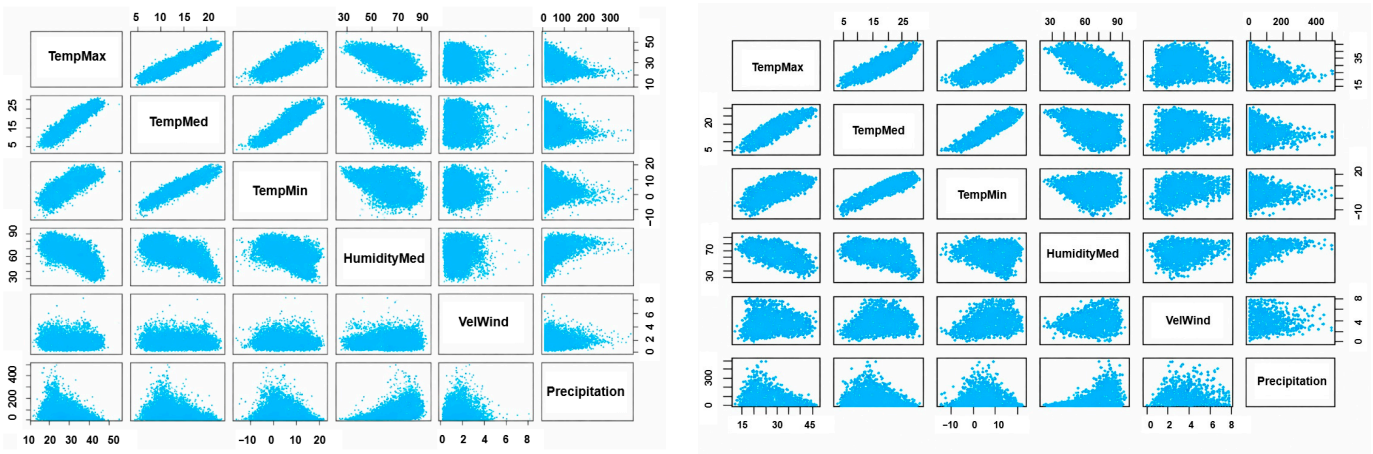
Figure 14. Number of meteorological stations from RIA in Andalusia for the years (**a**) 2000; (**b**) 2001; (**c**) 2002; and (**d**) 2003.

The cleaning and preprocessing were divided into two stages: cleaning and preprocessing of each data source, and their integration. During the preprocessing of the AEMET data, it was observed that there were a large number of missing values for years before 2000, making it unfeasible to use techniques for imputing missing values. To address this, instances from the database with mostly NULL values were removed (2068 out

of 10,248 instances were removed). After this cleaning process, it was possible to apply missing value imputation techniques, specifically missForest.

Regarding the data from the RIA, being a more recent data source, it did not present as many null values. Cleaning was done by province, starting with the detection of false values (NA) for the different meteorological variables. For most variables, a value of 0 corresponds to a false value. These values, along with other outliers or illogical values (e.g., wind speed exceeding 40 m/s), were imputed as NA for further treatment. The second step involved imputing these values using the instances available for the rest of the stations by means of the median of the distributions. A third step was to use Random Forest to detect possible outliers not detected in this initial manual search. The outForest library allows this pipeline to be performed, returning a ranking with the indices of the values furthest from the value predicted by Random Forest, as well as the predicted value in each case. Outliers are not removed; it depends on their nature. Those values that are outliers because they are false are identified, not because they are values found in the tails of the distributions. The values higher in the ranking returned by the algorithm are more likely to be false values as they are further away in the distributions than the actual outliers. After checking the nature of the outliers returned by the algorithm, those of interest were replaced with the predicted values.

After preprocessing both data sources, this study moved to their integration, where only the common variables were merged. For this purpose, a study of the distributions of these variables was conducted through pairs of scatter plots (Figure 15). In the analysis of these graphs, it was observed that the temperature distributions had similar trends, while differences were mainly observed in the ranges of precipitation and wind speed. In the RIA database, the maximum values are around 8 m/s, while in the AEMET database, they are 15 m/s. After delving into the study, it was found that all AEMET stations had average values of 8 m/s, except for station 6001 in Cádiz, located in Tarifa, so the values are correct. Regarding precipitation, a similar behavior was observed. The range in the AEMET data was increased by the Sierra de Grazalema weather station, where average precipitation is higher than at the other stations. After this analysis, it was decided to remove these stations from the AEMET data because, although they are real values, they distort the average values of the regions. Finally, a study of the distributions of the different variable was conducted again, where it can be seen that the variables had the same distributions and ranges.



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

Figure 15. Distributions of meteorological variables after removing the Tarifa and Sierra de Grazalema stations from AEMET: (**a**) AEMET; (**b**) RIA.

For the automation of the process, a scheduled task was created to download new data from RIA every month, preprocess it, clean it, and store it in the database.

3.3.2. Satellite Platforms

As a use case, time series of NDVI were obtained for olive grove polygons in the Andalusia region from 2017 to 2024. The data has a temporal frequency of 5 days, but only those with satellite images covering less than 10% cloudiness were stored.

