A Novel Decision Support System for the Interpretation of Remote Sensing Big Data

Wadii Boulila[1,2], Imed Riadh Farah[1,2], Amir Hussain[3]

1: RIADI Laboratory, National School of Computer Sciences, University of Manouba, Tunisia
2: ITI Department, Telecom-Bretagne, University of Rennes 1, France
3: Division of Computing Science & Maths, School of Natural Sciences, University of Stirling, United Kingdom

Abstract

Applications of remote sensing (RS) data cover several fields such as: cartography, surveillance, land-use planning, archaeology, environmental studies, resources management, etc. However, the amount of RS data has grown considerably due to the increase of aerial and satellite sensors. With this continuous increase, the necessity of having automated tools for the interpretation and analysis of RS big data is clearly obvious. The manual interpretation becomes a time consuming and expensive task.

In this paper, a novel tool for interpreting and analyzing RS big data is described. The proposed system allows knowledge gathering for decision support in RS fields. It helps users easily make decisions in many fields related to RS by providing descriptive, predictive and prescriptive analytics.

The paper outlines the design and development of a framework based on three steps: RS data acquisition, modeling, and analysis & interpretation.

The performance of the proposed system has been demonstrated through three models: clustering, decision tree and association rules. Results show that the proposed tool can provide efficient decision support (descriptive and predictive) which can be adapted to several RS users' requests. Additionally, assessing these results show good performances of the developed tool.

Keywords: Decision Support System, Remote Sensing Data, Image Interpretation, ETL Process, Data Warehouse, predictive analytic, descriptive analytics, prescriptive analytics

1. Introduction

Nowadays, the decision making process plays a vital role in the Remote Sensing (RS) field [1] [2] [3] [4]. It provides promising information that can be used in several fields such as: meteorology [5], resources management [6], regional planning [7], assessment of the environment [8], weather forecasting [9], measuring of the damages caused by natural disasters [10] [11], etc.

Analysis and interpretation of RS data can support users in decision making by providing efficient strategic and productive decisions leading to reduced environmental risks and improved decision making.
However, the amount of RS data has grown considerably due to the development of several data sources. Thus, traditional processing for RS data has become increasingly difficult especially with the emergence of this so called RS big data.
Big data refers to massive, heterogeneous, and unstructured digital content that is difficult to process using traditional data management tools and techniques [12]. Thus, the proposition and development of new frameworks and tools for RS big data interpretation and analysis has become a necessity for RS users.
However, any attempt to develop a Decision Support System (DSS) for RS big data faces several challenges:

- **Complexity of RS data**

  RS data are distinguished by several characteristics compared to other data. Due to these characteristics RS data are more complex than other types of data. Among them, we can list the high dimensionality nature of pixels [13]. In fact, each pixel in the image is plotted in a space in which several bands exist and therefore it can have several interpretations (i.e. it can belong to several land cover use). Moreover, the neighborhood concept in RS data affects the computation of each pixel. In [14], authors confirm that, in a context of RS data, several dependencies between pixels influence the processing and interpretation of images.
  Another important issue related to RS data is preprocessing such as radiometric correction, geometric correction and atmospheric correction. Preprocessing of RS data is a key step in the whole processing and interpreting chain [15].
  Thus, making decision with RS data cannot use currently conventional DSS tools. It requires the use of new scalable techniques to deal with cleansing, storage, queries over complex multimedia data, etc.

- **Modeling RS big data**

  Modeling RS data determines the way in which this data is represented and subsequently interpreted, making it a significant phase in analyzing RS data.
  RS data comes in different formats: image, alpha-numerical, grid, map, etc. In general, tools for interpreting RS data suffer from the semantic gap problem. Semantic gap is defined as the problem of properly describing images when moving from low-level visual features to high-level semantics of RS data [16]. This semantic gap becomes all the more critical as the amount and diversity of RS data increases [17]. DSS applied to RS big data should deal with this problem in order to ensure more relevant decision making [18].

- **Analyzing and interpretation of RS big data**

  Analyzing and interpretation of RS data is a highly challenging task. There are two major commonly used techniques: manual interpretation and automatic interpretation. The process of interpretation depends on several factors like texture, radiometry, shape, as

well as the relationship between objects in satellite images. Due to the high volume, complex, multi-scale RS data, as well as need for real/near real time decision making, the automatic interpretation of RS data is vital [19] [20]. However, the automatic interpretation process is generally a difficult task and subject to many problems such as segmentation of images and extraction of useful spatial and thematic information on objects of interest using human knowledge [21].

A challenging problem to RS community is to analyze and interpret RS big data while considering the complexity and the continuous increase of data size. The problem is exacerbated by the nature of RS data which is obtained from multiple sources, types, scales and locations with varying degrees of quality and reliability [22]. Traditional techniques and algorithms for analyzing RS big data take huge time to execute and need highly-performance platforms to run. Thus, new techniques and tools are extremely required to analyze and interpret RS big data and in many times more complex processing is required.

Based on these challenges, the main contribution of the present paper is to process and model complex RS big data in order to analyze, interpret and provide valuable decisions than can be used in many RS fields.

In this paper, we present a DSS tool for RS big data analytics. The proposed tool provides an environmental decision support dashboard. We develop a new framework that helps the user make step-by-step decisions while effectively utilizing previously dealt experiences.

The main aim is to develop an intelligent decision support tool that integrates RS data coming from different sources and reshapes it to assist users in diverse fields related to RS. The developed framework is capable of learning and implementing personalized policies to build descriptive, predictive and prescriptive analytics.

The remaining paper is organized as follows: Section 2 presents related research in the field of DSS using images and RS big data. Section 3 describes the proposed DSS tool for RS big data analysis. Section 4 discusses experimental results. Section 5 summarizes conclusion and future works.

2. Related Works

Research in related fields has been carried out using images as well as RS big data.

- **DSS in image fields**

DSS refers to a computer-based information system that supports decision-making activities. It integrates tools, techniques, models and decision-making procedures, which communicate with an information processing system [23]. The challenge for DSS is not providing enough information for decision making, but how to screen the information and select which is relevant to the decision [23]. DSS is used in many fields such as business, management, marketing and agricultural production. Many models of decision making process have been proposed in the literature which incorporates relevant image analysis as its integral part.

In [24], Alcón et al. proposed an automatic system for inspection of pigmented skin lesions and melanoma diagnosis. The developed system is based on a DSS for calculating a personal risk factor, and an image processing system for feature extraction and classification. Authors consider

non-professional applications using images acquired with a conventional (consumer-type) digital camera.

Alaa et al. in [25] developed a clinical DSS that learns a personalized screening policy from electronic health record data. The presented system groups patients' features into clusters and provides screening policies for every cluster of patients. Application of the proposed DSS is to provide recommendations for women with features of screening sequences can lead to a possible risk of breast cancer.

In [26], Dempere-Marco et al. presented a DSS in image understanding based on information extracted from the dynamics of saccadic eye movements. The proposed framework is composed by an image feature library and uses Markov model for analysis of dynamics of the visual search. The goal was to provide a general feature learning for decision support in image understanding. The proposed system was validated through a clinical scenario on the pulmonary vascular distribution with computed tomography images.

He et al. in [27] proposed to use Cellular Automata (CA) to discover dynamic transition rules automatically. The objective of the proposed model is to retrieve urban dynamic evolution rules over time. It takes advantages of self-adaptive CA model integrated with an artificial immune system to discover dynamic transition rules. Application of He et al. system was to simulate the urban land conversion of Guangzhou city, located in the core of China's Pearl River Delta.

In [28], Hwangbo et al. suggested a DSS to assist in the selection of the optimal classification method or scheme. The proposed system is based on case based reasoning to assist users in land-cover classification task. Four features are determined to ensure case retrieval which are: dataset, location, climate and class.

Ai et al. in [29] proposed a dynamic DSS. The proposed system combines GIS and social media to provide decision making in case of tsunami risk mitigation in Padang, Indonesia. Many actors can use Ai et al. system such as government policy makers, policy managers, policy executors and urban citizens impacted by disasters. The main purpose is to design tsunami risk maps and timely evacuation routes in regions that are exposed to tsunami risk.

Fegraus et al. in [30] presented an environmental dashboard based on geo-referenced livelihood data coming from household surveys and biophysical data collected from RS imagery. The aim of the proposed system is to compute a variety of metrics of ecosystem stress. Application of this system is to propose a DSS to monitor Tanzanian agriculture and ecosystem services.

