Using Data Mining for Predicting Cultivable Uncultivated Regions in the Middle East

Ahsan Abdullah[1], Ahmed Bakhashwain[2], Abdullah Basuhail[3], and Ahtisham Aslam[4]
[1]Department of Information Technology, King Abdulaziz University, Saudi Arabia
[2]Department of Arid Regions Agriculture, King Abdulaziz University, Saudi Arabia
[3]Department of Computer Science, King Abdulaziz University, Saudi Arabia
[4]Department of Information Systems, King Abdulaziz University, Saudi Arabia

Abstract: *Middle-East region is mostly characterized by a hot and dry climate, vast deserts and long coastlines. Deserts cover large areas, while agricultural lands are described as small areas of arable land under perennial grass pastures or crops. In view of the harsh climate and falling ground-water level, it is critical to identify which agriculture produce to grow, and where to grow it? The traditional methods used for this purpose are expensive, complex, prone to subjectivity, risky and are time-consuming; this points to the need of exploring novel IT techniques using Geographic Information Systems (GIS). In this paper, we present a data-driven stand-alone flexible analysis environment i.e., Spatial Prediction and Overlay Tool (SPOT). SPOT is predictive spatial data mining GIS tool designed to facilitate decision support by processing and analysing agro-meteorological and socio-economic thematic maps and generating crop cultivation geo-referenced prediction maps by predicative data mining. In this paper, we present a case study of Saudi Arabia by using decade old wheat cultivation data, and compare the historically uncultivated regions predicted by SPOT with their current cultivation status. The prediction results were found to be promising after verification in time and space using latest satellite imagery followed by on-site physical ground verification using GPS.*

Keywords: *Data mining, image processing, GIS, prediction, wheat, alfalfa.*

Received October 30, 2015; accepted June 1, 2016

1. Introduction

Most of the Middle-East region is characterized by a hot and dry climate and the presence of vast deserts and long coastlines [33]. Deserts cover large areas while agricultural lands are defined as the arable land that are under perennial grass pastures or crops, but covering small areas [19]. As per World Bank, the population growth rate in the Middle East is among the highest in the world, surpassing India and China during 1990-2008. Kingdom of Saudi Arabia (KSA) is no exception, and experienced 20% increase in population during the last decade (www.cdsi.gov.sa). This increase in population has resulted in an increase in demand for food and corresponding consumption. One option to meet this increase in food demand is to increase the production locally. However, when we look at this option, we observe an alarming situation, i.e. a sharp decline in arable land in the Kingdom--10% over the last decade [8]. In addition to this, different climatic factors and parameters e.g., soil type, elevation, humidity level; temperature (which is 41-48 degrees Celsius in the summer) and annual rainfall (which is only 100-150 mm) are also huge hurdles in locally increasing the production. With these parameters and climatic factors, using traditional predictive techniques is not enough for decision support for the Middle-East region in general and KSA in particular. Therefore, to address the combined problems of the decrease in arable land and the difficulties associated with cultivation due to the extreme climatic factors, we need to develop novel data-driven Expert System (ES) and Decision Support Systems (DSS) that can be used to identify currently uncultivated regions that are suitable for cultivation.

Several methodologies have been proposed in the past and software solutions developed to address the aforementioned problems. For example, in [1] authors propose the use of satellite images and classification of these images based on supervised and unsupervised learning and then to use them to extract viable parameters to make predictions and forecasting. In [23], authors presented an image processing and image mining approach for image feature extraction and then perform data mining on extracted features to make predictions. Similarly, another object and feature extraction based approach is discussed in [30], the authors propose the extraction of objects from image maps and then perform data miming on extracted features.

In addition to these theoretical approaches, different software solutions proposed to solve the described problem. For example, for the classical Geographical Information System (GIS) tools such as QGis or ArcGIS plugins proposed for performing prediction/classification, however, there are shortcomings too. For example, in [36] a Markov-Kalman filter model used for simulation of the dynamic changes of land use forecasting and a plug-in developed for use with QGis. Although, simplicity is one of the strengths of Markov chain models but this is also one of its biggest weaknesses. Markov chains are projection models i.e., they are not sensitive to policy changes, thus, it is not easy to incorporate the range of policy variables that could be interesting while considering their impact (say) for predicting cultivation. The SPACE extension for the popular GIS tool ArcGIS ver 10 is available for free; but ArcGIS is a commercial product, furthermore, SPACE is designed for crime analysis and forecasting instead of generating agriculture prediction maps as our proposed tool does.

Most of these methodologies and software solutions have their inherent strengths and weaknesses. Some of the foreseen limitations of existing solutions are as follows:

- Limited or no legend based data extraction from individual as well as overlaid maps.
- Regeneration of prediction maps from existing map datasets.
- Data extraction, forecasting and data mining support in one platform.
- Lack of support for some commonly used raster image formats (e.g., .bmp, .jpg, .tiff etc.,).
- Most of these excellent tools have a steep learning curve, thus requiring special time-intensive training.

To address the limitations in existing work, we propose and present the following contributions of our work:

- Support of data extraction from raster image maps based on maps legends and storing as text files.
- Read-Write Support for all common image formats (e.g., .bmp, .jpg, .tiff etc.,).
- Map overlay support to assign weights and predict combined effect of different parameters and generating prediction maps.
- Read-Write capability in CSV (comma separated values) allowing “interfacing” with popular data mining tools such as Weka, Orange and more.
- Easy to use intuitive Graphical User Interface (GUI) with fewer operational steps.

In this paper, we present our Spatial Prediction and Overlay Tool (SPOT) that endeavours to address the aforementioned limitations of the existing work by providing an integrated solution of proven information technology techniques of Data Mining, Image Processing and GIS. Using SPOT, firstly we ‘train’ the system by using the existing agro-meteorological data and then that particular predictive data-mining technique is used to predict suitable cultivation regions for a crop using selected parameters, the results are subsequently displayed in a GIS environment. Although in this paper we demonstrate our tool for predicting wheat cultivable regions, but depending on the available data and in view of crop rotations, the tool is flexible enough to be used for predicting cultivation regions for different crops, including fodder and pasture grass.

The remainder of the paper is organized as follows: section 2 provides the necessary background and key definitions. In section 3 we cover the related work from the perspectives of methodologies, theories and tools. Section 4 is materials and methods that describes the architecture and system details of the proposed SPOT tool and focusing on data processing and related issues. In section 5, we discuss the validation techniques and SPOT results of using real data. Finally, we conclude our work in section 6.

2. Background

In this section, we will discuss the favorability function [11] which forms the basis of the process of our prediction map generation, followed by overview of data mining and predictive data mining techniques.

In the end, a brief overview of the regional agriculture is presented.

