A review on the methods for big data analysis in agriculture

B I Evstatiev\textsuperscript{1} and K G Gabrovska-Evstatieva\textsuperscript{2}

\textsuperscript{1} Department of Electronics, University of Ruse Angel Kanchev, 8 Studentska Str., Ruse 7017, Bulgaria
\textsuperscript{2} Department of Computer Science, University of Ruse Angel Kanchev, 8 Studentska Str., Ruse 7017, Bulgaria

E-mail: bevstatiev@uni-ruse.bg

Abstract. The integration of information and communication technologies into agriculture lead to the development of precision agriculture. Nowadays it relies on Internet of Things devices, geospatial data, historical and real-time information, which has the potential to transform farming into smart farming. However the use of Big Data requires significantly different skills and knowledge, compared to what many farmers and agronomists possess, which is an obstacle for their effective use. This study aims to summarize and provide insight into the common methods used for data analysis in a wide variety of agricultural applications. Initially, the basic characteristics and sources of agricultural data are explained. Next, a review of the common data analysis methods (classification, clustering and regressions) is provided, containing information about the data sources used as well as the desired goal of the analysis. At the end of the paper, a summary is given on the applicability of data analysis methods depending on the desired goals.

1. Introduction

The rapid development of information and communication technologies in the last decades has had a tremendous impact on agriculture. The use of Internet of Things (IoT) sensors, online datasets and navigation systems opened many new opportunities for automation of agricultural processes, which have the potential to solve the problems with food shortage [1]. On the other hand, the availability of such enormous amount of data became a major challenge for people involved in agricultural activities for numerous reasons. It is necessary to analyse the available information in either real-time or near real-time if it is to be useful [1,2].

Numerous methods and tools for Big Data analysis were developed in the last decades, which are aimed at facilitating the process. However, their application requires significantly different knowledge and expertise, compared to what agronomists and other people involved in agriculture possess. The problem is exaggerated by the fact that many of them are not engineers, which limits the possibilities for effective application of precision agriculture [3]. Therefore, it is important that agricultural specialists are provided with adequate guidelines and information about the different methods for data analysis.

This study is aimed at performing a review on the common methods for analysis of agricultural data, as well as their area of application. It should analyse and summarize their characteristics and requirements in order to ease agricultural specialists when selecting the appropriate method based on the desired goals.
2. Big data in the agriculture

2.1. Big data characteristics
According to [2,4], big data can be characterized with the following dimensions:

- **Volume**: The size of data collected for analysis. For example, there are numerous open databases with remote sensing data with volume varying from terabytes to petabytes.
- **Velocity**: The time window in which data is useful and relevant – a certain data should be analysed in a reasonable amount of time in order to be useful. For example, the identification of pests and diseases would only be useful if identified in its early stages.
- **Variety**: The numerous data sources lead to significant difference in the properties of the collected data. It could vary in type (image, sensor data, etc.), time interval, resolution, etc.

Furthermore, other characterizations of big data also exist [1,6]:

- **Veracity**: This property refers to how accurate or truthful the data is. It could be considered from different points of view, such as is the data accurate, complete, consistent, objective, credible, etc.
- **Valorization**: Among the enormous amount of data, it is important to identify that one which is useful and has the potential to add value. Furthermore, it is important to determine how to make use of it.

The above dimensions and characteristics generally describe the big data; however, it is not mandatory that data analysis satisfy all of them as the analysis might get too complicated.

2.2. Sources of data in the agriculture
Several key sources of data used for analysis of agricultural process could be identified, which are analysed below.

**Ground sensor data**
Ground sensors are sensors that are installed literally on the ground. In the context of BIG data, they are considered as IoT devices. The main modules of an agricultural IoT system are responsible for data acquisition, data transmission and monitoring [6]. Different architectures exist for the data transmission, in terms of both communication medium as well as security [7,8].

**Remote sensing data**
Remote sensing is based mainly on the processing of images, obtained using satellites, airplanes, drones and radars. Different types of cameras are used, providing colour (RGB) images, infrared images, etc. [9].

**Weather stations**
Real-time information from weather stations is an important source for monitoring the environmental conditions in agricultural areas [1,5]. They could be considered as a set of ground sensors (temperature, humidity, rainfall, air velocity, pressure, radiation, etc.), providing remote data. Furthermore, weather station systems commonly provide forecasting services, which is crucial for optimization and management of the agricultural activities.