Most of works related to the DSS in image fields focus on medical related fields and the few of them that concentrate on RS field focus on a single area of application of their approaches. Besides, they use a single machine learning method to provide decision about RS domain and they don't consider the case of RS big data. In the present paper, a complete and detailed framework that allows processing RS big data is described. It can be, easily, transposed to others domains by modifying the features characterizing the considered application field. The proposed tool integrates several machine learning algorithms such as decision tree, neural networks, classification and association rules. This allows different use of the proposed tool and can support users by descriptive, prescriptive and predictive decisions.

- **RS Big Data**

RS refers to technology which measures object features or surfaces from a distance [31]. Nowadays, we have an extensive amount of RS data, due to the exponential growth of RS

sources (i.e. satellites). Also, the advance in spatial, temporal and spectral resolutions of sensors adds the volume of RS data. In many situations, the need to process a huge volume of data is a challenging problem for applications that require real time decisions [32].

In this paper, we consider big data as data characterized by 4 key attributes: volume, variety, velocity and value [33].

Several recent researches have been devoted to analyzing RS big data. Ma et al. in [31] gave an overview of existing techniques, tools and systems related to Big Data analysis, and especially to RS Big Data. The paper presents a state-of-the-art about system architecture, parallel file systems and databases, data managing tools and task scheduling for RS Big Data works. Giachetta in [34] presented a framework to process spatial and RS data in distributed environment. The proposed system is based on MapReduce and Hadoop paradigms. Rathore et al. in [35] proposed to proceed real-time and offline real-time and offline remote sensing big data. The proposed architecture is divided into three units: 1) RS big data acquisition unit (RSDU), 2) data processing unit (DPU), and 3) data analysis decision unit (DADU). Cavallaro et al. reviewed available parallelization techniques applied to RS field in [33]. The main objective of their paper is to identify which statistical data mining methods can be used in the field of remote sensing big data. The authors highlighted that parallel support vector machines (SVMs) are commonly used methods in data mining for remote sensing, and that the total time of the whole process can be significantly reduced by using parallelization methods.

RS big data can be used essentially for three main purposes: 1) descriptive analytics, 2) predictive analytics and 3) prescriptive analytics [36].

RS big data descriptive analytics is a deductive logic that aims to explain the relationship among different RS entities. In general, it regroups big data into smaller and more useful pieces of information. RS big data descriptive analytics summarizes three main questions: “what happened?”, “when it happened?” and “what is happening?”. Situations for RS big data descriptive analytics can be the description of land cover types, via descriptive statistics on satellite objects to determine the existence of unusual anomalies in the image data and discovering of contaminant and oil stains in sea surface.

The RS big data predictive analytics aims to create models in order to forecast trends. Based on a variety of statistical, modeling and machine learning techniques, RS big data predictive analytics exploits existing big data in order to predict future outcomes [37]. In many situations, incremental learning techniques are best suited to big data predictive analytics, as learning with a huge amount of data is difficult [38].

Examples of RS big data predictive analytics uses include predicting land cover changes, determining an unknown class label, predicting natural disasters using preventive decisions and alerting users for earthquakes.

The last type of RS big data is prescriptive analytics. This type addresses issues related to what we should do, why we should do it and what should happen with the best outcome under uncertainty [37]. The main difference between prescriptive analytics and predictive analytics is that the former provides multiple future decisions based on their decision-maker's actions. This type of analytics is recommended for situations where we need to predict consequences based on different choice of action. For example, RS big data prescriptive analytics can be used to help

environmental authorities, personnel, and engineers by providing them with strategies for land development, and resource management for land and regional planning.

3. Proposed approach

3.1. Proposed architecture

The main challenge of tools and systems in RS big data is being able to process a huge volume of data and providing the relevant decision in real time. As seen previously, RS data are combined from different sources and formats. This necessarily involves consolidation and cleaning steps before RS data can be used.

In this paper, we propose a new tool for DSS in RS big data. The goal of this tool is to provide valuable information and support for decision making in RS fields.

The proposed process is divided into the three main steps as shown by the Figure 1:

1. Step1: RS acquisition
   a. Collect RS big data coming from different sources.
   b. Solve complex problems related to heterogeneous RS data by applying several operations such as: cleaning, normalization, processing missing values, etc.
2. Step 2: modeling
   a. Load RS data into image data warehouse (DWH) and create a repository for images.
   b. Create RS data models: cubes, on line analytical processing, data marts and on line transactional processing.
3. Step 3: analyzing and interpretation
   a. Provide several types of report rendering: dashboard, reports, KPIs, statistics and decision support.

The main purpose of the first step is to regularly collect data coming from different sources so that it can be integrated during the next step named modeling step. During the first step many operations, generally, complex are achieved to prepare the data to be loaded into the data warehouse. The challenge is to integrate and consolidate large volume of RS data in a new unified data warehouse. The schema of the data warehouse is chosen during the modeling step according to the RS users' requirement. The modeling step requires the identification of all the main characteristics to meet most RS users' challenges and to respond to the different users' demands. A good modeling will enable fast querying in the analysis & interpretation step, ensure RS requirements to be met as well as to obtain accurate and meaningful decisions.

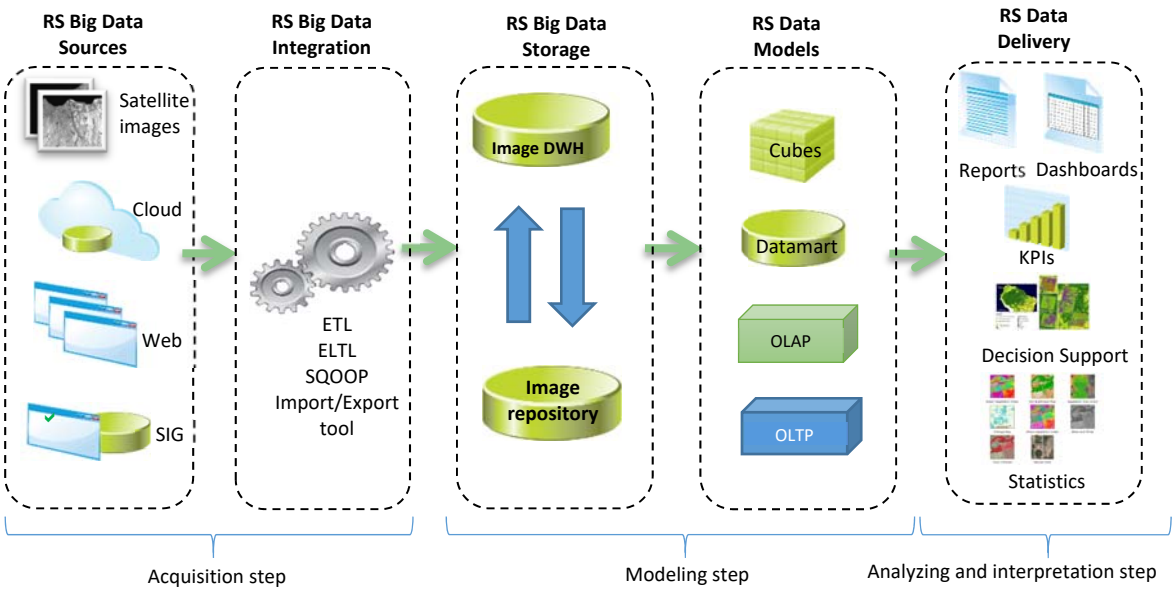


Figure 1. Proposed approach for RS Analysis

The main challenge of a DSS for RS data is to support users in making right decisions. This can be achieved by measuring the effectiveness of past decisions in order to optimize future recommendations. The proposed tool can be used in multiple fields such as: land cover classification, prediction of land cover change, disaster prevention, regional planning and management of resources. The frequency of data change in the current context is important. Thus, the proposed process is iterative and incremental as shown in Figure 2. The output of the DSS process can be translated into specific proactive or reactive information recommended for subsequent analysis with new coming RS data. In several cases, users' feedback must be taken into account in decision support process as it allows improving decision making in critical situation.