It was clearly observed that there were values outside the expected distribution (Figure 16), which were due to unreal data caused by cloud cover. Therefore, the first step in cleaning and preprocessing was to remove instances that were outside the range of the sinusoidal distribution. For the polygon shown in Figure 16, this range would be around 0.3 and 0.6. Therefore, initially, this first step involved setting a minimum value below which values are imputed as NA (NULL). The issue is that for each polygon, this minimum value, close to the minimum value of the sinusoidal distribution, is different. To perform this first step, the quantiles of the NDVI distributions for each polygon were obtained, and the minimum threshold value was set as 0.9 of the beginning of the second quantile. To clarify the process, let's consider the example of polygon 1–11. The NDVI distribution for this polygon can be seen in Table 2.

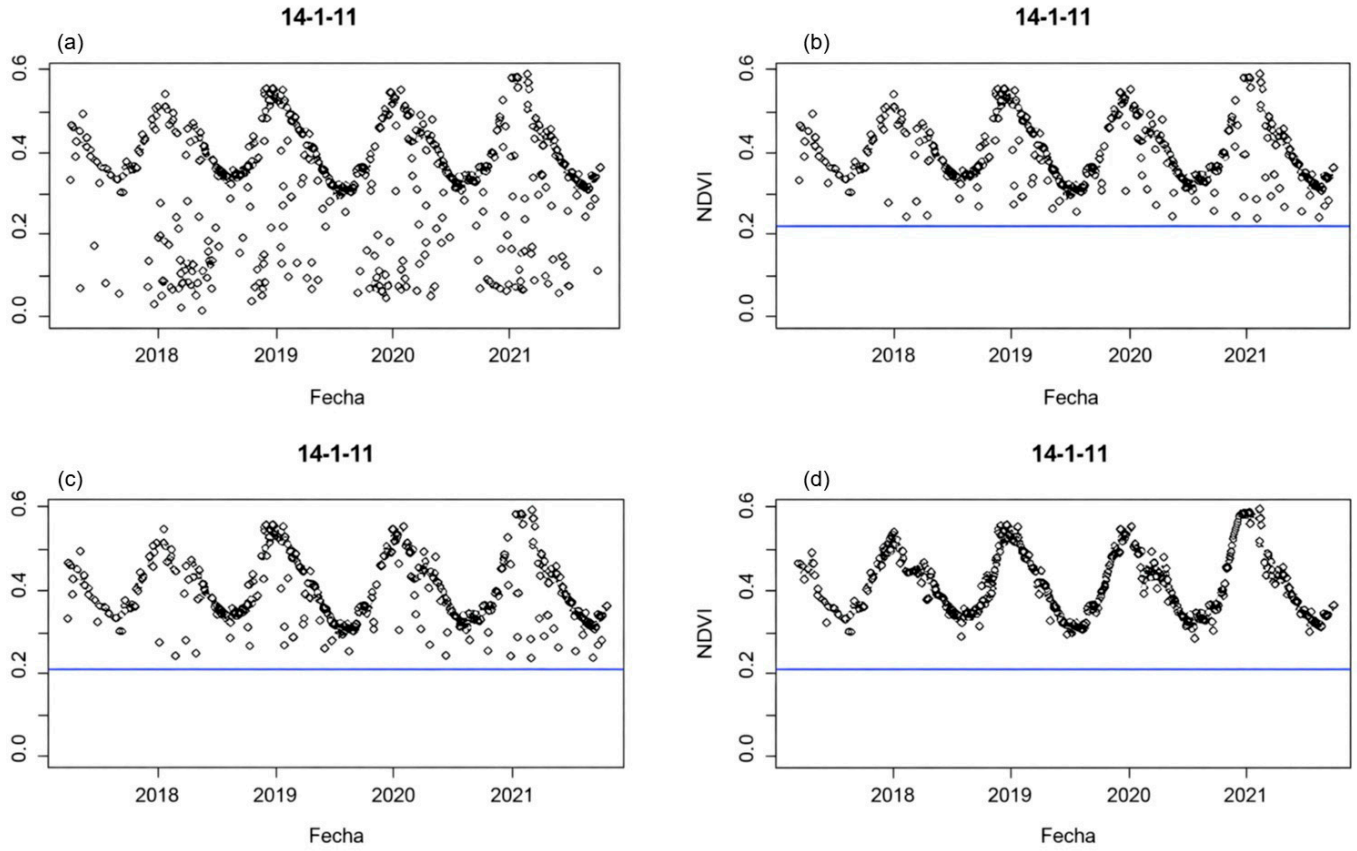


Figure 16. Time series of NDVI obtained from Sentinel 2 through the GEE platform: (**a**) raw data; (**b**) removal of instances outside the sinusoidal distribution interval; (**c**) elimination of outliers; (**d**) interpolation of false values. he blue line represents the minimum threshold.

Table 2. Distribution of NDVI for polygon 1–11.

0%	25%	50%	75%	100%
−0.05881548	0.22989626	0.32019703	0.40760768	0.59691647

Therefore, for this polygon, the threshold minimum value would be 0.9 × 0.22989626 = 0.2069066. Values below this threshold are considered false. Multiplying by 0.9 is done to increase the margin between the real minimums of the distribution and this threshold minimum.

