Methodology for Developing Comprehensive Data Acquisition, Processing, and Maintenance Methods for Data Science Applications in Precision Agriculture

Abstract and description

This work aims to develop a comprehensive methodology for designing and rapidly adjusting data acquisition, processing, and maintenance systems for data science applications in precision agriculture. While many publications provide solutions to specific problems, such as large amount different scale remote sensing data collection and processing, they often lack adaptable, scalable IT solutions that users can easily change and extend as modern technologies appear.

The primary goal is to create a flexible method that allows for the rapid customization of these systems to address various use cases in the ever-evolving global agricultural landscape. As new data collection tools and methods continuously appear, the system should enable efficient integration of both current and historical data. The prototype system developed in this research will serve as a model for how data can be systematically collected, processed, and maintained, while showing how this method can be adapted to changing technologies.

Agricultural data is inherently diverse, forming real-time and historical data, sensor-based and geospatial information, structured and unstructured formats, and varying dimensions. The methodology addresses key aspects such as dynamic data acquisition, processing automation, secure storage solutions, and continuous data validation to ensure accuracy over time.

Additionally, the system will incorporate advanced modeling techniques, such as digital twins and socio-technical simulations, allowing for the representation of real-world agricultural processes in a virtual environment. This approach will support seamless integration of machine learning, statistical analysis, and geospatial visualization, providing valuable insights for farmers and agricultural scientists.

Ultimately, the goal is to create a flexible, scalable method that can be rapidly adjusted for different use cases, allowing precision agriculture systems to keep pace with technological advancements and data science needs.

The primary data collection methods for this project included UAV-based multispectral imaging and LiDAR technology, which are crucial for capturing high-resolution, geospatial, and vegetation-specific data in precision agriculture.

For data storage, the Oracle Cloud Infrastructure (OCI) is used due to its wide functionality and scalability, including geospatial capabilities, AI and machine learning tools, and robust ETL processes where SQL is the main language, with support for Python and R. Data preparation tools employed include QGIS, Pix4D Mapper, Bentley Context Capture, and Hexagon Erdas Imagine.

Methodology provides ways to integrate historical records, field data in Excel, IoT sensor data, geospatial data, meteorological data, and statistical data like previous year harvests. It incorporates all data types into structured data model. This work has significantly drawn attention to the fact that metadata maintenance is an especially important task, for which a place is also provided in this data model.

Problem Statement:

The integration of multispectral, LiDAR, sensor, and historical data into a unified digital geospatial data model for use in precision agriculture faces significant challenges. There is a "gap" between experts in different fields, such as database and data transformation specialists, and agricultural professionals. This disconnect highlights the need for knowledge consolidation, a unified approach to solutions, and a shared understanding of data acquisition and transformation processes. Additionally, a clear methodology is needed for using tools in the rapidly evolving landscape of available technologies and solutions.

Subject:

Algorithms (sample protocols) for Integration, Processing, and Visualization of Geospatial and IoT Data in Precision Agriculture Using Remote Sensing Technologies, Data Transformation, and New Analytical Methods.

Aim and expected impact:

The goal is to develop a methodology for data acquisition, extraction from other systems, and transformation processes for solving tasks in precision agriculture. In essence, it is a methodology for developing methodologies, as there are always differences in how data is used, which depends on the available devices, software, and the users' existing experience. As main application it will be tested and verified in practice in AREI (The institute of Agricultural Resources and Economics in Latvia, which is an important bioeconomy industry research and leading field plant breeding institute with more than 100 years of history)

Hypotesis :

A specialist (or specialists) within a company or scientific institute can successfully adopt and effectively use rapidly changing technologies only if a well-adapted methodology, tailored to specific tasks and tools, has been developed and tested, with a significant emphasis on knowledge transfer.

Research methodology:

Data collection – real data, collected from the field using UAV and IoT sensors, open data. Use available systems and tools, adapt as much as possible they offer “from the box”. Strict design and definition data preparation protocols and preprocessing steps. Experimental prototype of the system. Choosing specific tools and methods are figured out by “real life” situation in the work environment e.g., institutional financial resources, projects, requirements and needs.

Research Questions:

1. How can a scalable and flexible system be designed to integrate diverse data sources (multispectral, LiDAR, IoT sensors) with historical data for precision agriculture?
   Focus: System architecture and modularity.
2. What are the most effective methodologies for transforming, analyzing, and visualizing complex, multimodal agricultural data?
   Focus: Data processing, transformation, and visualization.
3. How can real-time and historical data be seamlessly integrated into a database based geospatial model for precision agriculture to improve decision-making?
   Focus: Data integration and real-time processing, accessibility of results.
4. How can tools and technologies in precision agriculture be rapidly adapted to keep up with new data collection methods and rapidly evolving sensor technologies?
   Focus: Flexibility and adaptability in system design.
5. What key principles should guide the design of a modular, scalable system for long-term data storage and accessibility in cloud environments like Oracle Cloud Infrastructure?

Focus: Cloud scalability, data storage, and accessibility.

6. How can knowledge transfer between data scientists, agricultural experts, and technology providers be optimized to ensure efficient use of new and emerging technologies?

   Focus: Interdisciplinary collaboration and knowledge transfer.