**Historical information and datasets**
Historical information and datasets can have many dimensions and numerous sources, varying from archive weather information, different datasets with images and characteristics of plants, crops, diseases, agricultural areas, variation of prices and many more [1].

In theory everyone could make their own data analysis using the available historical data, however that requires a significant qualification. Therefore, for many agricultural specialists involved into decision making, it is easier to rely on statistical data, provided by the government, public or local authorities.

**Geographic Information System (GIS)**
The GIS technology commonly enriches the data obtained from other sources, by providing precise information about their location, thus identifying the agricultural objects and their parameters. This allows summarizing the big data on geographical maps [1,5].
**Data obtained from humans**

The human opinion is also an important source of data. People act as non-technical sensors relying on intelligence and comprehensive knowledge. The human “measurements” are created through personal and subjective observations, such as air quality, weather observations, statements on public safety, etc. This information could be collected using different tools, such as mobile or web applications, surveys, feeds from social media, etc. When the human opinion is analysed historically, it can provide valuable knowledge about the trends and expectations [1].

**3. Data analysis in the agriculture**

Considering the enormous amount and variety of agricultural data, it is very important to analyse and present it in an appropriate form that is easy for understanding even by non-engineer specialist. Several data analysis methods exist, which are explained below.

**3.1. Regression analysis**

Regression analysis is used to estimate the relationship between one or more independent variable(s) and a single dependent variable. It provides statistical means to:

- Investigate the significance of the relation between the dependent and independent variables;
- Assess the strength of impact of multiple independent variables on a single dependent one.

Furthermore, regression allows to compare variables on different scales with different units. Different regressions exist, depending on the number of independent variables, the shape of the regression line (linear, nonlinear, etc.), the type of the dependent variables (numeric, Boolean, etc.). A summary of the common applications of regression for analysis of agricultural data is presented in table 1.

| Sources of data | Type of regression | Goal of the analysis |
|----------------|--------------------|----------------------|
| Temperature; pH of soil; Nitrogen content in soil; Phosphorus content in soil; Potassium; Rainfall; Water required; Yield. | Multiple regression | Yield forecasting [10] |
| Milk yield; Temperature; Temperature-humidity index (THI); Delay impact | Nonlinear/Multiple regression | Dependency of milk yield on THI [11] |
| Irrigation; Fertilization | Linear regression | Dependency of crop yield on fertilization and irrigation [12] |
| Satellite images; Crop yield | Linear regression | Dependency of average crop yield on average NDVI [13] |
| Farm physical characteristics; Survey results; Management practices | Multiple regression | Influence of management practices [14] |

**3.2. Clustering**

Clustering is a set of methods that allow grouping together different agricultural objects. The clustering could be done for different reasons, such as optimization, identification of management zones, differentiating between object properties and quality. According to [15], a common clustering approach includes (figure 1):

- Step 1. Identification of the clustering parameters of the different agricultural objects;
- Step 2. Obtaining the optimal number of clusters;
- Step 3. Obtaining the centre of each cluster;
- Step 4. Obtaining the main characteristics of each cluster (profile, size, etc.).

An overview of common applications of cluster analysis in agriculture is presented in table 2.
Step 1. Identify the clustering parameters

Step 2. Obtain the optimal number of clusters

Step 3. Obtain the centre of each cluster

Step 4. Obtain the main characteristics of each cluster

Example, Soil type:
- Black and Red
- Black and Laterite
- Red & Sandy
- Granitic, Sandy

Figure 1. Flowchart diagram of a common clustering approach.