If an instance has an NDVI value that varies by more than 0.06 from its neighbors, it is considered a false value.

The third and final step consisted of imputing the false values specified in the previous two steps. For this purpose, linear interpolation between the previous and subsequent points of each Not Available (NA) was used. This imputation of missing values was performed using the imputeTS package. This procedure is executed monthly, after downloading the new NDVI data for that month.

3.3.3. Quantitative Assessment of Data Cleaning and Imputation

To complement the qualitative description of the data cleaning process, a quantitative assessment was conducted to evaluate the impact of outlier detection and missing value imputation on the main datasets included in the agronomic data space.

For climatological data, the proportion of detected anomalous values varied across variables and stations. On average, between 2.1% and 4.8% of the records per variable were flagged as potential outliers by the Random Forest-based detection algorithm. After domain-based validation, approximately 60–70% of these flagged instances were identified as false or inconsistent values and subsequently corrected using model-predicted values, while the remaining observations were retained as plausible extreme events.

Regarding missing value imputation, a validation procedure was performed by artificially removing a subset of observed values and comparing the imputed results with the original data. For multivariate climatological datasets imputed using Random Forest–based methods, the root mean square error (RMSE) ranged between 0.6 °C and 1.2 °C for temperature variables and between 0.8 mm and 1.5 mm for precipitation, depending on the station and temporal aggregation. These values are consistent with those reported in the literature for similar datasets and confirm the suitability of the adopted imputation approach. For vegetation index time series, short temporal gaps accounted for less than 5% of the observations per polygon. Linear interpolation yielded smooth and continuous NDVI profiles, preserving the expected seasonal patterns. After cleaning and imputation, the proportion of valid observations increased to over 98% across all analyzed polygons, enabling robust downstream analysis (see Table 3).

Table 3. Quantitative indicators of outlier detection and missing value imputation for the main datasets integrated into the agronomic data space.

Dataset	Metric	Value (Range)
Climatological data	Detected outliers	2.1–4.8%
Climatological data	Corrected false outliers	~60–70%
Climatological data	RMSE (temperature)	0.6–1.2 °C
Climatological data	RMSE (precipitation)	0.8–1.5 mm
NDVI time series	Missing values before cleaning	<5%
NDVI time series	Valid observations after cleaning	>98%

As an illustrative application of the proposed agronomic data space, the developed workflow was applied to olive groves in the Andalusia region, selected due to the high socio-economic relevance of this crop. In this context, heterogeneous datasets were integrated, including long-term historical meteorological records, geospatial parcel information,

high-resolution aerial imagery from the National Aerial Orthophotography Plan (PNOA), phytosanitary information on olive varieties and pests obtained from the TRIANA database, and time series of NDVI values segmented by olive grove polygons. The integration and preprocessing of these datasets enabled a consistent territorial-level characterization of olive groves and supported continuous monitoring of crop conditions. This example demonstrates the practical applicability of the proposed data space as a reusable and extensible framework, while more detailed analyses specific to olive cultivation are reported in a complementary study [16].

4. Discussion

This research highlights the importance of creating shared data spaces from open data sources, as well as the need to reuse data. In this sense, there are European initiatives where reference architectures that address these agricultural challenges are being proposed [1,13,20,31], facing different problems, including the lack of interoperability [4]. This study contributes to the development of algorithms, specifically process automation, metadata acquisition, processing and management of data, automatically, with the objective of adding value to such data, its reuse and applicability to specific cases of real needs, the need for which is highlighted in the research [2,39,67]. In that sense, this study exposes different methods and algorithm for automation, data cleaning and imputation of missing values, allowing a unified access to reliable and contrasted agronomic data, enabling it to be reused for a variety of purposes, aligning with the European Directive 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and reuse of public sector information [1].

On the other hand, in this study, a comprehensive analysis of the state of the art of various open data sources in the agronomic sector has been carried out. This analysis has facilitated a broader and more detailed understanding of these sources, allowing them to be compared and explored in depth, studying aspects as relevant as reliability, one of the main barriers to their deployment according to studies such as [68,69]. The results obtained provide an overview of the capabilities and limitations of the available data sources, which contributes to the knowledge and development of future research in the agronomic field and provides tools for the improvement of one of the main weaknesses detected by the studies [69–72], namely the lack of interoperability.

The study [73], which examines the dynamics of data spaces in the agricultural sector through scientific publications from 2001 to 2023, identifies a remarkable shift towards the practical application of research, underlining the growing importance of technology in the modernization of agriculture, but finds that there are still major barriers to the application of these technologies in all regions of the world. This study provides some tools for re-using the data, testing its reliability and providing the technological know-how to access it, thereby facilitating its interpretation and reducing the current barriers faced by small and medium-sized farmers.