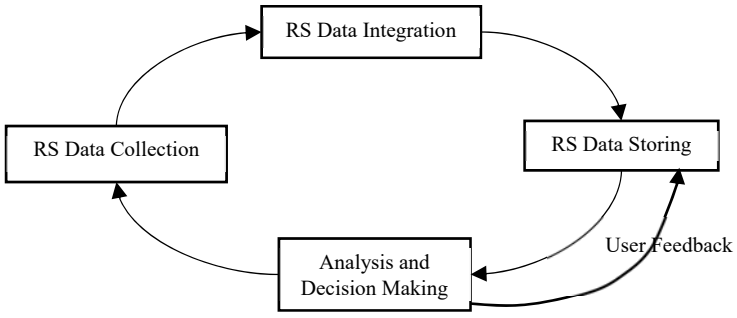


Figure 2. The proposed iterative and incremental process of DSS in RS big data

3.2.Acquisition step

3.2.1. RS Big Data Sources

RS data is collected from heterogeneous data sources. These sources include: satellite images, aerial photography, and GIS (Geographic Information System)

1. Satellite images: consist of images of the Earth collected by satellites, which aim to record, measure, analyze, and present information related to a phenomenon from a distance. Satellite images are coming from two types of sensors, namely passive and active. Most sensors record information about the surface of the Earth by measuring its resulting transmission of energy. Then, they provide specialized capabilities to manipulate, analyze, and visualize images. Mostly, satellite images are raster. This means that those images are made up of pixels (also referred to as grid cells). Each pixel stores a value representing the characteristic of the geographic feature at the cell location. Raster data is stored in various formats such as: IMG, TIFF, JPEG, ESRI grid, ADRG, BIL, CADRG, etc.
2. GIS: is a computer-based tool for capturing, storing, manipulating, analyzing, managing spatial or geographical data. It incorporates common database operations and geographical features with tabular data in order to analyze and assess world phenomena. Essentially, two types of data are used in GIS: raster and vector. Vector data is spatial data represented as geometrical shapes: points, lines and polygons. Each of these geometrical shapes are coupled to a row in a database known as attribute data. This information is housed in tabular format and it is useful to give more details about each spatial feature. Vector data is stored in various formats such as: DXF, GML, XYZ, GPX, SDC, VPF, KML, etc.

3.2.2. RS Big Data Integration

This step aims to combine data coming from various RS sources. This data is generally in different formats and characteristics, hence combining it first is essential for it to be processed by the proposed system. The challenging problem in RS data is the need for it to be processed before it can be analyzed and interpreted. For example, preprocessing is a key step in the satellite image chain which includes radiometric, geometric and atmospheric correction. Moreover, satellite images require generation of metadata to be queried.

The RS big data integration process includes three common tasks known as ETL (extract, transform and load). ETL is not a one-time process; as image databases are periodically supplied by new acquired images. Thus, in the proposed approach, RS integration process is an automated and an incremental process. This helps refresh the DWH with the new and the updated RS data. Data is extracted from data sources and loaded initially to staging image DWH after that to the image DWH. The proposed approach avoids including unnecessary data by identifying which data has been inserted or updated since last ETL cycle and limit extracting of this data.

Besides the three tasks of ETL, the RS big data integration process includes the process of extraction and storage of features from images. Figure 3 depicts the RS data integration process.

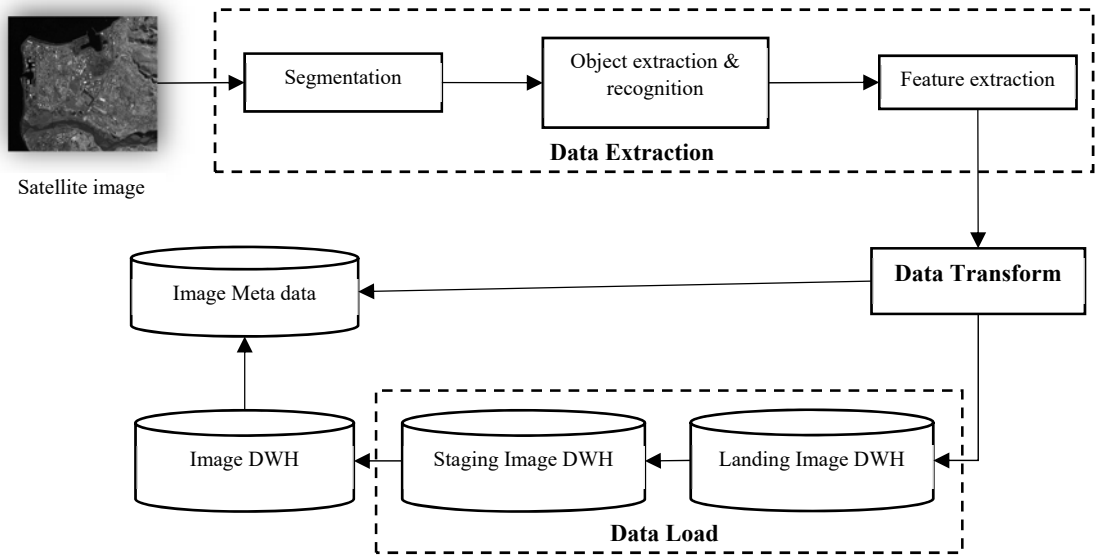


Figure 3. The proposed process for RS Big Data Integration

a. Data Extraction

The first task in the ETL process is data extraction. It aims to read useful data from RS images. To do this, three sub-tasks are sequentially performed which are: 1) segmentation, 2) object recognition and extraction, and 3) feature processing. The details are available in our previous work [39].

Segmentation aims to partition satellite images into homogeneous regions (also known as objects, clusters, segments). In this paper, a k-means method is used to segment images [40]. The second step determines meaningful objects from segmented images and assign a label to each object extracted from segmented images. The final step, feature processing, allows to describe extracted objects from images through a set of attributes.

In previous works, we proposed a set of attributes to describe objects extracted from satellite images [41] [42] [43]. The following attributes will be considered in this study:

- Spectral features: Mean Values and Standard Deviation values of Green (MVG, SDG), Red (MVR, SDR), NIR (MVN, SDN), SWIR (MVS, SDS) and Monospectral Bands (MVMB, SDMB).
- Textural feature: Homogeneity (Hom), Contrast (Cont), Dissimilarity (Dis), Entropy (Ent), Standard Deviation (SD) and Correlation (Cor).
- Shape features: Area (A), Length/Width (LW), Shape Index (SI), Roundness (R), Density (Den), Metric Relations (MR) and Direction Relations (DR) of each image object.
- Climate features: Temperature (Tem), Humidity (Hum) and Pressure (Pre).

- Vegetation feature: NDVI (Normalized Difference Vegetation Index) value for each object.

The considered features in this study can be substituted according to a specific field of application. For example, for the decision-making process in the medical field, features can be: patient features (name, gender, blood type, birth date,…), exam features (equipment, radio technology type, reason, ….), hospital features (name, address, city, …), etc.

b. Data Transformation

The main purpose of this task is cleaning and conforming of extracted data from satellite images to meet DWH requirements. Several rules are set according to final RS needs. In the proposed approach, several transformation actions can be made such as: split, union, merge, lockup and changes (data values, types and structures). Several transformation rules are added to ensure validation and cleaning of data.
In order to guarantee that RS data is suited to RS community usage, a knowledge-driven process for RS quality insurance is highly needed. This process is based on both computer-assisted and interactive ways to manage the integrity and quality of our RS data sources. The main goal of this process is to determine and build knowledge about our data. This knowledge can then be used in preprocessing tasks. The knowledge-driven process helps us to: 1) identify potentially incorrect data with an assessment of the likelihood that the data is in fact incorrect, 2) resolve issues related to incompleteness, lack of conformity, inconsistency, inaccuracy, invalidity, and data duplication.

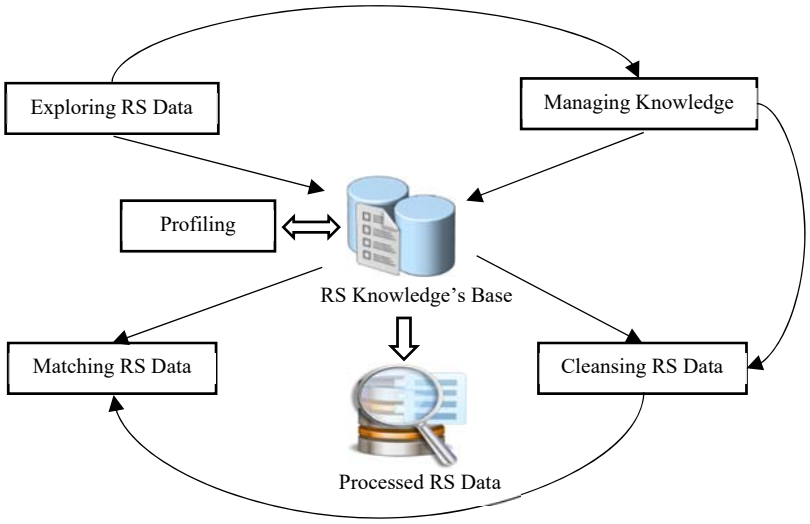


Figure 4. RS Quality Process

To ensure the quality of RS data, we build a semantic process which aims to analyze the data and decide its relevance. The built process allows resolving several issues related to RS data such as: incompleteness, lack of conformity, inconsistency, inaccuracy, invalidity, and data duplication.
Figure 4 presents the process of ensuring the RS data quality. This process is based on four steps: 1) exploring RS data, 2) managing knowledge, 3) cleansing RS data and 4) matching RS data.