Table 2. Application of Clustering for analysis of agricultural data.

| Sources of data                                | Methods used                  | Goal of the clustering                        |
|------------------------------------------------|-------------------------------|---------------------------------------------|
| Road distance between farms                    | Hierarchical clustering analysis | Optimization of supply chains by distance [15] |
| Map coordinates, NDVI                         | Clustering method              | Simplification of data [5]                  |
| Datasets with average temperatures, average rainfall, soil properties | Density based clustering (DBSCAN algorithm) | Clustering of regions based on similarities in temperature/rainfall/soil type [10] |
| Number of water tanks; Number of bore wells; Number of open wells | K-means algorithm             | Clustering of regions based on water level (to be used for yield forecasting) [16] |
| Dataset of images with plants and pests        | K-means algorithm              | Identification of plant pests [17]           |
| Soil physical properties                      | Fuzzy c-means clustering       | Defining irrigation management zones [18]    |

3.3. Classification

Classification is a different method, which is aimed at categorizing objects based on their properties, which are called predictors. It starts with finding an appropriate model and training it. Once the model is trained, it can be used to predict the classes of new objects, i.e. it allows their categorization. An overview of common applications of different classification methods with agricultural data is presented in table 3.

Table 3. Application of Classification for analysis of agricultural data.

| Sources of data used                                                                 | Methods used                  | Goal of the classification                        |
|--------------------------------------------------------------------------------------|-------------------------------|---------------------------------------------|
| Data from multispectral, RGB and thermal sensors; Coordinates; Enhanced vegetation index | Decision trees                | Monitoring fertilizer efficiency [19]          |
| Images of plants with weed and soil on the background                                | Support Vector Machine (SVM)  | Weed detection [20]                           |
| Images of plants under different outdoor conditions                                 | k-Nearest Neighbour SVM       | Detection of plants [21]                      |
| Different vegetation indices (NDI47, NDI45, EVI, MCARI and GNDVI)                   | SVM                           | Detection of land and crop type [22]          |
| Numerous crop images classified as either healthy or with a certain disease          | Convolutional Networks (CNN)  | Crop disease detection [23]                   |
| Images with daily development of diseases; Different indices (NDVI; Simple ratio; Structure insensitive vegetation index; etc.) | SVM; Convolutional Networks (CNN) | Disease early detection [24]                   |
| C-band radar backscatter data (satellite)                                            | Image processing; Deep learning | Crop type detection [25]                      |
4. Conclusions
The present study was aimed at performing a review of the common data analysis methods used for agricultural applications. Initially, the key characteristics of Big Data are explained and the main sources of agricultural data are overviewed (ground sensor data, remote sensing data, weather stations, historical information and datasets, GIS geospatial data and human data).

Next, the key data analysis methods are analysed. The performed review showed that regression analysis is the obvious solution when the goal is to investigate the dependencies between agricultural data presented in numerical form. Image sources are also applicable though they are also converted to numerical arrays, such as vegetation indices and coordinates. Examples for successful application of regression analysis is yield forecasting, assessment of management routines, etc.

The clustering approach can be used with any type of input data when the goal is to define zones by some of their properties or by their mutual distances. Examples for such applications are optimization of supply chains, simplification of classified data, categorization of zones (management, yield, rainfall, etc.), identification of pests and other.

The third group of methods is classification analysis, and can be used with any type of input data. Its key agricultural applications include object detection and categorization by one or several properties. It is commonly used for detection of weeds, plants, crops, diseases and pests as well as for categorization of agricultural objects for easier understanding and analysis.

The review conducted in this work could be useful for agronomists and people involved into agricultural activities, who wish to get initial guidelines on the methods for analysis of farming data. The presented results and conclusions provides information about the recommended data analysis methods based on the goals of the investigation.