Although the proof of concept presented in this study has been implemented for a specific region, its underlying methodology has been designed to be largely independent of the geographical context. The proposed agronomic data space is based on a modular architecture and standardized processing workflows that can be adapted to other regions depending on data availability and institutional settings. At the national level, several components of the data space are directly transferable across Spain. In particular, climatological data processing pipelines can be applied to other Spanish regions using national or regional meteorological networks with similar structures. In addition, agronomic datasets obtained from national information systems such as TRIANA, SIGPAC, or RAIF are available across the country, enabling the replication of crop- and plot-level analyses beyond the regional

case study. High-resolution aerial orthophotography from the Spanish National Plan for Aerial Orthophotography (PNOA) further supports nationwide scalability by providing homogeneous spatial coverage.

Beyond the national context, the proposed approach is also scalable to international and global settings. Satellite-based remote-sensing products and vegetation indices used in this study rely on global Earth observation infrastructures, such as the Copernicus Programme and platforms like Sentinel Hub, which provide worldwide coverage. As a result, the satellite data acquisition and processing pipelines can be directly extrapolated to other regions in Europe and to agricultural systems worldwide. While the algorithmic workflows for data acquisition, cleaning, validation, and integration are largely agnostic to local climatic conditions and crop types, differences in institutional data governance, data availability, and agricultural practices must be considered when transferring the approach to other contexts. Explicitly addressing these factors enables the proposed agronomic data space to be adapted to diverse regions while preserving its core principles of interoperability, robustness, and reproducibility.

However, there are certain limitations to the work carried out in terms of architecture, as this is not its objective. This research aims to expose the multitude of existing open data sources, in which gaps in standardization have been evidenced, highlighting the need for adaptation to the new unified European and global data markets still in the definition and development phase. This adaptation is key for precision agriculture and will allow it to benefit from digitization to make improvements in sustainability, to achieve the objectives of the European Green Pact [6]. A more detailed study of these architectures can be found in [39,74].

This work is based on the four fundamental pillars, Data, Governance, Technology, and Participants, outlined in the study [21]. As in the previously mentioned study, ensuring the interoperability of data from different systems while maintaining data quality remains a key challenge for investigation. This study achieves improved data access through the development of standardized algorithms specifically designed for this purpose. In terms of governance studies, study [75] shows that different governance approaches (e.g., hierarchical, market-based, collaborative, and their hybrids) influence site and farm conditions, which in turn influence agricultural production practices.

Building on this framework, the agronomic data space proposed in this study primarily relies on institutionally governed open data sources, reflecting a hierarchical governance model that supports standardization, data reliability, and regulatory compliance. At the same time, the modular and interoperable design of the proposed architecture allows for its evolution towards more collaborative or hybrid governance schemes, in which additional stakeholders may progressively contribute data under shared rules and responsibilities. This flexibility is essential to ensure the long-term sustainability and scalability of agronomic data spaces across different institutional contexts.

Despite their high value, open agronomic data sources are subject to potential biases and limitations, including heterogeneous spatial and temporal coverage. For example, agroclimatic monitoring networks such as the Agroclimatic Information Network of Andalusia (RIA) may present regional coverage gaps that can affect data representativeness. From an ethical perspective, although the proposed data space exclusively processes non-personal and publicly available data, equitable access to data-driven tools remains a relevant concern, as uneven data availability may reinforce existing disparities across regions or crops. Explicitly acknowledging these aspects strengthens the responsible deployment and transferability of the proposed approach.

Finally, open data spaces are conceived as a tool to support agricultural research by providing reliable information sources and processed data that enhance interpretability.

This enables third-party applications to focus on developing new decision support tools in specific contexts, such as olive groves [16] or Dehesa [76], using data generated by the algorithms developed in this study. These studies, based on the developed data space, demonstrate the scalability and utility of the various data sources analyzed, as well as the effectiveness of the algorithms proposed in this work.

5. Conclusions

To achieve the objective of this research, a prototype of a shared data space was developed through the study, integration, preprocessing, cleaning and interpretation of different open data sources, from which data warehouses were created, which will serve as support for decision-making tools. To this end, this work has focused on the study of open data sources, and their interpretation and standardization for new data models. As a result, configurable algorithms have been developed that extract key agronomic information for different farming expectations, including (i) preprocessed meteorological information; (ii) crop and phytosanitary information; (iii) access to high-resolution aerial photogrammetry; (iv) access to remote-sensing images; (v) pest information; and (vi) time series of the main vegetation indices.

The results of this study will be integrated into the CITRIDATA project, which aims to create a federated data space and design models and services for the citrus value chain in Andalusia. This project is funded by the Ministry for Digital Transformation and Public Administration as part of the Sectoral Data Spaces Promotion Plan, which seeks to foster innovation, competitiveness, and secure data exchange across sectors. Future work includes analyzing common data access interfaces proposed by European reference architectures and adapting developments to align with the European data strategy, facilitating their integration into unified data spaces. These actions will be carried out through the CITRIDATA project and the strategic Data Spaces for the Agri-Food Sector initiative, within the framework of the ENIA International Chair awarded to the University of Córdoba.

Author Contributions: Conceptualization, C.M.-R. and R.L.-B.; methodology, C.M.-R. and R.L.-B.; software, C.M.-R.; validation, C.M.-R.; formal analysis, C.M.-R.; investigation, C.M.-R. and R.L.-B.; resources, C.M.-R. and R.L.-B.; writing—original draft preparation, C.M.-R. and R.L.-B.; writing—review and editing, C.M.-R., R.L.-B. and A.P.-G.; supervision, C.M.-R., R.L.-B. and A.P.-G.; funding acquisition, C.M.-R. All authors have read and agreed to the published version of the manuscript.