The first step determines all RS data sources involved in the quality process. It also evaluates types and ranges of data in these sources. The second step which is knowledge management, is to create, modify and review the knowledge to be inserted in the knowledge base. The third step in the process of RS data quality, aims to analyze and make changes to the data. The final step enables reducing duplication between several RS data sources. The RS knowledge's base contains information related to RS data. The main goal of this base is to build semantic representations about frequently used data such as : synonym associations, term relationships, validation and business rules, and matching policies [44]. An additional task is added to the quality process which is profiling. It communicates with the RS knowledge's base in each of the four tasks mentioned above. Profiling helps improving the effectiveness of the quality processes. It generates notifications that in general cases lead to a recommeded action for processed RS data.

In this paper, in order to set up the RS semantic quality, we use the DQS provided by Microsoft [44]. The process of quality insurance is based on two steps: building the knowledge base (create, manage and train the knowledge base) and create the data quality project. For more details, interested readers can refer to [45] [46].

c. Data Load

RS data involves loading large volumes of data. A three-stage ETL process is recommended in this case. RS data is, first, extracted, integrated and loaded to a landing zone which has a similar architecture to data sources. Here, the main advantages are that users can explore, visualize and analyze RS data before validating it. Then, RS data is moved to the staging zone which has an architecture close to the DWH. Staging zone is required because the process of reading RS data is incremental, and many transformations and validation are needed before consuming the data in the warehouse. Another important issue is that the process of ETL in our case is time-consuming; therefore the three-stage ETL is a good alternative to resolve the problem.

3.2.3. RS Big Data Storage

To enforce scalability, data integrity, availability and high query performances, we consider three main operations in the physical design of the data warehouse and the RS big data storage: distribution, indexing and partitioning.

For distribution, we dole out RS data across several physical devices. Time is reduced by using a parallel thread processing when reading multiple sources. An architecture based on redundant array of independent disks (RAID 10) in a storage area network (SAN) is adopted in our context. The choice of RAID 10 is preferred since it combines disk mirroring and disk striping to protect data.

For indexing, we develop two types: dimension table and fact table indexing, to reduce data processing. For dimension table indexes, we choose to create nonclustered indexes for frequently demanded columns. We also created clustered indexes for year, month and date of the DimDate dimension. For fact table indexes, we created nonclustered for foreign key columns.

Partitioning aims to spread RS data over multiple nodes. The main purpose of this operation is to improve query performances, improve index manageability and increase flexibility of backup and recovery of RS data.

3.3.Modeling Step

The DWH is designed as a star schema. It contains a fact table (called FactState) and seven dimension tables (i.e. DimObject, DimNDVI, DimTexturalFeatures, DimSpectralFeatures, DimDate, DimClimateFeatures, DimShapeFeatures). The data warehouse designed in this paper allows different points of view according to specific features selected. Thus, we search for similar objects according to textural, spectral, NDVI, climate or shape features. This is very important in our approach and it can be useful in many cases, such as following land cover changes of an object according to specific features, or determining the features that most influence the object's changes.

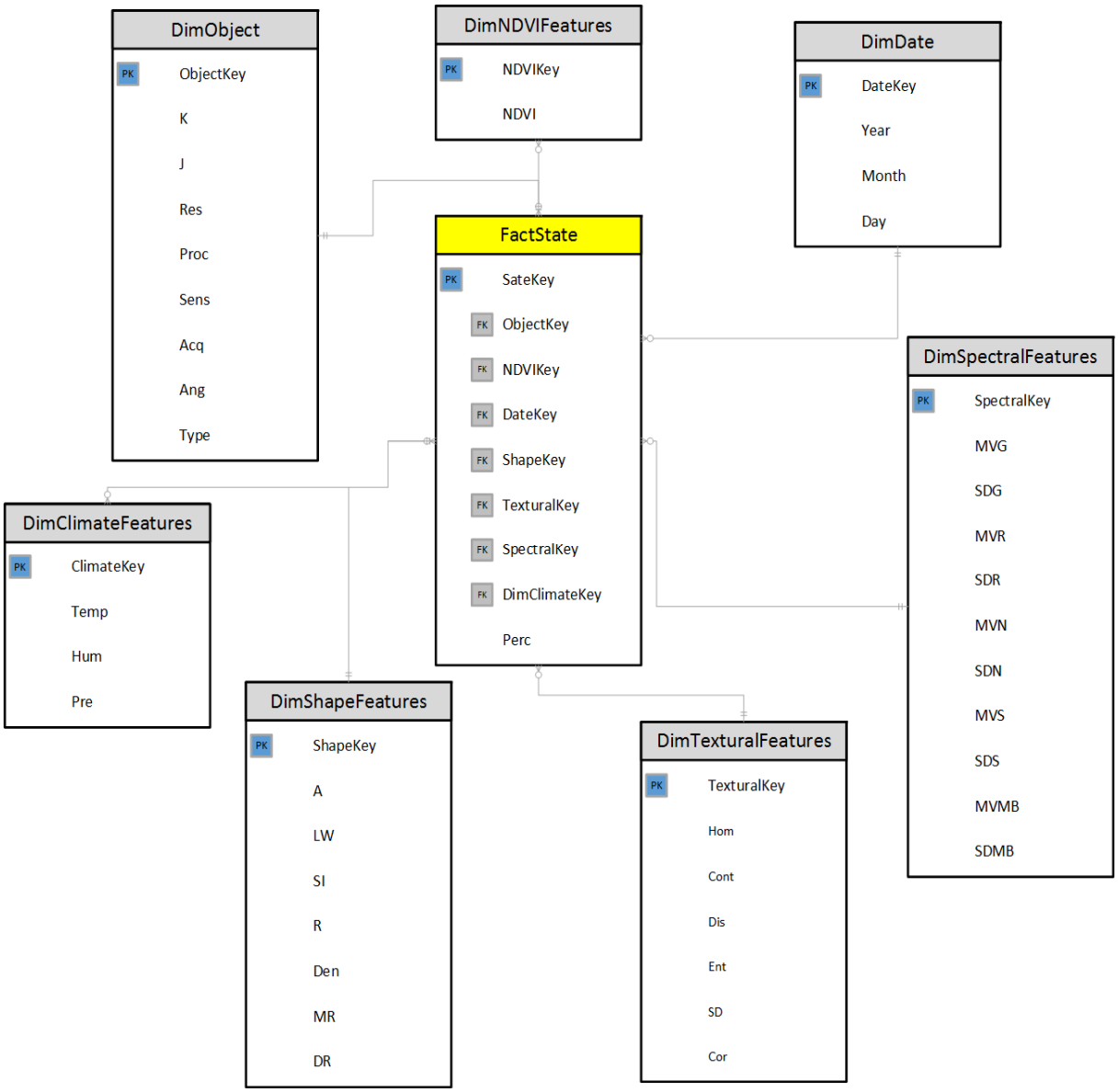


Figure 5. Proposed data warehouse

Figure 5 depicts the proposed data warehouse. The fact table is placed in the center of the star schema [47]. It contains a foreign key issued from the seven dimension tables and measurements necessary for several RS applications. Dimension tables are around the fact table and contain primary keys and attributes describing objects.
The proposed DWH architecture allows obtaining of several subsets, related to features representing states of a given object. Therefore, users can assess object states in different manners and evaluate changes of these states according to spectral, textural, climate or shape features. This will help evaluate the object behavior with regard to its features.