2.1. Favorability Function

Consider developing a cultivation potential map to identify the areas that are suitable for future cultivations. The spatial distribution of regions to be selected for future cultivations is then defined as the target pattern. The area for which we want to have information on the target pattern is defined as the study area. The target pattern is defined as an unobservable and binary spatial pattern i.e., as per present data cultivation does not exist. While we may know the features and sites of past cultivations in the study area, but these locations are not considered to be part of the target pattern as per the definition.

Although the target patterns are the unknowns that we wish to know, however, we can generate the prediction images for the target patterns using prediction models. The prediction models are thus defined as the logical or analytical procedures utilized to create prediction images for the given target patterns. The prediction models range from simple ad-hoc procedures, to complex mathematical or statistical techniques (predictive data mining), and to models that are based on an expert’s qualitative knowledge (Fuzzy inference). We endeavour to generate prediction images, which “approximate” as close as possible to the unknown and unobservable target patterns. Thus, a prediction pattern is defined as a thematic representation or a map of the prediction image to visualize and realize the prediction results.

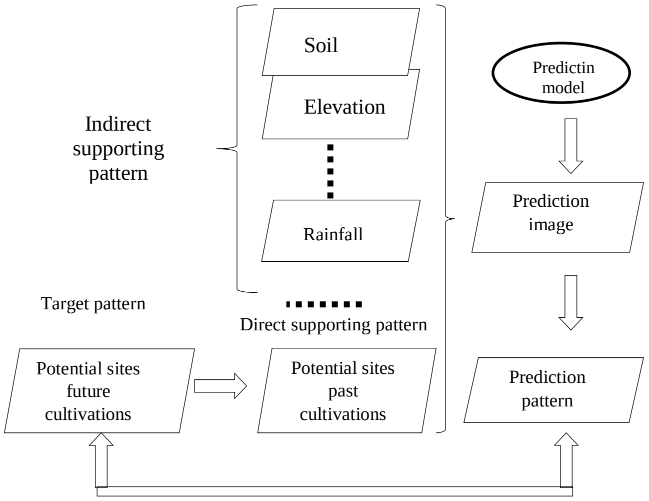

Soil
Elevation
Rainfall
Indirect supporting pattern
Predictin model
Prediction image
Target pattern
Direct supporting pattern
Potential sites future cultivations
Potential sites past cultivations
Prediction pattern

Figure 1. Favorability function approach showing the propositional relationships among different definitions related to generating prediction pattern.

Our approach is based on starting from parametric maps all the way to generating prediction pattern. This approach is based on and rooted at the same needs and procedures that are followed by domain experts to generate thematic maps using spatial data. More formally, the approach of prediction modelling is the “favorability function” approach of [11] and is shown in Figure 1. The favorability function approach is based on two assumptions i.e., 1) the target pattern can be categorized based on the agro-meteorological and socio-economic spatial data, and 2) the incidents related to the target pattern occur in future under similar agro-meteorological and socio-economic conditions.

2.2. Data Mining

Data mining is knowledge discovery in databases. Unlike locating data of interest in a database (say) using Structured Query Language (SQL), data mining is about finding what you do not know that you do not know. There are different methods for data classification that can be used for predictive data mining: Decision Trees (DT), Rule-Based Methods, Logistic Regression (LogR), Linear Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Artificial Neural Networks (ANN), Linear Classifier (LC) and so forth [18, 34, 35]. Some examples of popular data mining tools being Weka, RapidMiner, Knime, Orange, Oracle Data Miner and more. Various classification methods have shown different efficacies and accuracies based on the kind of data sets that were used [20].

2.3. Predictive Data Mining

Prior to selecting a prediction technique for coding in SPOT, number of prediction and predictive data-mining techniques were evaluated. Overview of some of these techniques is given here.

GIS and Multi Criteria Evaluation (MCE) techniques based tools are mostly used to solve the site selection problems. However, GIS and MCE have their own limitations and cannot be used mutually exclusively to get the best solution. GIS is mainly used for physical spatial fitness analysis, while MCE usually deals with analyzing decision problems and assessing different options based on the decision maker’s preferences.

Integration of GIS with MCE has been done through loose and tight coupling techniques, but has many drawbacks and limitations [16]. In [15] a GIS-based multicriteria evaluation site selection tool is reported that was developed for ArcGIS 9.3 using COM technology to achieve software interoperability.

In our study, we have however, developed the SPOT tool from scratch, though integrating SPOT with ArcGIS could be part of future work.

Markov chain predictions are based on the current state and subsequent transitions from this state.

Therefore, depending on the current state, it is required to exactly match a known state in the model. Thus, if the current state has not been reached, it is not possible to have transitions from that state and no future states can be reached i.e. predictions cannot be made. SVMs don’t perform well when the features space is greater than the number of samples.

Furthermore, SVM provide direct probability requiring additional cross-validation methods [9]. C45 rules don’t scale well for ‘noisy’ and large datasets [28]. Similarly, the major shortcoming of J48 is its run-time complexity which increases with increase in depth of the corresponding ‘tree’ [37]. The reliability of Decision Trees is dependent on data quality i.e. the robustness is not one of the strengths of Decision Trees. Even a small change or ‘error’ in the input data can cause large variations in the corresponding results.

Another major limitation of decision trees is that the decisions/results are based on expectation and a small change in a rational expectation can lead to the major variation in results [25]. Regression could not be considered suitable for prediction because the technique only considers linear relationships between dependent and independent variables, which is hardly the case for complex real-life problems that may result in unrealistic outcomes [22].

After comparison with these predictive data mining techniques, we found weighted Naïve Bayesian to be a superior prediction technique, mainly because the technique does not have the issues associated with these techniques and also caters to specialized unsupervised learning. Furthermore, Bayesian techniques provide formalism of thought under ambiguous conditions, such that degrees of belief are represented as numerical parameters, which are subsequently combined as per rules of probability theory.

2.4. Spatial Data Mining

Spatial data mining is the use of formal data mining methods and procedures for knowledge discovery in spatial data sets. In other words, spatial data mining is applying data mining to spatial models. In spatial data mining, decision makers use spatial or geographical data to extract business intelligence and related results.

The main distinction between data mining in spatial databases and relational databases is the impact of the neighboring attributes of some of the objects of interest, which may influence the objects under consideration and are therefore also required to be considered.

Because of the categorical requirements of location and extension of spatial objects, inherent relationships among spatial neighborhoods (e.g., distance, topological, and bearing relationships) are used in spatial data mining algorithms. Thus, novel techniques and tools are required to effectively and efficiently mine spatial data.

2.5. Regional Agriculture

As per the earth policy institute (www.earth-policy.org), grain production is stagnating across the Arab world, yet grain demand is growing rapidly as population expands. Since 1960, the region's population has nearly quadrupled to 360 million. By 2050 the region is projected to add another 260 million people, dramatically increasing pressure on already stressed land and water resources. This is one of the main reasons why we have considered wheat in this study; furthermore, the wheat-alfalfa cycle confirms the cultivation of fodder in the predicted regions for raising livestock.