Funding: Smart-o-live: Agriculture, milling and consumption, smart sustainable and healthier olive oils in the new, healthier olive oils in the new agribusiness of the future" (REF.: EXP—00145726/MIG-20211025). This call is included among the actions foreseen in the Spanish National Recovery, Transformation and Resilience Plan, which will receive funding from the "Next Generation EU" funds, including the Recovery and Resilience Mechanism.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Some climatological data, from TRIANA, GEE (specifically obtaining NDVI), and downloading of PNOA images, processed and unprocessed in the repository (https://bitbucket.org/dataspace_cmr/data/src/main/, accessed on 28 January 2026), have been presented. Due to the amount of data obtained in this study, it is not possible to show them all; for additional data, please consult the authors.

Acknowledgments: The authors acknowledge the International Chair ENIA in Agriculture—University of Córdoba (TSI-100921-2023-3), within the framework of the European Recovery, Transformation and Resilience Plan, funded by the European Union—Next Generation EU (Ministry for Digital Transformation and Public Administration).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Diario Oficial de la Unión Europea Directiva (UE) 2019/1024 Del Parlamento Europeo Y Del Consejo Relativa a Los Datos Abiertos y La Reutilización de La Información Del Sector Público. 2019. Available online: https://www.boe.es/doue/2019/172/L00056-00083.pdf (accessed on 6 February 2026).
2. AGRIDATASPACE: Restreinte Building a European Framework for the Secure and Trusted Data Space for Agriculture D6.3 Final Report WP6: Project Management and Coordination [Internet]. Available online: https://agridataspace-csa.eu/ (accessed on 9 April 2022).
3. 175 Zettabytes By 2025 [Internet]. Available online: https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/ (accessed on 9 January 2025).
4. Amjad, A.; Azam, F.; Anwar, M.W.; Haider Butt, W. A Systematic Review on the Data Interoperability of Application Layer Protocols in Industrial IoT. *IEEE Access* **2021**, *9*, 96528–96545. [CrossRef]
5. Comisión Europea. Objetivos de Desarrollo Sostenible [Internet]. Available online: https://ec.europa.eu/info/strategy/international-strategies/sustainable-development-goals_es (accessed on 9 April 2022).
6. Consejo Europeo Pacto Verde Europeo—Consilium [Internet]. Available online: https://www.consilium.europa.eu/es/policies/green-deal/ (accessed on 9 April 2022).
7. A European Strategy for Data. Brussels, 2020. Available online: https://digital-strategy.ec.europa.eu/en/policies/strategy-data (accessed on 9 January 2025).
8. Comisión Europea: Espacios Comunes Europeos de Datos Para La Agricultura y La Movilidad | Configurar El Futuro Digital de Europa [Internet]. Available online: https://digital-strategy.ec.europa.eu/es/library/common-european-data-spaces-agriculture-and-mobility (accessed on 9 June 2024).
9. European Data: Use Case [Internet]. Available online: https://data.europa.eu/es/impact-studies/use-cases (accessed on 8 April 2022).
10. Digital Europe Programme (DIGITAL) Call for Proposals Preparatory Actions for Data Spaces-Preparatory Actions for Data Spaces HISTORY OF CHANGES. 2021. Available online: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/digital/wp-call/2021/call-fiche_digital-2021-prepacts-ds-01_en.pdf (accessed on 9 January 2025).
11. H2020 CYBELE Project: Fostering Precision Agriculture and Livestock Farming through Secure Access to Large-Scale Hpc-Enabled Virtual Industrial Experimentation Environment Empowering Scalable Big Data Analytics [Internet]. Available online: https://cordis.europa.eu/project/id/825355/es (accessed on 9 April 2022).
12. Perakis, K.; Lampathaki, F.; Nikas, K.; Georgiou, Y.; Marko, O.; Maselyne, J. CYBELE–Fostering Precision Agriculture & Livestock Farming through Secure Access to Large-Scale HPC Enabled Virtual Industrial Experimentation Environments Fostering Scalable Big Data Analytics. *Comput. Netw.* **2020**, *168*, 107035. [CrossRef]
13. H2020 ATLAS Project: Agricultural Interoperability and Analysis System [Internet]. Available online: https://cordis.europa.eu/project/id/857125/es (accessed on 9 April 2022).
14. Roussaki, I.; Doolin, K.; Skarmeta, A.; Routis, G.; Lopez-Morales, J.A.; Claffey, E.; Mora, M.; Martinez, J.A. Building an Interoperable Space for Smart Agriculture. *Digit. Commun. Netw.* **2022**, *9*, 183–193. [CrossRef]
15. Leal, A.C. Data Analytics in Agriculture. In *Digital Agriculture*; Springer: Cham, Switzerland, 2024; pp. 519–539. [CrossRef]
16. Martínez-Ruedas, C.; Guerrero-Ginel, J.E.; Fernández-Ahumada, E. Methodology for the Automatic Inventory of Olive Groves at the Plot and Polygon Level. *Agronomy* **2022**, *12*, 1735. [CrossRef]
17. Martinez-Ruedas, C.; Flores-Arias, J.-M.; Moreno-Garcia, I.M.; Linan-Reyes, M.; Bellido-Outeiriño, F.J. A Cyber–Physical System Based on Digital Twin and 3D SCADA for Real-Time Monitoring of Olive Oil Mills. *Technologies* **2024**, *12*, 60. [CrossRef]
18. Duan, Y.; Edwards, J.S.; Dwivedi, Y.K. Artificial Intelligence for Decision Making in the Era of Big Data–Evolution, Challenges and Research Agenda. *Int. J. Inf. Manag.* **2019**, *48*, 63–71. [CrossRef]
19. Piot-Lepetit, I. Digitainability and Open Innovation: How They Change Innovation Processes and Strategies in the Agrifood Sector? *Front. Sustain. Food Syst.* **2023**, *7*, 1267346. [CrossRef]
20. H2020 DEMETER Project: Building an Interoperable, Data-Driven, Innovative and Sustainable European Agri-Food Sector [Internet]. Available online: https://cordis.europa.eu/project/id/857202/es (accessed on 9 April 2022).
21. Šestak, M.; Copot, D. Towards Trusted Data Sharing and Exchange in Agro-Food Supply Chains: Design Principles for Agricultural Data Spaces. *Sustainability* **2023**, *15*, 13746. [CrossRef]
22. Catedra ENIA-AGRI. Universidad de Córdoba [Internet]. Available online: https://www.uco.es/servicios/actualidad/noticiasactualidaddia/item/153506-la-universidad-de-cordoba-consigue-1-2-millones-de-euros-para-promover-la-inteligencia-artificial-en-el-sector-agroforestal (accessed on 9 June 2024).
23. AINIA Packaging Hub: Innovación y Sostenibilidad. Objetivo H2030 [Internet]. Available online: https://www.ainia.es/packaging-hub (accessed on 9 June 2024).