3.4.Analyzing and Interpretation Step

The final step in the proposed approach is RS data analysis and reporting. The main goal of the analysis step is to provide significant information that the RS community has requested. In general, this step aims to extend the capability of the DWH by adding useful business measures and key performance indicators (KPIs). The built data model provides an abstraction of RS fields according to the users' view. This helps information workers to easily understand the database schema design. A custom layer can be added to the data model. For example, defining KPIs to compare business measures with their targets, adding aggregation, and setting calculated fields.
In this paper, we used a multidimensional data model, which can create dimension tables, fact tables and one or more cubes, where each of them can depict RS users' requirements. In RS data analytics, a key point is to ensure a good understanding of the application domain.
The reporting step aims to deliver information gained from the analytic solution. This step is the most important, since it summarizes the whole effort made during the modeling process. Reporting task aims to present visual representations, create informative and elicit decisions. Reports and dashboards generated in this task assist RS users in decision making, reducing risks and summarizing information. Several types of reports can be used such as parameterized reports, linked reports, snapshot reports, cached reports and drilldown reports. RS users can view these reports in a web browser, custom application, excel or sent as email attachments.
Figure 6 depicts the proposed reporting architecture. The proposed architecture is divided into three main steps: 1) semantic query engine, 2) reporting processing and 3) reporting rendering. The first step aims to obtain data coming essentially from two sources (relational and multidimensional). In the report building step, the type of reports according to the user's requirements is chosen. The final step is report rendering. It aims to choose the delivery of built reports which can be for: web browser, email, office or custom application.

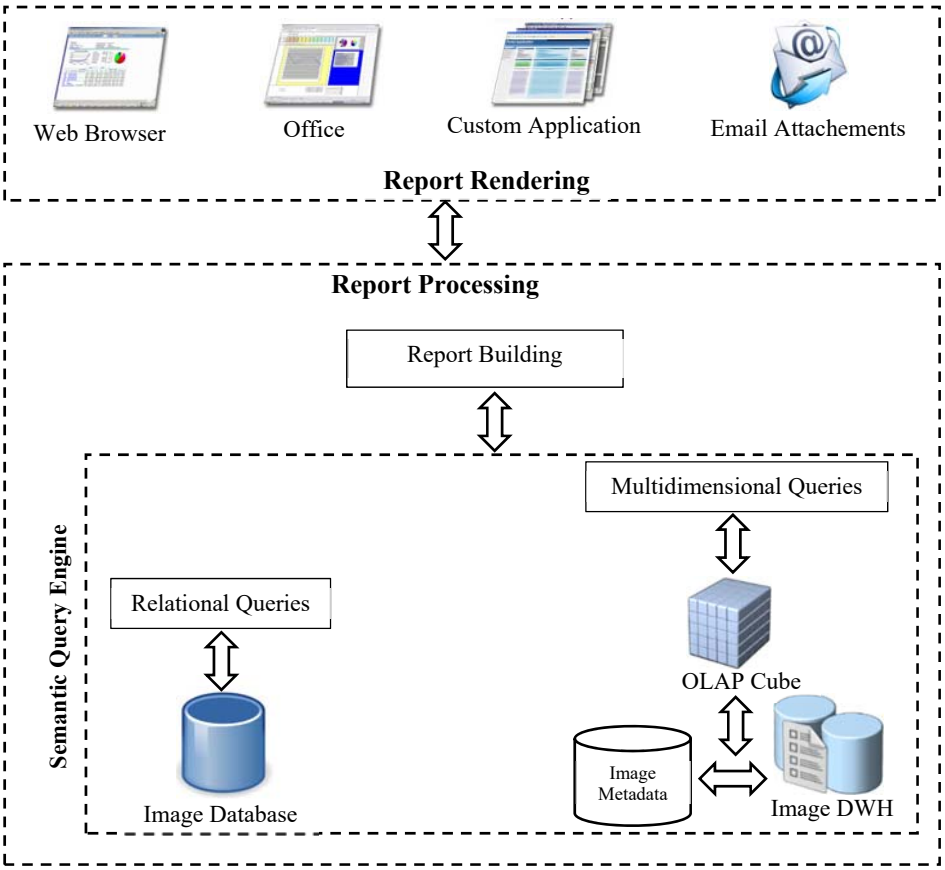


Figure 6. Proposed Reporting Architecture

4. Experimental results

To validate the proposed approach, we choose three different models which are: the classification model, the decision tree model and the association rules' model. Microsoft SQL Server 2012 is used to implement these three models.

4.1.Classification Model

The first application of the proposed tool is to group objects in the image DWH depending on their attributes. Clustering helps finding similar groups (clusters) inside the data especially if clusters are not obvious (case of hidden variables that accurately classifies the data).

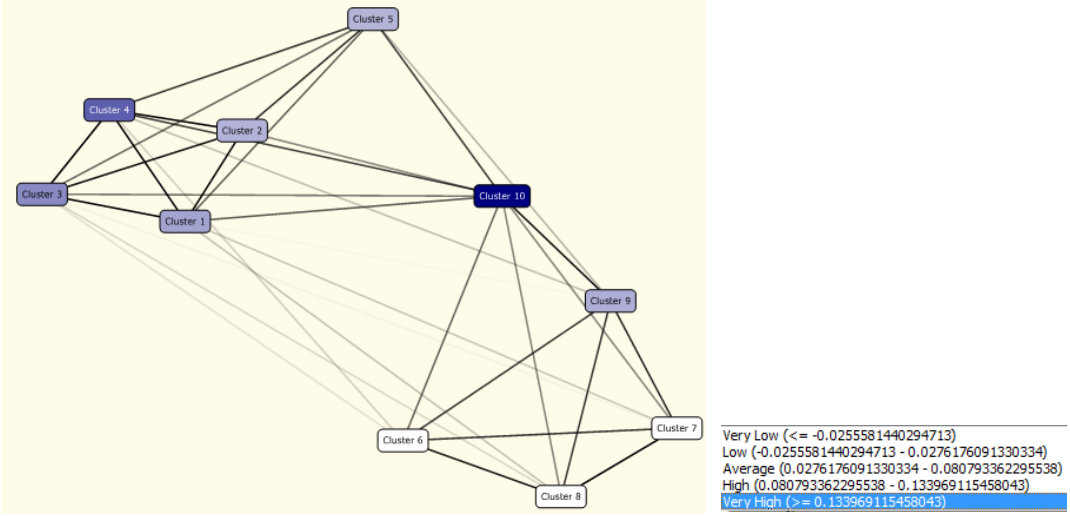
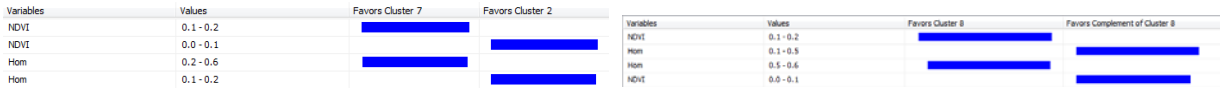


Figure 7. (a) Clustering of satellite objects according to the NDVI attribute and (b) different state of the NDVI attribute

Figure 7(a) presents a clustering of satellite objects in the image DWH according to the NDVI attribute. We obtain on each cluster a set of objects grouped by their features. Figure 7(b) depicts the different states of the NDVI attribute. Clusters will be shaded accordingly when modifying the NDVI state. Lines between two clusters represent the relationship between these two clusters (that means the two clusters are similar although they are distinct). The intensity of the line shows the degree of similarity; if the thickness is high it means that there is a strong relationship between the two clusters or those two clusters are close to each other.

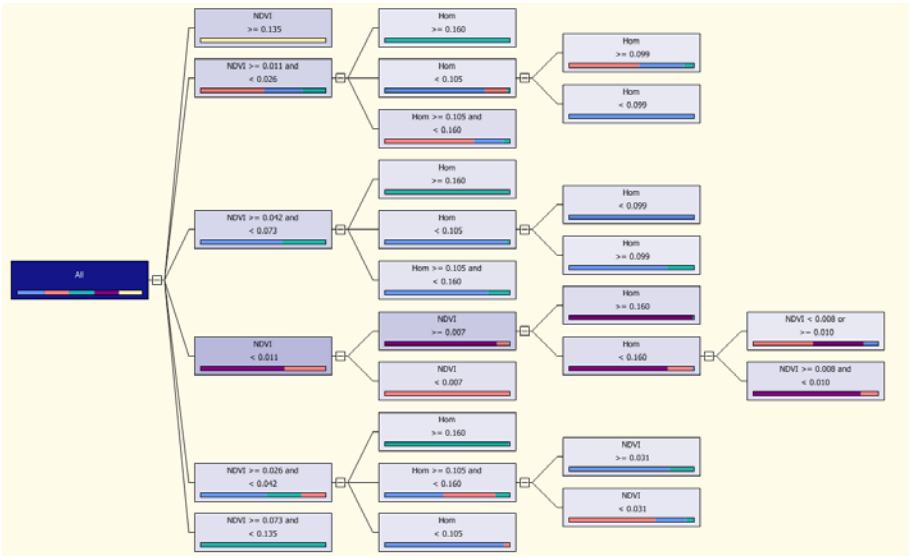


(a) Compared to another cluster (b) Compared to complement of the cluster

Figure 8. Clusters' comparison. (a) Compared to another cluster. (b) Compared to complement of the cluster

Figure 8 presents a clusters' comparison according to the different attribute values which can be done by using cluster discrimination. Figure 8(a) shows a comparison of a cluster 7 with cluster 8 according to attributes NDVI and Homogeneity. Figure 8(b) shows a comparison of a cluster with a complement of the cluster. This help to identify the given cluster features versus all the other clusters.