The critical decision of selecting the best crop to grow can be made by taking into consideration multiple factors such as the management of soil fertility, water, land use, the climate's response to available inputs as well maintaining the nutrient content to maximize the productivity, just to name a few. In the context of water for example, farmland may be used for the production of drought-tolerant wheat varieties (Al-gaimi, Sindy-1, Sindy-2, and Hab-Ahmar) grown in the Kingdom.

Similarly, soil can be protected and nurtured to ensure its long-term productivity and stability. The objective of our proposed solution is also to facilitate decision-makers in arriving at such decisions using a data-driven approach and use prediction when considering the long-term perspective.

3. Related Work

In this section, we will discuss some existing work that describes the use of IT and its related technologies to address agriculture-related problems. The reviewed prior work is divided into two sub-sections i.e., the methodologies and the tools.

3.1. Methodologies

A comparison of four classification methods to extract land use and land cover from raw satellite images (specifically focusing on Saudi Arabia) is discussed in [1]. In this work, authors propose the use of satellite images to extract varying parameter data for remote arid areas in Saudi Arabia. In this work, supervised and unsupervised learning methods are used for the classification of maps. An important issue with this work is that authors propose the visual comparison of developed maps to evaluate the accuracy of results, which may not be suitable in the long run.

A methodology for image processing and object oriented classification of satellite images for agriculture is discussed in [30]. The focus of proposed methodology is on defining objects from images and then exhaustively extracting set of features from every object. Features extracted from different types of information e.g. spectral, textural, structural, shape etc. Like the work discussed in [23] this approach loses the data existing at pixel level instead of features and object level.

In [2] the problem of identifying a new landfill is discussed using an intelligent system based on fuzzy inference. Several factors are considered in the process of identification of the suitable site, such as topography and geology, natural resources, socio-cultural aspects, economy and safety. The system has the ability to accommodate new information on the landfill site selection by updating its knowledge base.

The system subsequently ranks the sites on a scale of 0–100%, with 100% being the most appropriate one.

Unlike our proposed system which works unsupervised, fuzzy inference involves supervised learning in the form of writing IF…THEN fuzzy rules which involves subjectivity.

In [32], a joint solution of DSS and ES is presented that can help small-holding farmers in decision-making by minimizing the impact of climactic risk factors and assisting farmers by suggesting the best possible options for improving crop productivity. The proposed solution also enables the stakeholders to share this agricultural information among themselves (by using mobile connectivity) and to keep the system updated with the prevailing climactic conditions and weather factors. Taking into consideration the impact of different parameters, the DSS uses a rule-based approach for prediction and shares the prediction results and suggestions with the end-users.

The DSS in [13] takes into account different factors such as weed control, pests, crop selection, soil factors (e.g., soil flora and soil fauna), etc., and helps in decision-making to handle conventional and genetically-modified crops. Instead of going to the micro-level functions of individual species, soil-fauna modelling was done at the community level by making use of machine- learning methods such as regression trees, model trees and linear equations. Finally, the proposed system was used for the prediction of climactic factors and their possible effects on crop production.

3.2. Tools

Geographic Resources Analysis Support System (GRASS) GIS [4, 26, 27] is an open source software suite which provides lot of useful features and functions in domain of geospatial data management and analysis, image processing, graphics and map production, spatial modelling and visualization. GRASS provides modules for maps creation, 3D and 2D maps visualization and maps overlay and different image processing operations (e.g., edge detection, image rectification, resolution enhancement, image rectification etc.,). Even though with numerous features, unlike SPOT GRASS lacks the datasets extraction features to extract datasets from different sources (e.g., agricultural maps) that can be used as input for different data mining tools as training and evaluation datasets. Applying data mining techniques on input data sets result in output datasets which can be visualized by text to map conversion. The GRASS set of tools also lack such kind of features supported by SPOT.

eStation [3] is a data collection and processing service designed and developed to automatically deal with receiving, processing, analysing and dissemination of key environmental parameters (e.g., rainfall, active fires, water index, land surface temperature, vegetation index etc.,). eStation is limited in its capacity to processing the identified environmental variables only.

This approach does takes into consideration the data that is available in the form of digital maps. Even though extensive data on different environmental variables (e.g., rainfall, temperature, soil type, humidity, elevation etc.,) is available as. Joint Photographic Experts Group (JPG) digital maps, but extracting this data from maps into usable form to make precise predictions is a difficult task; which SPOT is designed to do due to its unique image processing features.

SPIRITS is another tool for processing image time series data for agriculture monitoring [14].

SPIRITS is specifically designed for environmental monitoring for crop production analysis so as to help decision and policy makers. SPIRITS makes use of remote sensing image time series data, estimates the impact of anomalies on the crop production and shares the analysis report with concerned stakeholders for decision-support. SPIRITS also helps in extracting useful statistics from images, stores them in internal database and then use to create graphs. Three main limitations of this SPIRITS as compared to SPOT are 1) does not support the common image formats (e.g., .bmp, .jpg, .tiff etc.,) and most of the agricultural maps are available in any one of these formats and 2) does not performs image to table transformation and 3) does not performs prediction; this being the major theme of our work.

In [31], the authors present an intelligent system FAUSY that can be used to capture the relationship between environmental variables and sap-flow measurements. The system makes use of the “fuzzy algorithm” to extract relationships from a set of input-output environmental observations and applies algorithmic techniques for learning and forecasting the predicted results via a simulated model. These predicted results help farmers and policy-makers in planning future programmes of cultivation. Unlike SPOT does not performs predictive data mining using “dirty” raster maps.

Agro Genius [17] is an Expert System designed to help different stakeholders at Informatory and Advisory Levels. At the Informatory Level, users can obtain static information about the advantages and disadvantages of organic and inorganic cultivation under certain weather conditions. At the Advisory Level, users can interact with the system in the form of a question-answer session, resulting in a flow chart-like tree structure. The generated ‘tree’ shows nodes, leaves, etc. as climactic effects, weather forecasts, possible crops along with the possibility of their successful growth under these climactic conditions.

WinDisp [7] is a software package that can be used for the display and analysis of satellite images, maps and related databases. It is also equipped with basic functionality for early warning for food security.

WinDisp can also be used for multiple images comparison, extract and graph trends from satellite images. From overlay formation point of view, this tool does not provide comprehensive functionality of maps overlay and for assigning weights to different maps in overlay formation for analysis by inspection as SPOT does. Furthermore, WinDisp is also limited in data extraction from the overlaid maps which is main aspect of forecasting the future based on past data.