24. Datos.Gob.Es [Internet]. Available online: https://datos.gob.es/es/noticia/plan-de-impulso-de-los-espacios-de-datos-sectoriales (accessed on 2 December 2024).
25. Durmus, H.; Gunes, E.O. Integration of the Mobile Robot and Internet of Things to Collect Data from the Agricultural Fields. In Proceedings of the 2019 8th International Conference on Agro-Geoinformatics, Agro-Geoinformatics, Istanbul, Turkey, 16–19 July 2019. [CrossRef]
26. Caruso, A.; Chessa, S.; Escolar, S.; Barba, J.; Lopez, J.C. Collection of Data with Drones in Precision Agriculture: Analytical Model and LoRa Case Study. *IEEE Internet Things J.* **2021**, *8*, 16692–16704. [CrossRef]
27. Madhumathi, R.; Arumuganathan, T.; Shruthi, R. A Survey on Wireless Sensor Networks and Instrumentation Techniques for Smart Agriculture. *Lect. Notes Data Eng. Commun. Technol.* **2022**, *68*, 453–467. [CrossRef]
28. Madhumathi, R.; Arumuganathan, T.; Shruthi, R. Internet of Things in Precision Agriculture: A Survey on Sensing Mechanisms, Potential Applications, and Challenges. *Lect. Notes Netw. Syst.* **2022**, *213*, 539–553. [CrossRef]
29. Klaina, H.; Guembe, I.P.; Lopez-Iturri, P.; Campo-Bescós, M.Á.; Azpilicueta, L.; Aghzout, O.; Alejos, A.V.; Falcone, F. Analysis of Low Power Wide Area Network Wireless Technologies in Smart Agriculture for Large-Scale Farm Monitoring and Tractor Communications. *Measurement* **2022**, *187*, 263–2241. [CrossRef]
30. Joint, R.; Kumar Singh, R.; Hasan Rahmani, M.; Weyn, M.; Berkvens, R. Joint Communication and Sensing: A Proof of Concept and Datasets for Greenhouse Monitoring Using LoRaWAN. *Sensors* **2022**, *22*, 1326. [CrossRef]
31. Gaia-X Domain Agriculture [Internet]. 2021. Available online: https://www.bundeswirtschaftsministerium.de/Redaktion/EN/Publikationen/Digitale-Welt/211116-pp-agriculture.pdf?__blob=publicationFile&v=3 (accessed on 9 January 2025).
32. European Commision Shaping Europe's Digital Future. 2020. Available online: https://commission.europa.eu/system/files/2020-02/communication-shaping-europes-digital-future-feb2020_en_4.pdf (accessed on 6 February 2026).
33. Giray, G.; Catal, C. Design of a Data Management Reference Architecture for Sustainable Agriculture. *Sustainability* **2021**, *13*, 7309. [CrossRef]
34. Murakami, E.; Saraiva, A.M.; Ribeiro, L.C.M.; Cugnasca, C.E.; Hirakawa, A.R.; Correa, P.L.P. An Infrastructure for the Development of Distributed Service-Oriented Information Systems for Precision Agriculture. *Comput. Electron. Agric.* **2007**, *58*, 37–48. [CrossRef]
35. Steinberger, G.; Rothmund, M.; Auernhammer, H. Mobile Farm Equipment as a Data Source in an Agricultural Service Architecture. *Comput. Electron. Agric.* **2009**, *65*, 238–246. [CrossRef]
36. Ouafiq, E.M.; Saadane, R.; Chehri, A.; Jeon, S. AI-Based Modeling and Data-Driven Evaluation for Smart Farming-Oriented Big Data Architecture Using IoT with Energy Harvesting Capabilities. *Sustain. Energy Technol. Assess.* **2022**, *52*, 102093. [CrossRef]
37. Zhaoliang, L.; Huang, W.; Wang, D. Functional Agricultural Monitoring Data Storage Based on Sustainable Block Chain Technology. *J. Clean. Prod.* **2021**, *281*, 124078. [CrossRef]
38. Debauche, O.; Mahmoudi, S.; Manneback, P.; Lebeau, F. Cloud and Distributed Architectures for Data Management in Agriculture 4.0: Review and Future Trends. *J. King Saud Univ.—Comput. Inf. Sci.* **2021**, *54*, 7494–7514. [CrossRef]
39. Ngo, V.M.; Le-Khac, N.-A.; Kechadi, M.-T. Designing and Implementing Data Warehouse for Agricultural Big Data. In *BigData 2019*; Springer: Cham, Switzerland, 2019.
40. Ouafiq, E.M.; Saadane, R.; Chehri, A. Data Management and Integration of Low Power Consumption Embedded Devices IoT for Transforming Smart Agriculture into Actionable Knowledge. *Agriculture* **2022**, *12*, 329. [CrossRef]
41. Ministerio De Agricultura. P. y A. Mapa.Org.Es [Internet]. Available online: https://www.mapa.gob.es/es/ministerio/default.aspx (accessed on 5 March 2022).
42. Junta de Andalucía Red de Información Agroclimática de Andalucía. API Rest Del Servicio Web de RIA. Spring Boot REST API RIA [Internet]. Available online: https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaws/swagger-ui.html (accessed on 5 March 2022).
43. AEMET OpenData y de Estadísticas. Plan RISP. [Internet]. Available online: https://www.aemet.es/es/datos_abiertos (accessed on 5 March 2023).
44. Copernicus: Europe's eyes on Earth European Union's Earth Observation [Internet]. Available online: https://www.copernicus.eu/en (accessed on 5 March 2022).
45. Ministerio de Transporte. PNOA: Plan Nacional de Ortofotografía Aérea. [Internet]. Available online: https://pnoa.ign.es/ (accessed on 5 March 2022).
46. Sentinel Hub EO Browser Cloud API for Satellite Imagery. [Internet]. Available online: https://apps.sentinel-hub.com/eo-browser/?zoom=5&lat=44.33957&lng=-0.70312&themeId=DEFAULT-THEME&toTime=2021-06-25T07:33:04.102Z (accessed on 5 March 2022).
47. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
48. EOS Land Viewer Download Satellite Imagery. [Internet]. Available online: https://eos.com/landviewer/?lat=37.89160&lng=-4.77270&z=11 (accessed on 5 March 2022).