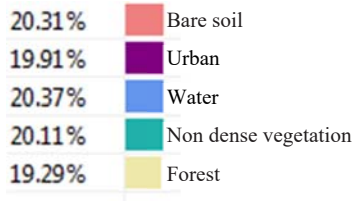
4.2.Decision Tree Model



(a) Fuzzy identification of land cover types based on NDVI and Hom

```
NDVI < 0.0105 (00000000200)
  NDVI < 0.0074 (0000000020000)
  NDVI >= 0.0074 (0000000020001)
    Hom < 0.15954 (000000002000100)
      NDVI >= 0.00802 and < 0.00988 (0000000020001 0000)
      NDVI < 0.00802 or >= 0.00988 (0000000020001 0001)
    Hom >= 0.15954 (000000002000101)
NDVI >= 0.0105 and < 0.026 (00000000201)
  Hom < 0.10482 (0000000020100)
    Hom < 0.099348 (000000002010000)
    Hom >= 0.099348 (000000002010001)
  Hom >= 0.10482 and < 0.15954 (0000000020101)
  Hom >= 0.15954 (0000000020102)
NDVI >= 0.026 and < 0.0415 (00000000202)
  Hom < 0.10482 (0000000020200)
  Hom >= 0.10482 and < 0.15954 (0000000020201)
    NDVI < 0.03065 (000000002020100)
    NDVI >= 0.03065 (000000002020101)
  Hom >= 0.15954 (0000000020202)
NDVI >= 0.0415 and < 0.0725 (00000000203)
  Hom < 0.10482 (0000000020300)
    Hom < 0.099348 (000000002030000)
    Hom >= 0.099348 (000000002030001)
  Hom >= 0.10482 and < 0.15954 (0000000020301)
  Hom >= 0.15954 (0000000020302)
NDVI >= 0.0725 and < 0.1345 (00000000204)
NDVI >= 0.1345 (00000000205)
```

(b) Rule describing the classification of land cover types based on NDVI and Hom



(c) Land cover types and their fuzzy membership based on NDVI and Hom

Figure 9. Identification of object land cover types. (a) Fuzzy identification of land cover types based on NDVI and Hom, (b) and (c)

The proposed tool can be used to determine land cover types of extracted objects from satellite images. One of the important advantages of the proposed tool is to investigate the effect of specific objects features in identifying the class of a given object. Also, the proposed tool provides a fuzzy membership to each land cover type at each tree node. Several paths can be matched to classify objects. Each path provides a fuzzy fitting of the considered object to the different land cover types. These paths can be translated into rules.

In the current example, we only use NDVI and Homogeneity (Hom) to classify objects. Figure 9(a) depicts the fuzzy classification tree based on NDVI and Hom. As we note, at each tree node, we have an indication of which land cover types (colors in the node) a given attribute can lead to. Figure 9(b) illustrates the translation of the fuzzy classification into a rule describing the classification of different land cover types based on NDVI and Hom. Figure 9(c) shows the different land cover types and their fuzzy membership.

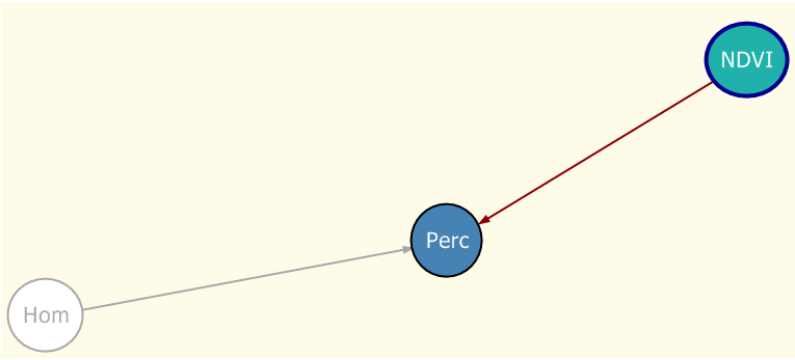


Figure 10. Influence of each attribute in the decision

Figure 10 shows the influence of each attribute in the decision. In our example, we can identify which attribute has the strongest influence in the percentage changes. Here, we note that the percentage of change depends deeply on the NDVI attribute.

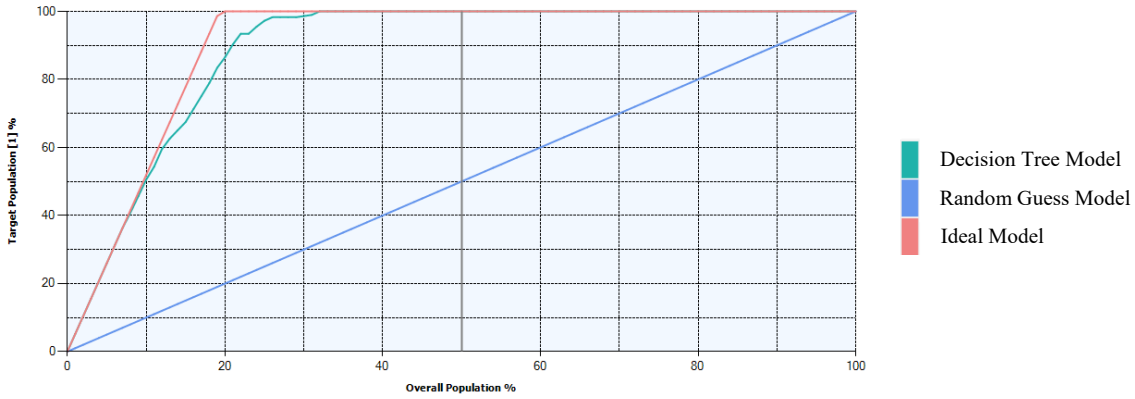


Figure 11. Evaluation of the prediction model using lift chart

In order to measure the effectiveness of the proposed model, we use two methods: lift chart and classification matrix.

Figure 11 depicts the lift chart method. This method assesses the proposed decision tree compared to a random guess. We compare different proposed models to determine the best one. In figure 11, the x-axis of the chart represents the percentage of the test dataset that is used to evaluate the predictions. The y-axis of the chart represents the percentage of

predicted values. The blue line refers to a random guess (diagonal line). A good model must be on top of this line. The red line represents perfection (ideal values) and the blue line refers to our decision tree model. We note that our model is close to the ideal model.

	Bare soil	Urban	Water	Non dense vegetation	Forest
Bare soil	242	23	4	0	0
Urban	0	257	0	0	0
Water	45	12	270	0	14
Non dense vegetation	0	0	0	325	0
Forest	0	4	15	0	289

Table 1. Classification Matrix for the decision tree model

The second method to assess the effectiveness of the prediction model (decision tree) proposed by our tool is the classification matrix. This method aims to determine whether predicted values match the actual values. This is a standard method for evaluation of statistical models.
Table 1 depicts the chart corresponding to the classification matrix. The rows represent predicted values for the decision tree model whereas the columns represent actual values. The classification matrix is assessing results of the prediction model in an easily understandable manner. From Table 1, we note that the percentage of good classification is 92.2%, which demonstrates good performance of the proposed prediction model.

4.3. Association Rules' Model

The proposed tool also supports use of association rules method to analyze RS big data. This method is widely used in machine learning for mining interesting relationships between variables in large data. One important use of this method is to build associations between satellite object features. Discovered association rules help RS interpreters and decision makers discover correlated features of objects and their impact.
Figure 12 depicts the association rules technique used to help identify linked object features which lead to each land cover change. As shown, we obtain different values of land cover change (percentage of changes), different values of attributes involved in changes and confidence in these changes (values near the blue bars).
Figure 13 provides a classification of values for each attribute and its impact for a given change in the type of land cover. This is very important since it allows understanding the link between specific values of land cover changes and values of object features.