A visual information solution that consists of suite of image processing software is provided by EXELIS [5]. Jagwire [6] is a web-based geospatial data management solution that can help domain experts to search data, transform it into information and make it available for concerned stakeholders regardless of their physical location. This software is very useful for efficient and quick sharing of data between distributed stakeholders but is not very effective when the target is extraction of right data from different data sources (e.g., maps) followed by predictive data mining as per SPOT.

We believe that our idea, it’s implementation and subsequent verification of results by ground-truthing is indeed novel, because as per literature review, unlike our work there is hardly any indigenous application work reported for the Middle East region using a tool with built-in predictive data mining capability like ours while taking into consideration the JPG vs. TIF format legend issues in the public domain data.

4. Materials and Methods

The architecture of the SPOT tool is shown in Figure-2. The figure shows the main components of the tool. Very briefly, Modules 1 and 6 (Load and Save) allows conversion of raster maps into Comma Separated Values (CSV) and vice-a-versa for reading and saving files in different formats. Furthermore, CSV formats allows mathematical and statistical operations that are not possible on raster data. In the current paper, we discuss the prediction engine i.e., Module 5 in detail.

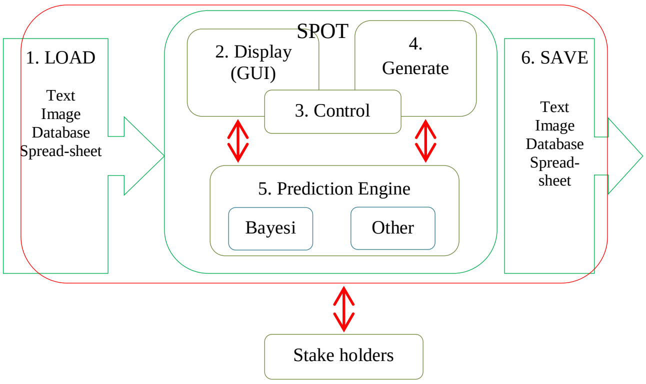


Figure 2. Architecture of the SPOT tool showing the main modules.

4.1. Prediction Engine

This section is centred on the seminal paper [21]. The Naïve Bayesian (NB) learning uses Baye's theorem to calculate the most likely class label of the new unclassified instance. Let a_1, a_2, · · ·, a_n be feature (or parameter) values of a new unclassified instance. Let *C* to be the feature that characterizes the class value, and *c* corresponds to one of the values of *C*. A new i.e., unclassified instance *d* is classified based on the maximum posterior probability. Formally, class assignment of *d* is expressed as follows in Equation (1):

$$V_{map}\ (d) = arg\,max_c\ P(c)P(a_1,a_2,...,a_n\,|\,c) \tag{1}$$

In NB learning, all parameters or features are considered to be independent for the given class value, therefore, as in Equation (2)

$$P(a_1,a_2,...,a_n\,|\,c) = \prod_{i=1}^{n} P(a_i\,|\,c) \tag{2}$$

The maximum posterior classification of NB is given as in Equation (3)

$$V_{nb}(d) = arg\,max_c\ P(c)\prod_{i=1}^{n} P(a_i\,|\,c) \tag{3}$$

The assumption that all features are independent is clearly almost always wrong. There are two main approaches to relax the assumption of independence in NB learning. The first method of relaxing this notion consists of selecting a feature subset in data. The predictive accuracy of NB is known to improve by eliminating unneeded or highly correlated parameters.

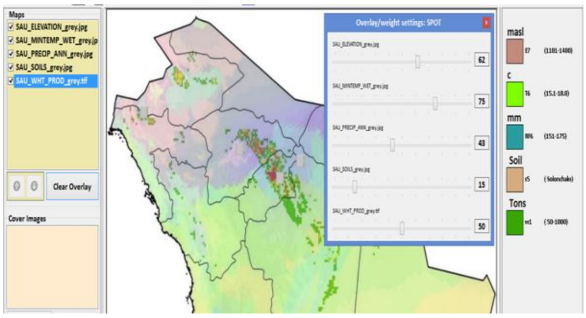
This is understandable as the correlated features violate the notion of naivety i.e., feature independence. Feature subset selection applied to NB learning is formally defined as in Equation (4):

$$V_{fsnb}(d, I(i)) = arg\,max_c\ P(c)\prod_{i=1}^{n} P(a_i\,|\,c)^{I(i)} \tag{4}$$

Where *I*(*i*) ∈{0,1}, assigning weights to features for NB inference is another way of relaxing the feature independence assumption. Assigning real valued weights to each parameter or feature is more generic as compared to feature deselection/selection. Thus, feature selection can be considered as a special case of feature weight assignment, where only binary weights are used i.e., 0 or 1. The NB classification with feature weight assignment can thus be reformulated as

$$V_{fwnb}(d, w(i)) = arg\,max_c\ P(c)\prod_{i=1}^{n} P(a_i\,|\,c)^{w(i)} \tag{5}$$

Where w(*i*) ∈ $\Re^+$, In Equation-5, unlike traditional NB, each feature or parameter *i* has a specific weight *w(i)*. The weight *w*(*i*) can be a positive real number that represents the importance of feature *i*. Because feature weight assignment is a general representation of feature selection, therefore, this comprises a considerably larger solution search space as compared to just feature selection.



Overlay/weight settings: SPOT

- ☑ SAU_ELEVATION_grey.jpg
- ☑ SAU_MINTEMP_WET_grey.jp
- ☑ SAU_PRECIP_ANN_grey.jpg
- ☑ SAU_SOILS_grey.jpg
- ☑ SAU_WHT_PROD_grey.tif

SAU_ELEVATION_grey.jpg 62

SAU_MINTEMP_WET_grey.jpg 75

SAU_PRECIP_ANN_grey.jpg 43

SAU_SOILS_grey.jpg 15

SAU_WHT_PROD_grey.tif 50

Figure 3. SPOT interface showing parameter maps loaded with variable transparency control and legend values.

Our SPOT tool allows analysis by assigning weights to the overlaid parameter maps i.e. controlling map transparency as shown in Figure 3. Here the transparency (t_i) values for the parameters of elevation, min wetness temperature, precipitation, soil type and wheat variety being t_1=62%, t_2=75%, t_3=43%, t_4=15% and t_5=50%, respectively. For NB inference, SPOT subsequently converts the transparency values to feature weights as follows:

$$w(i) = \frac{t_i}{\sum w(i)} \text{ here } t_i \in \mathrm{T} \tag{6}$$

Here *T* is possible transparency weights and t_i is one such transparency weight for the i^{th} parameter map.

Thus the feature weights corresponding to the parameter transparency values considered in Figure 3 as per Equation (6) will be w_1=27.31, w_2=30.61, w_3=17.55, w_4=6.12 and w_5=20.41.