49. Datos Obtenidos Del Seguimiento de Plagas y Enfermedades En Las Estaciones de Control Biológico [2006–2021]. Red de Alerta e Información Fitosanitaria (RAIF) [Internet]. Available online: https://www.juntadeandalucia.es/datosabiertos/portal/dataset/raif (accessed on 5 March 2022).
50. Programa TRIANA: Tratamientos Integrados En Andalucía En Agricultura. [Internet]. Available online: https://www.juntadeandalucia.es/organismos/agriculturaganaderiapescaydesarrollosostenible/areas/agricultura/sanidad-vegetal/paginas/produccion-integrada-atrias-programas-triana.html (accessed on 5 March 2022).
51. Descarga de Información Geográfica SIGPAC. [Internet]. Available online: https://www.juntadeandalucia.es/organismos/agriculturaganaderiapescaydesarrollosostenible/servicios/sigpac/visor/paginas/sigpac-descarga-informacion-geografica-shapes-provincias.html (accessed on 5 March 2022).
52. RDocumentation: Scatter Plots Function [Internet]. Available online: https://www.rdocumentation.org/packages/plot3D/versions/1.4/topics/Scatter%20plots (accessed on 25 March 2022).
53. RDocumentation: ImputeTS Package [Internet]. Available online: https://www.rdocumentation.org/packages/imputeTS/versions/3.2 (accessed on 25 March 2022).
54. RDocumentation: OutForest Package [Internet]. Available online: https://www.rdocumentation.org/packages/outForest/versions/0.1.1 (accessed on 25 March 2022).
55. Iberpix: Visualizador Cartográfico Publicado Por El Centro Nacional de Información Geográfico. [Internet]. Available online: https://iberpix.cnig.es/iberpix/visor/ (accessed on 5 March 2022).
56. Artículo 122 de La Ley 37/1988, de 28 de Diciembre, de Presupuestos Generales Del Estado Para 1989 y Está Adscrito al Ministerio de Transportes, Movilidad y Agenda Urbana, a Través de La Dirección General Del Instituto Geográfico Nacional (IGN). [Internet]. Available online: http://www.ign.es/web/ign/portal/qsm-cnig (accessed on 5 March 2022).
57. Ministerio de Transporte. PNOA: Comparador de Fotos. [Internet]. Available online: https://www.ign.es/web/comparador_pnoa/index.html (accessed on 5 March 2022).
58. FEGA: Fondo Español de Garantía Agraria. Visualizador PNOA. [Internet]. Available online: https://sigpac.mapa.gob.es/fega/visor/ (accessed on 5 March 2022).
59. Catálogo de Servicios de Visualización Inspire Del SIGPAC: Servicios WMS. [Internet]. Available online: https://www.mapa.gob.es/es/cartografia-y-sig/ide/directorio_datos_servicios/agricultura/servicios-wms-sigpac/wms_sigpac.aspx (accessed on 5 March 2022).
60. Ministerio de Transporte. PNOA: WMTS Capabilities. [Internet]. Available online: https://www.ign.es/wms-inspire/pnoa-ma?Request=GetCapabilities&Service=WMTS (accessed on 5 March 2022).
61. Ministerio de Transporte Movilidad y Agencia Urbana Centro Descargas PNOA: Instituto Geográfico Nacional. [Internet]. Available online: https://pnoa.ign.es/pnoa-imagen/productos-a-descarga (accessed on 5 March 2022).
62. Sentinel Application Platform (SNAP). [Internet]. Available online: http://step.esa.int/main/toolboxes/snap/ (accessed on 6 February 2026).
63. Visor de Red de Alerta de Información Fotosanitaria. [Internet]. Available online: https://www.juntadeandalucia.es/agriculturaypesca/visorraif/ (accessed on 5 March 2022).
64. Sistema de Información Geográfica de Parcelas Agrícolas (SIGPAC). [Internet]. Available online: https://www.mapa.gob.es/es/agricultura/temas/sistema-de-informacion-geografica-de-parcelas-agricolas-sigpac- (accessed on 6 February 2026).
65. AEMET: Variables. Estadística de Las Variables Meteorofenológicas. Available online: https://www.aemet.es/es/datos_abiertos/estadisticas/estadistica_meteorofenologicas (accessed on 5 March 2022).
66. Copernicus: Swagger UI. TimeSeriesServiceV11: Computes Times Series for a Product over a given Polygon or Point. [Internet]. Available online: https://proba-v-mep.esa.int/api/timeseries/apidocs/ (accessed on 5 March 2022).
67. Schuetz, C.G.; Schausberger, S.; Schrefl, M. Building an Active Semantic Data Warehouse for Precision Dairy Farming. *J. Organ. Comput. Electron. Commer.* **2018**, *28*, 122–141. [CrossRef]
68. Groher, T.; Heitkämper, K.; Umstätter, C. Digital Technology Adoption in Livestock Production with a Special Focus on Ruminant Farming. *Animal* **2020**, *14*, 2404–2413. [CrossRef]
69. da Silveira, F.; da Silva, S.L.C.; Machado, F.M.; Barbedo, J.G.A.; Amaral, F.G. Farmers' Perception of the Barriers That Hinder the Implementation of Agriculture 4.0. *Agric. Syst.* **2023**, *208*, 103656. [CrossRef]
70. Boros-Papp, S.; Várallyai, L. The Opinion of Farmers and Small and Medium-Sized Enterprises on the Importance of ICT in Hajdú-Bihar County, Hungary. *J. Agric. Inform.* **2019**, *10*, 60–67. [CrossRef]
71. Alfonsi, R.M.; Naidoo, M.; Gasparatos, A. Adoption and Desirable Characteristics of Information and Communication Technologies for Urban Small-Scale Food Producers in South Africa. *Front. Sustain. Food Syst.* **2024**, *8*, 1332978. [CrossRef]
72. Ayim, C.; Kassahun, A.; Addison, C.; Tekinerdogan, B. Adoption of ICT Innovations in the Agriculture Sector in Africa: A Review of the Literature. *Agric. Food Secur.* **2022**, *11*, 22. [CrossRef]
73. Chamorro-Padial, J.; Virgili-Gomá, J.; Gil, R.; Teixidó, M.; García, R. Agriculture Data Sharing Review. *Heliyon* **2025**, *11*, e41109. [CrossRef]

74. Mutia, I.; Sitanggang, I.S.; Annisa, A.; Astuti, D.A. Application of Spatial Data Warehouse for Agriculture: Challenge and Future Trends. In Proceedings of the 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE, Depok, Indonesia, 14–15 September 2021; pp. 277–282. [CrossRef]
75. Bethwell, C.; Sattler, C.; Stachow, U. An Analytical Framework to Link Governance, Agricultural Production Practices, and the Provision of Ecosystem Services in Agricultural Landscapes. *Ecosyst. Serv.* **2022**, *53*, 101402. [CrossRef]
76. Martínez-Ruedas, C.; Guerrero-Ginel, J.E.; Fernández-Ahumada, E. A Methodology for Automatic Identification of Units with Ecological Significance in Dehesa Ecosystems. *Forests* **2022**, *13*, 581. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.