References
[1] Kamilaris A, Kartakoullis A, Prenafeta-Boldú F 2017 A Review on the practice of big data analysis in agriculture Computers and Electronics in Agriculture 143
[2] Chi M, Plaza A, Benediktsson J, Sun Zh, Shen J and Zhu Y 2016 Big data for remote sensing: challenges and opportunities Proceedings of the IEEE 104 (11) pp 2207-2219
[3] Madhuri J, Indiramma M 2019 Role of big data in agriculture International Journal of Innovative Technology and Exploring Engineering (IJITEE) 9 (2) pp 3811-3021
[4] Ivanova E, Iliiev T, Stoyanov I 2020 Modelling and simulation of big data networks 2020 7th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE), 12-14 November 2020, Ruse, Bulgaria, pp 1-4
[5] Saiz-Rubio V and Rovira-Más F 2020 From smart farming towards agriculture 5.0: a review on crop data management Agronomy 10 207 doi:10.3390/agronomy10020207
[6] Zhang Ch and Liu Z 2019 Application of big data technology in agricultural Internet of Things Advanced Big Data and Artificial Intelligence Technologies for IoT Systems 15 (10)
[7] Alkhatib A, Baicher G 2012 Wireless sensor network architecture 2012 International Conference on Computer Networks and Communication Systems(CNCS 2012) IPCSIT 12 IACSIT Press, Singapore
[8] Hristova T 2020 Blockchain application in mining management E3S Web of Conferences 168 00009.
[9] Montes de Oca A, Arreola L, Flores A, Sanchez J and Flores G 2018 Low-cost multispectral imaging system for crop monitoring International Conference on Unmanned Aircraft Systems (ICUAS)
[10] Majumdar J, Naraseyappa S and Ankalaki S 2017 Analysis of agricultural data using data mining techniques: application of big data J Big Data 4 (20)
[11] Türkmenn E, Yasilioğlu E 2018 The effect of heat stress on milk yield in a Ipard supported dairy cattle enterprise Proceedings of the XIX World Congress of CIGR Turkey pp 51-61
[12] Maharjan B, Liang W Z, Panday D and Qiao X 2019 Evaluating crop sensor in maize grown in semi-arid condition under varying irrigation and nitrogen levels The 12th European Conference on Precision Agriculture, France
[13] Paixão L, Marques da Silva J R, Terron J M, Ramiro A and Ordóñez F 2019 Pisum sativum L. (Pea) yield modelling using Sentinel-2 NDVI maps The 12th European Conference on Precision Agriculture, France

[14] Townsley R and Parker W 1987 Regression analysis of farm management survey data New Zealand Journal of Experimental Agriculture 15 (2) pp 155-162

[15] Bălan A-V, Toma E, Dobre C and Soare E 2015 Organic farming patterns analysis based on clustering methods Agriculture and Agricultural Science Procedia 6 pp 639-646

[16] Manjula E and Djodiltachoumy S 2017 A model for prediction of crop yield International Journal of Computational Intelligence and Informatics 6 (4)

[17] Fina F, Birch F, Young R, Obu J, Faithpraise B and Chatwin C 2013 Automatic plant pest detection and recognition using k-MEANS clustering algorithm and correspondence filters International Journal of Advanced Biotechnology and Research 4 (2) pp 189-199

[18] Jiang Q, Fu Q and Wang Z 2011 Study on delineation of irrigation management zones based on management zone analyst software Computer and Computing Technologies in Agriculture IV. CCTA 2010. IFIP Advances in Information and Communication Technology 346 Springer

[19] Florence A, Revill A, Gibson-Poole S, Vigors B, Rees RM, MacArthur A, Barnes AP, Hoad SP and Williams M 2019 Advanced technologies for efficient crop management (ATEC) The 12th European Conference on Precision Agriculture, France

[20] Wellhausen C, Pflanz M, Pohl J-P and Nordmeyer H 2019 Automatic weed recognition for site-specific herbicide application The 12th European Conference on Precision Agriculture, France

[21] Arroyo J, Guijarro M and Pajares G 2016 An instance-based learning approach for thresholding in crop images under different outdoor conditions Computers and Electronics in Agriculture 127 pp 669–679

[22] Devadas R, Denham R J and Pringle M 2012 Support vector machine classification of object-based data for crop mapping, using multi-temporal landsat imagery International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, ISPRS Congress, Australia

[23] Sambasivam G and Opiyo G 2020 A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks Egyptian Informatics Journal https://doi.org/10.1016/j.eij.2020.02.007

[24] Rumpf T, Mahlein A-K, Steiner U, Oerke E-C, Dehne H-W and Plümer L 2010 Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance Computers and Electronics in Agriculture 74 pp 91-99.

[25] Teimouri N, Christiansen M P, Sørensen C A G and Jørgensen R N 2019 A deep learning-based approach for crop classification using dual-polarimetric C-band radar data The 12th European Conference on Precision Agriculture, France.

Acknowledgments
This work was supported by the Erasmus+ Program of the European Union under contract no 609544-EPP-1-2019-1-PS-EPPKA2-CBHE-JP “Boosting Innovation in Education aNd REsearch oF Precision AgriculturE in Palestine (BENEFIT)”.