4.2. Classes and their Description

The SPOT application developed using C# (.NET 4.5 framework).

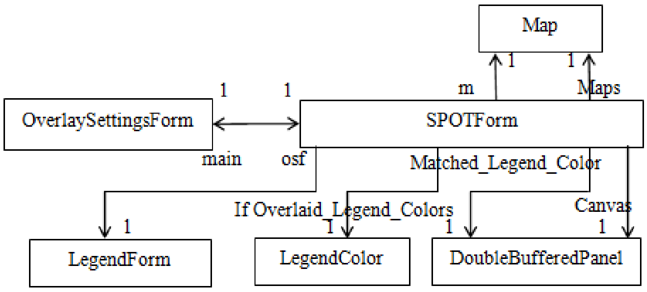


Figure 4. Class diagram SPOT (association).

Figure 4 is the Class diagram (Association) that shows all the classes maintained and manipulated in SPOT along with their corresponding associations.

Since SPOT is a GUI based windows application, therefore, does not requires an extended object oriented approach. The associations in the diagram show SPOTForm class used along with fields from LegendForm, LegendColor, DoubleBufferedPanel and Map classes.

4.3. Fields, Methods and their Description

The SPOTForm is the main container that provides all GUI control of the application. Among rest of the classes only OverlaySettingsForm class has used a field of SPOTForm class. Names of the classes, fields and methods are self-explanatory.

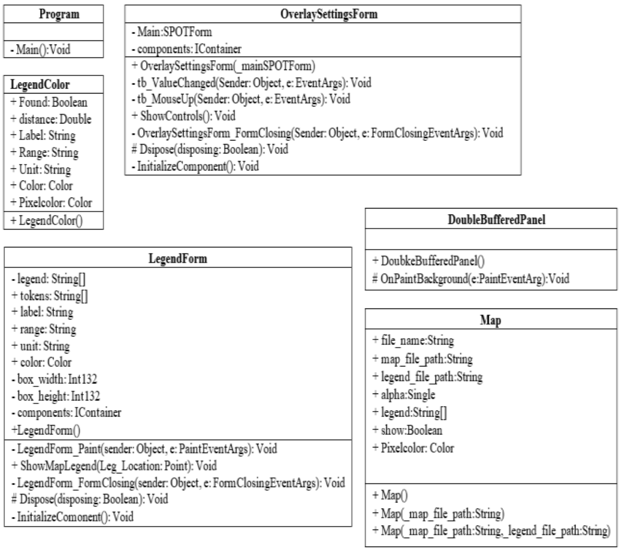

Program
- Main():Void
OverlaySettingsForm
- Main:SPOTForm
- components: IContainer
+ OverlaySettingsForm(_mainSPOTForm)
- tb_ValueChanged(Sender: Object, e: EventArgs): Void
- tb_MouseUp(Sender: Object, e: EventArgs): Void
+ ShowControls(): Void
- OverlaySettingsForm_FormClosing(Sender: Object, e: FormClosingEventArgs): Void
# Dsipose(disposing: Boolean): Void
- InitializeComponent(): Void
LegendColor
+ Found: Boolean
+ distance: Double
+ Label: String
+ Range: String
+ Unit: String
+ Color: Color
+ Pixelcolor: Color
+ LegendColor()
DoubleBufferedPanel
+ DoubkeBufferedPanel()
# OnPaintBackground(e:PaintEventArg):Void
LegendForm
- legend: String[]
+ tokens: String[]
+ label: String
+ range: String
+ unit: String
+ color: Color
- box_width: Int132
- box_height: Int132
- components: IContainer
+LegendForm()
- LegendForm_Paint(sender: Object, e: PaintEventArgs): Void
+ ShowMapLegend(Leg_Location: Point): Void
- LegendForm_FormClosing(sender: Object, e: FormClosingEventArgs): Void
# Dispose(disposing: Boolean): Void
- InitializeComonent(): Void
Map
+ file_name:String
+ map_file_path:String
+ legend_file_path:String
+ alpha:Single
+ legend:String[]
+ show:Boolean
+ Pixelcolor: Color
+ Map()
+ Map(_map_file_path:String)
+ Map(_map_file_path:String,_legend_file_path:String)

Figure 5. Fields and methods of SPOT.

Figure 5 shows the fields and methods of each class except SPOTForm. Figure 5 shows the field and/or methods for OverlaySettingsForm, LegendColor, DoubleBufferedPanel, LegendForm, Map and Program with main ( ) method.

Descriptions of methods for all classes (including Map, LegendColor, LegendForm, OverlaySettingsForm and DoubleBufferedPanel) are provided in corresponding tables, for example Tables 1 and 2.

Table 1. Methods of Class “Map”.

Methods	Description
Map (+2 Overlaid)	1. Without specifying image and legend. 2. With image only. 3. With both image and legend.

Table 2. Methods of class “LegendForm”.

Methods	Description
Legend Form	Constructor
LegendForm_FormClosing	Function preventing closing of the form on button marked “x”
LegendForm_Paint	Function to Parse each line of the legend file, Extract the color, range, units, label and then Draw the legend box for the corresponding color and display other values.
ShowMapLegend	Takes a point specifying the display position of current window as input, then sets the current window position as per input parameter and then reDraws the legends.

5. Results and Discussion

5.1. Validation of Results

While performing prediction modelling, the utmost important and utterly essential part is performing validation of the prediction results. In the absence of any kind of validation, the prediction model along with the prediction image is useless and hardly has any scientific value. Here we describe some relatively simple validation procedures, so that the prediction results can be inferred unambiguously with respect to the future cultivation potential of the study area.

After obtaining a prediction image, a suitable validation is the one that is based on the assessment across the prediction results and the unknown target pattern i.e., the potential areas identified for future cultivations. This is based on the assumption that the study area consists of past cultivations, which is true in our case. Because the target pattern (the future cultivations) *is* unknown, a direct assessment w.r.t the target pattern is infeasible. Therefore, the next best option is to simulate the assessment by using a part of the past cultivation regions as if they characterise the target pattern. To simulate the assessment, we must limit the use of all data of the past cultivations for the study area. This is achieved by splitting the data into two sets i.e., one subset of data that will be used for acquiring the prediction image and the other subset for comparison with the prediction results for validation.

Observe that, without the data splitting, it would be impossible to validate the prediction results. As far as the interpretation of the prediction results are concerned, the assessment of the prediction results with the un-partitioned past cultivations in the study area does not makes sense. Thus, splitting of the past cultivations data is the foundation of the validation techniques discussed here.

5.1.1. Time-based Validation

It is assumed that we have the complete time and space distribution of the past cultivations for the study area.