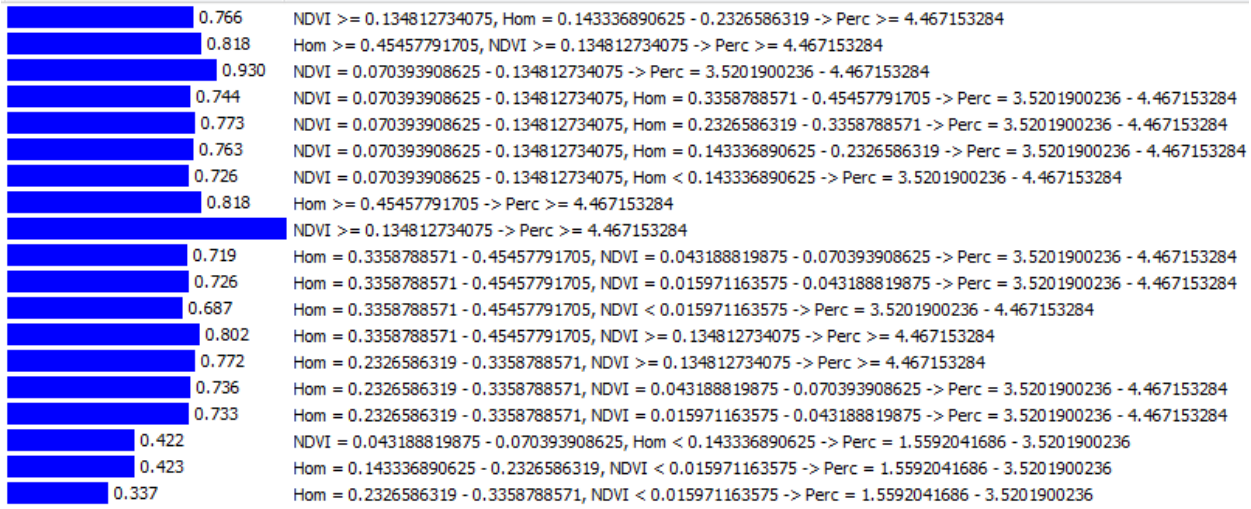


Figure 12. Identifying land cover changes based on an association rules technique

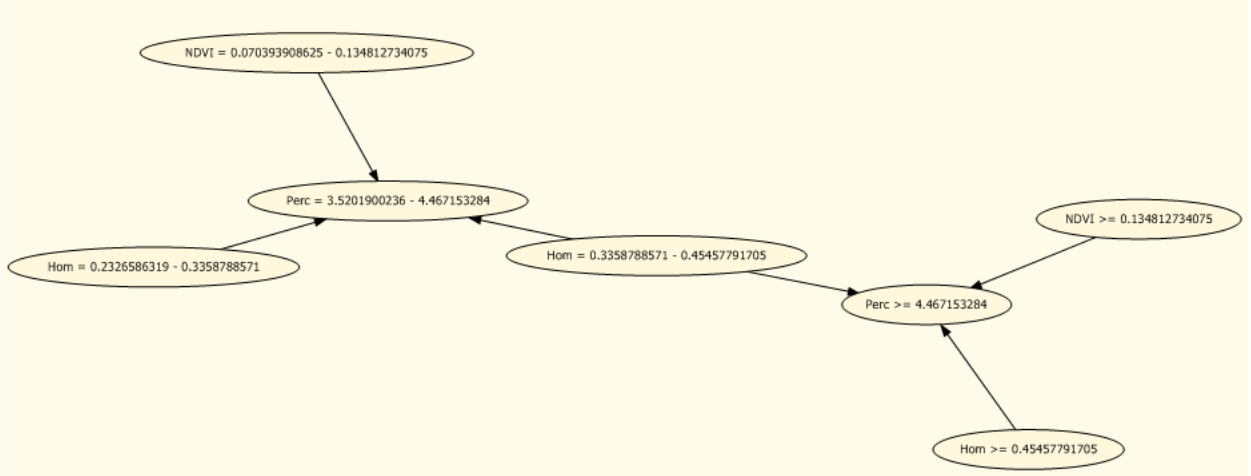


Figure 13. Identification of object land cover types

5. Conclusions and Future Works

In this paper, we proposed a new decision support tool for RS big data analytics. The proposed system offers an environmental dashboard that effectively supports decision making. The proposed tool may have diverse applications in fields of: cartography, resources management and regional planning. The developed tool assists RS users to build descriptive, predictive and prescriptive analytics.

The main challenges for the proposed system were: 1) complexity of RS data, 2) modeling RS big data and 3) analyzing and interpreting RS big data.

To overcome these challenges, a three step approach has been followed: RS data acquisition, modeling, and analyzing and interpretation.

The challenge of analyzing the complexity of RS big data is resolved by proposing an iterative and incremental process of data integration. This is a three-step process (loading image DWH, staging image DWH and final image DWH) which is recommended for complex data.
Additionally, we developed a RS semantic quality process which analyzes and determines the relevance of data.
For the modeling of data, we opted a multidimensional model based on a star schema for the image DWH. We also proposed techniques such as: distribution, indexing and partitioning to enhance querying and retrieval of RS big data.
Three different applications (e.g. classification, decision tree and association rules) of the proposed tool were detailed which depict the multiuse of the proposed tool. It can be adapted to several users' requests and provides efficient decisions in different RS fields. Besides, assessing the decision tree model demonstrates consistently good performance of the developed tool.
Future work will integrate uncertainty modeling in the proposed tool. RS big data is marred by several types of uncertainty which can affect the reliability of provided decision. Imperfection related to RS data can be uncertainty, imprecision, vagueness, conflict, etc. This imperfection is accentuated with the huge amount of big data. Therefore, building reliable decision support systems in RS fields requires modeling imperfection in different stages of the decision support process [43] [48] [49].

References

[1] T. Platt, S. Sathyendranath, Ecological indicators for the pelagic zone of the ocean from remote sensing, Remote Sensing of Environment, vol. 112, pp. 3426-3436, 2008.

[2] A. Bhardwaj, L. Sam, A. Bhardwaj, F. J. Martín-Torres, LiDAR remote sensing of the cryosphere: Present applications and future prospects, Remote Sensing of Environment, vol. 177, pp. 125-143, 2016.

[3] F. E. Fassnacht, H. Latifi, K. Stereńczak, A. Modzelewska, M. Lefsky, L. T. Waser, C. Straub, A. Ghosh, Review of studies on tree species classification from remotely sensed data, Remote Sensing of Environment, vol. 186, pp. 64-87, 2016.

[4] S. Khanal, J. Fulton, S. Shearer An overview of current and potential applications of thermal remote sensing in precision agriculture, Computers and Electronics in Agriculture, vol. 139, pp. 22-32, 2017.

[5] Jinyoung Rhee, Jungho Im, Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data, Agricultural And Forest Meteorology, vol. 237-238, pp. 105-122, 2017.

[6] O. Ait El Mekki, N. Laftouhi, Combination of a geographical information system and remote sensing data to map groundwater recharge potential in arid to semi-arid areas: the Haouz Plain, Morocco, Earth Science Informatics, vol. 9, pp. 465-479, 2016.

[7] Y. Aimaiti, A. Kasimu, G. Jing, Urban landscape extraction and analysis based on optical and microwave ALOS satellite data, Earth Science Informatics, vol. 9, pp. 425-435, 2016.

[8] S. Zhao, Q. Wang, Y. Li, S. Liu, Z. Wang, L. Zhu, Z. Wang, An overview of satellite remote sensing technology used in China's environmental protection, Earth Science Informatics, vol. 10, pp. 137-148, 2017.

[9] J. M. Ramírez-Cuesta, M. Cruz-Blanco, C. Santos, I. J. Lorite, Assessing reference evapotranspiration at regional scale based on remote sensing, weather forecast and GIS Tools, International Journal of Applied Earth Observation and Geoinformation, vol. 55, pp. 32-42, 2017.

[10] Y. Liu, L. Wu, Geological Disaster Recognition on Optical Remote Sensing Images Using Deep Learning, Procedia Computer Science, vol. 91, pp. 566-575, 2016.

[11] M. A. Hoque, S. Phinn, C. Roelfsema, I. Childs, Tropical cyclone disaster management using remote sensing and spatial analysis: A review, International Journal of Disaster Risk Reduction, vol. 22, pp. 345-354, 2017.

[12] D. Talia, Clouds for scalable big data analytics, Computer, vol. 46, no. 5, pp. 98–101, 2013.

[13] G. A. Licciardi, F. Del Frate, Pixel Unmixing in Hyperspectral Data by Means of Neural Networks, IEEE Transactions on Geoscience and Remote Sensing, vol. 49, no. 11, pp. 4163-4172, 2011.