Objectives:

1. To develop a modular and adaptable system for integrating real-time and historical agricultural data (from UAV multispectral imagery, LiDAR, IoT sensors, etc.) using cloud-based infrastructure.

   Outcome: A flexible system design for precision agriculture data.

2. To create a methodology for efficient data processing, transformation, and storage using Oracle Cloud Infrastructure, using free open-source tools like QGIS and/or Pix4D Mapper, Bentley Context Capture, Erdas Imagine (could be any others, depends on licenses bought).

   Outcome: A clear, adaptable workflow for data integration and storage.

3. To define principles and guidelines for selecting appropriate tools and technologies for specific data collection, processing, and visualization tasks in precision agriculture.

   Outcome: A framework for tool choice and methodology development.

4. To explore and confirm innovative approaches for multimodal data fusion and analysis that integrate various agricultural data types (e.g., sensor data, historical records, and geospatial information) in a single system.

   Outcome: A tested and validated approach to data fusion for precision agriculture.

5. To help knowledge transfer and collaborative development between data science experts, agricultural specialists, and technology providers, enhancing the usability and efficiency of emerging technologies.

   Outcome: A methodology to streamline interdisciplinary collaboration.

Results:

- Comparable data collected over several years, integrated into a unified information system and presented in a format understandable to a wide range of users, using databases and geospatial information system tools for visualization, analysis, and reporting.
- Algorithms and methodology developed for the creation of such a system.
- Structuring, classification, and analysis of test data sets.

Scientific Contribution:

In practice, users need tools and systems that allow them to effectively carry out specific tasks and solve problems. To build such a system, a methodology is needed that leverages available, proven, and most suitable tools and techniques. ***The scientific contribution of this work*** *is the development of a method that offers a way to assemble such a system while defining key principles and offering recommendations for selecting the right tools. Additionally, it introduces principles for the development of new data analysis methods*.

These principles include:

- Modularity: The ability to divide the system into smaller, independent components, helping easier maintenance and customization.
- Service-oriented architecture: Building the system around services that ensure flexibility and integration.
- Integrated structured data storage: Ensuring data consistency and accessibility for various users and processes.
- Data availability: Providing data in a simple and readily accessible manner, based on the needs of system users.
- System customization tool integration: Enabling system adaptation with minimal coding, allowing users to easily tailor system functionality to their needs.
- System scalability: The ability of the system to scale from small to large operations, ensuring sustainability and efficiency.

These elements ensure a system that is suitable, scalable, adaptable, and efficient for users, capable of responding to the rapidly evolving challenges of data collection and analysis in precision agriculture. The decision to choose the right platform (system) as base is especially important because changes later normally are extremely expensive. OCI was chosen because of a good free plan with various upgrade options.

In the context of outlined in the Scientific Contribution, there are several areas that may still need further development or ongoing research, particularly in rapidly evolving fields like precision agriculture, data science, and geospatial technologies. Here are shown a few potential gaps or challenges that might not be fully developed yet:

1. Real-time Integration of Diverse Data Sources.

While methods for integrating structured data from various sources exist, fully realizing a seamless, real-time integration of IoT sensor data, UAV multispectral imagery, LiDAR, and historical data into a single coherent system is still a developing field. The challenge lies in synchronizing data with varying formats, update rates, and quality in a scalable way.

2. Automated Knowledge Transfer Mechanisms.

Although knowledge transfer is identified as critical, fully automating this process—where a system can automatically adapt to new tools, datasets, or analytical methods based on user feedback or emerging technologies—has not yet been fully achieved. Current systems often require manual intervention or user training, which can slow down adoption of new methodologies.

3. Advanced Modular Systems with Minimal Coding.

While modular systems and low-code/no-code platforms are progressing, the level of customization and flexibility needed for precision agriculture (especially with diverse data types like geospatial and multispectral) may not be fully supported in current systems. There is ongoing development to make such systems more user-friendly and adaptable without requiring significant programming skills.

4. Scalable Geospatial Data Processing in Real-Time.

While geospatial data processing tools are available (e.g., QGIS, Oracle Cloud's geospatial capabilities), the challenge of handling large-scale, multi-source, real-time data streams in a scalable manner are still not fully resolved. Processing and visualizing high-resolution LiDAR or multispectral data across vast areas in real time, while ensuring accuracy and speed, stays an open issue.

5. Machine Learning Integration with Geospatial Systems.

Integrating advanced machine learning algorithms directly within geospatial data processing systems for precision agriculture is still a growing area. The ability to seamlessly apply AI/ML models on real-time geospatial data streams (e.g., predicting crop yields or detecting anomalies) in a scalable way is a field where significant advancements are expected.

6. Standardization Across Tools and Systems.

Another challenge lies in the standardization of methodologies, tools, and formats across various platforms. While there are tools that help data integration, there is no widely adopted standard for ensuring that different systems (such as those collecting UAV data or IoT sensor data) can communicate and integrate efficiently without extensive customization.

These are some areas that might still require further development or are in the first stages of becoming more robust solutions. Addressing these gaps could further enhance the proposed methodology, making it more versatile, scalable, and adaptable for precision agriculture and other data-intensive fields.