While using this time-based partition technique, we are able to estimate the probability of successful future cultivations within a certain time constrain such as "for the next 15 years". Briefly, this validation technique consists of four steps. First step i.e., 1) constructing a prediction model and generating a prediction image using past cultivation data of the study area, 2) using the satellite imagery of the reference year, dividing the past cultivations into two time periods i.e. the ones that have occurred before the reference year and the those that occurred after that year, 3) using prediction of first step and cultivations of the first period only, generating the prediction image and finally 4) compiling statistics, by comparing the prediction results of the third step at the pixel-level with distribution of the occurrences of the final period.

5.1.2. Space-based Validation

It is also assumed that we have the spatial distribution of the past cultivations over the study area considered. Subsequently, the study area is divided into two disjoint sub-areas i.e., area_1 and area_2. One of two sub-areas is selected for constructing a prediction model, while the other sub-area is used for validating the prediction.

While using this space-based partition technique, the current prediction model of the study area can be extended to the surrounding areas or areas with similar agro-meteorological properties [12].

For validation of SPOT results, a combination of time and space-based partition approach is adopted i.e., we use the decade old data for making cultivation predictions. Subsequently, the prediction results in space are compared with the cultivation as of today i.e., time-based with additional level of validation based on actual site visit and ground truth verification. Thus, our validation is in addition to the pixel-based approach.

5.2. Predicting Wheat Cultivation Regions

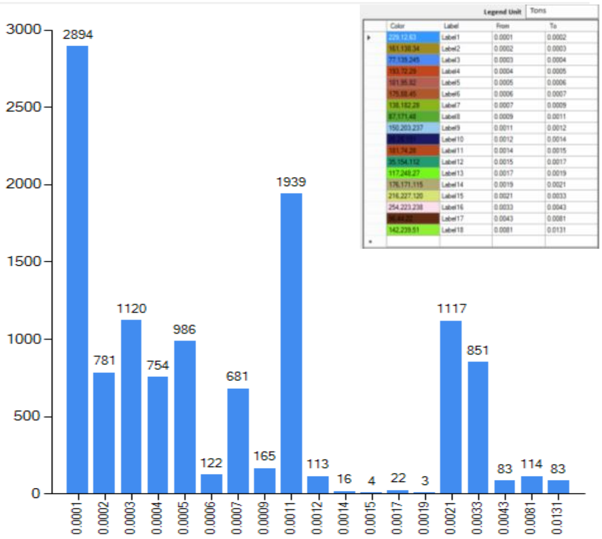
To demonstrate the prediction capability of the SPOT tool, we first use historical geo-referenced parameter maps (JPG format 750x750 pixels at 250 dpi) available at International Wheat and Maize Improvement Center (www.cimmyt.org). Details of the maps considered are as follows:

1. *Elevation*: SRTM Data v.3 CIAT, 2006.
2. *Min Temp Wettest quarter*: WorldClim v. 13, 2004.
3. *Annual Rainfall*: WorldClim v. 13, 2004.
4. *Soil*: FAO classification v. 2.1, 2002.
5. *Wheat*: Spatial Production Allocation Model (SPAM), v. 3, HarvestChoice, 2000.

The five raster maps considered show the location of nine wheat trial sites, these raster maps identify the wheat trial sites by different colors, which we set to a single color as shown by grey colored boxes in Figure 3.

Figure 6 shows the SPOT posterior probability distribution results generated for the cultivation of low yield wheat variety (50-1,000 tons) along with system generated color-codes assigned to the probabilities.

Here the probability distribution is based on equal interval basis [10] with zero results not shown, we subsequently adjust the interval range and color-codes to as per our requirements. Note that the user has the option to perform prediction for any of the five different wheat types/varieties, just by selection from the drop-down menu; not shown here. After adjusting the probability distribution ranges as per range of interest and corresponding color assignment, the prediction map is generated, as shown in Figure 7.



Legend Unit Tons

Color	Label	From	To
229.12.63	Label1	0.0001	0.0002
161.138.34	Label2	0.0002	0.0003
77.139.245	Label3	0.0003	0.0004
193.72.29	Label4	0.0004	0.0005
181.95.82	Label5	0.0005	0.0006
175.88.45	Label6	0.0006	0.0007
138.182.28	Label7	0.0007	0.0009
87.171.48	Label8	0.0009	0.0011
150.203.237	Label9	0.0011	0.0012
[illegible]	Label10	0.0012	0.0014
181.74.28	Label11	0.0014	0.0015
35.154.112	Label12	0.0015	0.0017
117.248.27	Label13	0.0017	0.0019
176.171.115	Label14	0.0019	0.0021
216.227.120	Label15	0.0021	0.0033
254.223.238	Label16	0.0033	0.0043
96.44.22	Label17	0.0043	0.0081
142.239.51	Label18	0.0081	0.0131

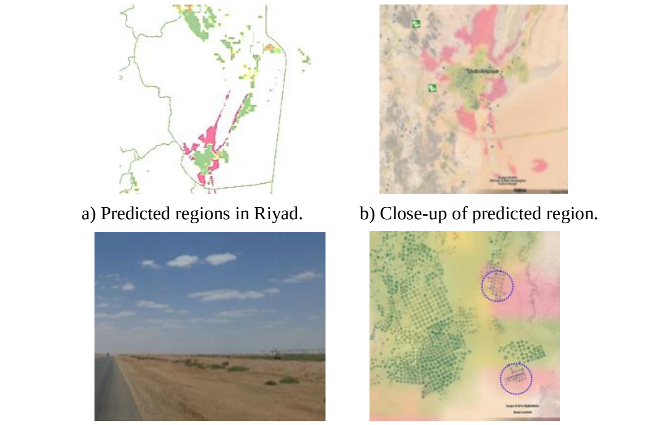
Figure 6. Posterior probability distribution results for low yield wheat variety (50-1,000 tons), here x-axis is probability and y-axis probability count.

We started the experiments with the first legend value i.e., low yield wheat variety (50-1,000 tons) using weights as per Figure 3. The regions predicted to be suitable for that variety are highlighted by pink color mostly for the Riyadh province as shown in Figure 7-a.

The prediction map of Figure 7-a was overlaid on the Satellite image of the KSA using Google Earth, the results are shown in Figure 7-b. From Figure 7-b observe that the predicted regions identified fall in and around the cultivated regions of Wadi Aldwasir.

Figure 7-c is a close-up of Figure 7-b which clearly shows present day cultivation in the pink predicted regions with farms identified by blue dashed circles i.e., validation in time and space. However, the next obvious question is, the cultivation seen in satellite imagery of Figure 7-c is watermelons, or wheat or some other crop? For this purpose ground truth survey was required for which the metalled road of highway 177 was travelled that goes through the Wadi, as shown in Figure 6-d.