[14] Y. Ma, L. Wang, P. Liu, R. Ranjan, Towards building a data-intensive index for big data computing – A case study of Remote Sensing data processing, Information Sciences, vol. 319, pp. 171–188, 2015.

[15] C. Bodart, H. Eva, R. Beuchle, R. Raši, D. Simonetti, H. J. Stibig, A. Brink, E. Lindquist, F. Achard, Pre-processing of a sample of multi-scene and multi-date Landsat imagery used to monitor forest cover changes over the tropics, ISPRS Journal of Photogrammetry and Remote Sensing, vol. 66, pp. 555-563, 2011.

[16] J. Zhang, T. Li, X. Lu, Z. Cheng, Semantic Classification of High-Resolution Remote-Sensing Images Based on Mid-level Features, IEEE Journal of Selected Topics in Applied Earth Observations and Remote sensing, vol. 9, no. 6, pp. 2343- 2353, 2016.

[17] S. Réjichi, F. Chaabane, F. Tupin, Expert Knowledge-Based Method for Satellite Image Time Series Analysis and Interpretation, IEEE Journal of Selected Topics in Applied Earth Observations and Remote sensing, vol. 8, no. 5, pp. 2138- 2150, 2015.

[18] W. Boulila, I. R. Farah, B. Solaiman, H. Ben Ghézala, Interesting spatiotemporal rules discovery: application to remotely sensed image databases, VINE Journal of Information and Knowledge Management Systems, vol. 41, no. 2, pp. 167-191, 2011.

[19] J. Verbesselt, Big Data: Techniques and Technologies in Geoinformatics, H.A. Karimi (Ed.). CRC Press, Taylor & Francis, London (2014), ISBN: 978-1-466-59651-2, International Journal of Applied Earth Observation and Geoinformation, vol. 35, part B, pp. 368–369, 2015.

[20] W. Boulila, K. S. Ettabaa, I. R. Farah, B. Solaiman, H. Ben Ghézala, Towards a multi-approach system for uncertain spatio-temporal knowledge discovery in satellite imagery, International Journal on Graphics, Vision and Image Processing, vol. 9, no. 06, pp. 19-25, 2009.

[21] L. Moller-Jensen, Classification of urban land cover based on expert systems, object models and texture, Computers, Environment and Urban Systems, vol. 21, pp. 291–302, 1997.

[22] R.T. Kouzes, G.A. Anderson, S.T. Elbert, I. Gorton, D.K. Gracio, The changing paradigm of data-intensive computing, Computer, vol. 42, no. 1, pp. 26-34, 2009.[23] A. Sharifi, Remote sensing and decision support systems, Spatial Statistics for Remote Sensing, Remote Sensing and Digital Image Processing, vol. 1, pp. 243-260, 1999.

[24] J. F. Alcón, C. Ciuhu, W. T. Kate, A. Heinrich, N. Uzunbajakava, G. Krekels, D. Siem, G. D. Haan, Automatic Imaging System With Decision Support for Inspection of Pigmented Skin Lesions and Melanoma Diagnosis, IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 1, pp. 14-25, 2009.

[25] A. M. Alaa, K. H. Moon, W. Hsu, Member, M. V. D. Schaar, Confident Care: A Clinical Decision Support System for Personalized Breast Cancer Screening, IEEE Transactions on Multimedia, vol. 18, no. 10, pp. 1942-1955, 2016.

[26] Y. Heb, B. Ai,Y. Yao, F. Zhong, Deriving urban dynamic evolution rules from self-adaptive cellular automata with multi-temporal remote sensing images, International Journal of Applied Earth Observation and Geoinformation, vol. 38, pp. 164–174, 2015.

[27] L. Dempere-Marco, X. P. Hu, S. L. S. MacDonald, S. M. Ellis, David M. Hansell, G. Z. Yang, The Use of Visual Search for Knowledge Gathering in Image Decision Support, IEEE Transactions on Medical Imaging, vol. 21, no. 7, pp. 741-754, 2002.

[28] J. W. Hwangbo, K. Yu, Decision Support System for the Selection of Classification Methods for Remote Sensing Imagery, KSCE Journal of Civil Engineering, vol. 14, no. 4, pp. 589-600, 2010.

[29] F. Ai, L. K. Comfort, Y. Dong, T. Znati, A dynamic decision support system based on geographical information and mobile social networks: A model for tsunami risk mitigation in Padang, Indonesia, Safety Science, vol. 90, pp. 62–74, 2016.

[30] E. H. Fegraus, I. Zaslavsky, T. Whitenack, J. Dempewolf, J. A. Ahumada, K. Lin, S. J. Andelman, Interdisciplinary Decision Support Dashboard: A New Framework for a Tanzanian Agricultural and Ecosystem Service Monitoring System Pilot, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 5, no. 6, pp. 1700-1708, 2012.

[31] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, W. Jie, Remote sensing big data computing: Challenges and opportunities, Future Generation Computer Systems, vol. 51, pp. 47–60, 2015.

[32] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, G. Trianni, Recent

advances in techniques for hyperspectral image processing, Remote Sensing of Environment, vol. 113, pp. S110–S122, 2009.

[33] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, A. Plaza, On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 10, pp. 4634- 4646, 2015.

[34] R. Giachetta, A framework for processing large scale geospatial and remote sensing data in MapReduce environment, Computers &Graphics, vol. 49, pp. 37–46, 2015.

[35] M. M. U. Rathore, A. Paul, A. Ahmad, B. W. Chen, B. Huang, W. Ji, Real-Time Big Data Analytical Architecture for Remote Sensing Application, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 10, pp. 4610-4621, 2015.

[36] Z. Sun, H. Zou, K. Strang, Big Data Analytics as a Service for Business Intelligence, Open and Big Data Management and Innovation, Lecture Notes in Computer Science, vol. 9373, pp. 200-211, 2015.

[37] M. Minelli, M. Chambers, A. Dhiraj, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, Wiley Publishing, 2013.

[38] Z. K. Malik, A. Hussain, J. Wu, An online generalized eigenvalue version of Laplacian Eigenmaps for visual big data, Neurocomputing, vol. 173, pp. 127–136, 2016.

[39] W. Boulila, I. R. Farah, K. Saheb Ettabaâ, B. Solaiman, H. Ben Ghézala: Spatio-Temporal Modeling for Knowledge Discovery in Satellite Image Databases, *CORIA* COnférence en Recherche d'Information et Applications, Sousse, Tunisia, pp. 35-49, 2010.

[40] J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp. 281–297, 1967.

[41] I. R. Farah, W. Boulila, K. Saheb Ettabaa, B. Solaiman, M. Ben Ahmed, Interpretation of multi-sensor remote sensing images: Multi-approach fusion of uncertain information, TGRS IEEE Transaction on Geoscience and Remote Sensing, vol. 46, no.12, pp. 4142 – 4152, 2008.

[42] W. Boulila, I. R. Farah, K. S. Ettabaa, B. Solaiman, and H. B. Ghézala, A data mining based approach to predict spatiotemporal changes in satellite images," International Journal of Applied Earth Observation and Geoinformation, vol. 13, no. 3, pp. 386–395, 2011.

[43] A. Ferchichi, W. Boulila, I. R. Farah, Propagating aleatory and epistemic uncertainty in land cover change prediction process, Ecological Informatics, vol. 37, pp. 24-37, 2017.
[44] Microsoft: Data quality services, SQL Server 2012 books online,

http://msdn.microsoft.com/en-us/library/ff877925.aspx

[45] A. Leonard, M. Masson, T. Mitchell, J. M. Moss, M. Ufford, Data Cleansing with Data Quality Services, SQL Server 2012 Integration Services Design Patterns, Apress publishing, pp. 101-122, 2012.

[46] A. Leonard, M. Masson, T. Mitchell, J. M. Moss, M. Ufford, Data Correction with Data Quality Services, SQL Server 2012 Integration Services Design Patterns, Apress publishing, pp. 101-123, 2014.[47] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, John Wiley & Sons, 2013.

[48] W. Boulila, A. Bouatay, I. R. Farah, A Probabilistic Collocation Method for the Imperfection Propagation: Application to Land Cover Change Prediction, The Journal of Multimedia Processing and Technologies, vol. 5, no.1, pp. 12-32, 2014.

[49] A. Ferchichi, W. Boulila, I. R. Farah, Towards an uncertainty reduction framework for land-cover change prediction using possibility theory, Vietnam Journal of Computer Science, vol. 4, no. 3, pp. 195-209, 2017.