Highway 177 passes along the historically cultivated regions and close to the historically uncultivated regions as of year 2000 i.e. the regions now predicted to be suitable for cultivation. Cultivation could be seen, inspected and verified along the highway for the accessible farms. When large private farms along the predicted region on highway 177 were approached, the personnel on duty did not allowed visiting the farms or taking pictures. However, those farm personnel provided confirmed information about the crop being cultivated there; it was alfalfa. This validates the SPOT results and our initial premise in view of the wheat-alfalfa cycle.



a) Predicted regions in Riyad.

b) Close-up of predicted region.

d) Driving on highway 177 in Wadi Aldwasir.

c) Cultivation on predicted regions in Wadi Aldwasir.

Figure 7. Potential wheat cultivation regions identified by SPOT in the riyadh province of the kingdom of saudi arabia.

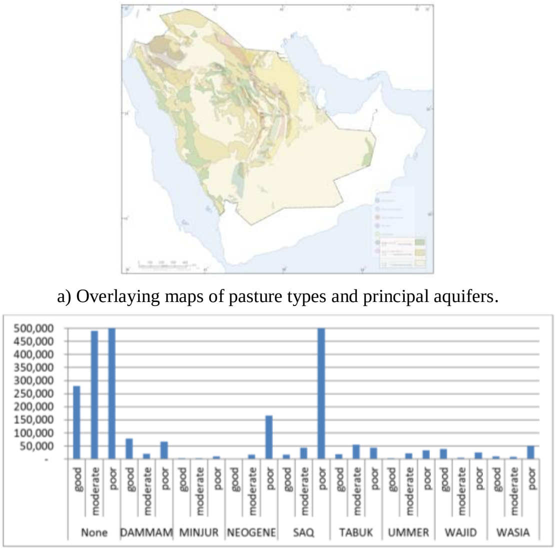
Additional experiments conducted for high yield variety (2,501-4,500 tons); the areas identified in Hail province of KSA, for which the cultivation could only be verified by latest satellite imagery. Because in the absence of a metalled road, those areas were inaccessible without a 4x4 vehicle and required expert driver with sufficient knowledge of the area.

Furthermore, some of the predicted areas were close to animal sanctuaries i.e., although predicted to be suitable for cultivation, but not used for this purpose.

5.3. Impact of Aquifers on Pastures

For this we use SPOT with high resolution latest TIF raster maps with one-to-one matching of map colors with legend colors. Using SPOT the pasture map was overlaid on the principal aquifer map as shown in Figure 8-a with transparency set to 50%.

From Figure 8-a, it can be observed that there is little or no overlap between most of pastures and aquifers, but this level of observed overlap cannot be used for scientific investigation. Therefore, we did raster to CSV conversion of the overlaid maps, loaded the results in a database and quantified this overlap by running SQL group-by queries (legend values not included). Subsequently a graph (two-level histogram) was created based on the query outcome; results are shown in Figure 8-b.



a) Overlaying maps of pasture types and principal aquifers.

b) Grouping of results of overlaid maps of pasture types and principal aquifers (x-axis is class of aquifer and y-axis record count for Figure 8-a.

Figure 8. The impact of principal aquifer on pastures and vegetation communities in KSA.

From Figure 8-b it can be observed that most number of records are for poor pastures, when there are no aquifer i.e., none that count is 1.6 million records and it is 1.8 million for SAQ aquifer. For the sake of comprehension, in Figure 8-b record count is only shown for upto 0.5 million records. The interesting observation here is that in the presence of aquifers the good and moderate pastures are less, while poor pastures are more. The ratio between with and without aquifers being 1:1.63, 1:2.51 and 1:0.75 for good, moderate and poor pastures, respectively. Thus principal aquifers do not contribute to pasture growth, with potential to untap this groundwater resource.

5.4. Openness of SPOT

Tools such as Weka (www.cs.waikato.ac.nz/ml/weka/) and Orange (www.orange.biolab.si) support input of CSV (Comma Separated Value) file format and have several predictive data mining techniques to choose from. Since SPOT converts raster maps to CSV file format, therefore, these CSV files can be used as input these data mining and machine learning software suites. The predicative data mining results generated by these data mining tools in CSV format can subsequently be loaded into SPOT.

6. Conclusions

In this paper, we presented a multi-disciplinary, integrated solution followed by subsequent implementation of the SPOT (Spatial Prediction and Overlay Tool) tool for predicting suitable (wheat/alfalfa) cultivation. The proposed SPOT tool is based on the Naïve weighted Bayesian predictive data mining technique and the tool's accuracy is as per the accuracy of the said technique. As a case study, using historical wheat cultivation data, we demonstrated the application of SPOT by successful identification of those regions in the Kingdom where crops and fodder have not been grown historically, but have a high probability of growth success as per current cultivation on the predicted regions. Wheat yield and protein content are improved when wheat planted after alfalfa (fodder) instead of wheat-on-wheat [29]. Thus our prediction results are equally applicable for predicting suitable regions for cultivating alfalfa. As part of future work we plan to enhance our work by following the framework [24] for analyses and improvement of Data-fusion algorithms used and making SPOT web-enabled.

Acknowledgements

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH)-King Abdulaziz City for Science and Technology - the Kingdom of Saudi Arabia-award number 12-AGR2709-3. The authors also, acknowledge with thanks Science and Technology Unit, King Abdulaziz University for technical support.

References

[1] Al-Ahmadi F. and Hames A., "Comparison of Four Classification Methods to Extract Land Use and Land Cover from Raw Satellite Images for Some Remote Arid Areas, Kingdom of Saudi Arabia," *Journal of King Abdulaziz University, Earth Sciences*, vol. 20, no.1, pp. 167-191, 2008.

[2] Al-Jarrah O. and Abu-Qdais H., "Municipal Solid Waste Landfill Siting using Intelligent System," *Waste Management*, vol. 26, no. 3, pp. 299-306, 2006.

[3] Anonymous, http://estation.jrc.ec.europa.eu/, Last Visited, 2015.

[4] Anonymous, http://grass.osgeo.org, Last Visited, 2014.

[5] Anonymous, http://www.exelisvis.com/, Last Visited, 2015.

[6] Anonymous, http://www.exelisvis.com/ProductsServices/Jagwire.aspx, Last Visited, 2015.

[7] Anonymous, http://www.fao.org/giews/english/windisp/windisp.htm, Last Visited, 2015.

[8] Anonymous, http://www.tradingeconomics.com/saudi-arabia/arable-land-percent-of-land-area-wb-data.html, Last Visited, 2015.

[9] Auria L. and Rouslan A., *Support Vector Machines (SVM) as a Technique for Solvency Analysis*, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin, 2008.

[10] Chung C. and Fabbri A., "Probabilistic Prediction Models for Landslide Hazard Mapping," *Photogrammetric Engineering and Remote Sensing*, vol. 65, no. 12, pp. 1389-1399, 1999.

[11] Chung C. and Fabbri A., "The Representation of Geoscience Information for Data Integration," *Nonrenewable Resources*, vol. 2, no. 2, pp. 122-139, 1993.

[12] Chung C. and Fabbri A., "Validation of Spatial Prediction Models for Landslide Hazard Mapping," *Natural Hazards*, vol. 30, no. 3, pp. 451-472, 2003.

[13] Demšar D., Džeroski S., Henning Krogh P., and Larsen T., "Using Machine Learning to Predict the Impact of Agricultural Factors on Communities of Soil Microarthropods," *Metodološki Zvezki*, vol. 2, no. 1, pp. 147-159, 2005.

[14] Eerens H., Haesen D., Rembold F., Urbano F., Tote C., and Bydekerkedf L., "Image Time Series Processing for Agriculture Monitoring," *Environmental Modelling and Software*, vol. 53, pp. 154-162, 2014.

[15] Eldrandaly K., "Developing a GIS-Based MCE Site Selection Tool in ArcGIS Using COM Technology," *The International Arab Journal of Information Technology*, vol. 10, no. 3, pp. 276-282, 2013.

[16] Eldrandaly K., *Spatial Decision Making: An Intelligent GIS-Based Decision Analysis Approach*, VDM Verlag, 2010.

[17] Kannan P. and Hemalatha K., "Agro Genius: An Emergent Expert System for Querying Agricultural Clarification Using Data Mining Technique," *International Journal of Engineering and Science*, vol. 1, no. 11, pp. 34-39, 2012.

[18] Kantardzic M., *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons Publishing, 2003.

[19] Killmann W., *Climate Change and Food Security: a Framework Document*, Food and Agriculture Organization, 2008.

[20] Kim Y., "Comparison of the Decision Tree, Artificial Neural Network, and Linear Regression Methods based on the Number and Types of

Independent Variables and Sample Size," *Expert Systems with Application*, vol. 34, no. 2, pp. 1227-1234, 2008.

[21] Lee C., Gutierrez F., and Dou D., "Calculating Feature Weights in Naive Bayes with Kullback-Leibler Measure," *in Proceedings of IEEE 11th International Conference on Data Mining*, Vancouver, pp. 1146-1151, 2011.

[22] Linear Regression Model, CAMO, http://www.camo.com/rt/Resources/linear_regression_model.html, Last Visited, 2015.

[23] Lu K. and Yang D., "Image Processing and Image Mining using Decision Trees," *Journal of Information Science and Engineering*, vol. 25, pp. 989-1003, 2009.

[24] Nassar M., Kanaan G., and Awad H., "Framework for Analysis and Improvement of Data-fusion Algorithms," *in Proceedings of 2nd IEEE International Conference on Information Management and Engineering*, Chengdu, pp. 379-382, 2010.

[25] Nayab N. Disadvantages to Using Decision Trees. http://www.brighthubpm.com/project-planning/106005-disadvantages-to-using-decision-trees/, Last Visited, 2015.

[26] Neteler M. and Mitasova H., *Open Source GIS: A GRASS GIS Approach*, Springer, 2008.

[27] Neteler M., Bowman M., Landa M., and Metz M., "GRASS GIS: A Multi-purpose Open Source GIS," *Environmental Modelling and Software*, vol. 31, pp. 124-30, 2012.

[28] Quinlan J., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 2014.

[29] Rotational Value of Alfalfa, http://www.americasalfalfa.com/alfalfa/media/PDFs/Rotational_Value_v4, Last Visited, 2014.

[30] Ruiz L., Recio J., Hermosilla T., and Sarria A., "Identification of Agricultural and Land Cover Database Changes using Object-oriented Classification Techniques," *in Proceedings of 33rd International Symposium on Remote Sensing of Environment*, Stresa, 2009.

[31] Siqueira J., Paço T., Silvestre J., Santos F., Falcão A., and Pereira L., "Generating Fuzzy Rules by Learning from Olive Tree Transpiration Measurement-An Algorithm to Automatize Granier Sap Flow Data Analysis," *Computers and Electronics in Agriculture*, vol. 101, pp. 1-10, 2014.

[32] Spatial data warehouse. http://www.spatial-eye.com/Engels/Applications/Spatial-DWH/page.aspx/117, Last Visited, 2015.

[33] Steeg van de J. and Tibbo M., *Livestock and Climate Change in the Near East Region*, Food and Agriculture Organization, 2012.

[34] Tan P., Steinbach M., Karpatne A., and Kumar V., *Introduction to Data Mining*, Addison-Wesley Publishing, 2006.

[35] Witten I. and Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishing, 2005.

[36] Xia H., Liu H., and Zheng C., "A Markov-Kalman Model of Land-Use Change Prediction in XiuHe Basin, China," *in Proceedings of Geo-Informatics in Resource Management and Sustainable Ecosystem*, China, pp. 75-85, 2013.

[37] Zhao Y. and Zhang Y., "Comparison of Decision Tree Methods for Finding Active Objects," *Advances in Space Research*, vol. 41, no. 12, pp. 1955-1959, 2007.

Ahsan Abdullah did his PhD in data mining from the University of Stirling UK, MSc Computer Sciences and MSc Computer Engineering both from the University of Southern California USA and BS Electrical Engineering (with distinction) from Univ. of Engineering and Tech. Lahore. He is at the Dep. of IT, Faculty of Computing and IT, King Abdulaziz University, Jeddah. He is the PI of the project funded by the King Abdulaziz City for Science and Technology titled "Using Data Mining for Predicting Long Term Productivity of Pastures in the Kingdom of Saudi Arabia". He has been the lead guest editor of special issue on Big Data of the Springer Cognitive Computation Journal with IF=3.47; in 2015 he had seven ISI indexed impact factor journal papers as first author.

Ahmed Bakhashwain did his PhD in Plant Physiology from King Saud University Riyadh, MSc in Renewable Natural Resources from University of Arizona, USA and BS in Biological Science from, King Abdulaziz University, Jeddah. He is currently serving in the Department of Arid Regions Agriculture Faculty of Meteorology, Environment and Arid Land Agriculture King Abdulaziz University, Jeddah. He is the CO-I of the KACST funded project with the primary author.

Abdullah Basuhail did his PhD in Computer Science, Digital Image Processing (Wavelet Transform) from Florida Institute of Technology, USA. He is currently serving in the Department of Computer Science Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah He is the CO-I of the KACST funded project with the primary author.

Ahtisham Aslam did his PhD in Semantic Web from University of Leipzig Germany, MS in Computer Science from Hamdard University Lahore and BSc from the Punjab University Lahore. He is currently serving in the Dept. of IS, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah.