Predict the Main Factors that Affect the Vegetable Production in Palestine Using WEKA Data Mining Tool*

Dr. Yousef Abuzir**

*Received: 2/10/2017, Accepted: 31/10/2017.
** Associate Professor/ Al-Quds Open University/Palestine

Abstract:

This research presents an applied study using data mining to discover some factors affecting agricultural vegetable production and predicting the yield production in Palestine. In this research, we are interested in finding some factors that will influence the agricultural production to increase the amount of production to benefit the farmers in particular and individual, society in general. To achieve this goal we used Waikato's Knowledge Analysis Environment (WEKA) tool and algorithms such as K-Means, Kohonen's Self Organizing Map (KSOM) and EM to identify the most influential factors that increase the production of agricultural vegetable. This research has proved that K-Means is worthwhile to increase the efficiency and reliability of the prediction process of determining the factors that affect the yield production and KSOM the most accurate to predict the yield production.

Keywords: Data Mining, K-means, EM Algorithm, Kohonen's Self-Organizing Map (KSOM) and Clustering.

ملخص:

يقدم هذا البحث دراسة تطبيقية باستخدام استخراج البيانات لاكتشاف بعض العوامل التي تؤثر على إنتاج الخضراوات الزراعية والتنبؤ بكمية المحصول الزراعي في فلسطين. في هذا البحث، نهتم بإيجاد بعض العوامل التي من شأنها التأثير على الإنتاج الزراعي لزيادة كمية الإنتاج لصالح المزارع بشكل خاص والفرد والمجتمع بشكل عام. لتحقيق هذا الهدف استخدمنا أداة ويكا WEKA والخوارزميات مثل ك–مينز K-Means، خريطة التنظيم الذاتي كوهونين KSOM وخوارزمية أي إم EM لتحديد العوامل الأكثر تأثيرا والتي تزيد من إنتاج الخضروات الزراعية. وقد أثبت هذا البحث أن خوارزمية ك–مينز هي الاجدر بالاهتمام لزيادة كفاءة وموثوقية عملية التنبؤ عن طريق تحديد العوامل التي تؤثر في الإنتاج الزراعي، وان خوارزمية كوهين KSOM هي الأكثر دقة في التنبؤ بالمحصول الزراعي.

INTRODUCTION

In the past, we faced the problem of the scarcity of information or the difficulty to access it. Nowadays, the situation is reversed in light of information revolution that became a huge flood of information and the real wealth of the institution. We can neglect or exploit these data. The matter does not stop when storing this data, but how to exploit them.

Data mining is certainly a very important topic. The widespread and easy availability of information technology has inflated the volume of information, making the issue of large data on the Internet controversial. This has increased the need for the development of powerful tools to deal with this huge amount of information for analyzing data and extracting information and knowledge. Therefore, we use intelligent tools to manipulate these data. Data mining is a technique aims at extracting knowledge from vast amounts of data. This technique based on mathematical algorithms, which is the basis for data mining. Data mining derived from many sciences such as statistics, mathematics, logic, artificial intelligence, expert systems, neural networks, pattern recognition, machine learning and other sciences that are considered intelligent and non-traditional sciences (Mohammed, 2016).

Several core techniques used data mining by describing the type of mining and data recovery operation. Different studies and solutions always share or use the following Data mining Techniques (Brown, 2012), (Patel et al., 2014):

- Association
- Classification
- Clustering
- Prediction
- Sequential patterns
- Decision trees

The importance of this research originates from the idea of using data mining techniques to create predictive results that benefit the different beneficiary in the agricultural sector. In addition, working according to this technique will provide advice and forecasts an increase in the agricultural production by finding the main

factors that influence the agricultural vegetable crops in Palestine.

This study is concerned with datasets resulting from 11 agricultural seasons sites in Palestine (Bureau, 2017). We applied K-Means, Kohonen's Self Organizing Map (KSOM) and EM on the dataset. According to the results of the different types of tests conducted on these models, the results show that K-Means is the most accurate model in determining the factors that affect the agricultural production and KSOM is the most accurate algorithm to predict yield production.

The focus of this study is to identify key attributes that affect agricultural vegetable crop. This paper uses data mining techniques to help in predicting and finding the combination of factors required to identify high performance vegetable crop for the farmers and researchers in agricultural fields.

Moreover, this research identifies the most influential factors that increase the production of agricultural vegetable. In general, area, fertilizers, equipment and evaporation are the most important factors using K-Means algorithms. When EM and KSOM are used instead, different factors are considered, these are temperature, speed, humidity. This research shows that K-Means are worthwhile to increase the efficiency and reliability of the prediction process. Both EM and KSOM produced similar outputs to predict the main factors that affect the vegetable production in Palestine. Comparing these factors with real situations, the results of these two methods were less useful than those results produced by K-means.

Literature Survey

Data mining emerged in the late 1980s and proved to be a successful solution for analyzing large amounts of data, transforming them from simple data that was accumulated and not understood into information that could then be exploited and utilized.

Shu et al. in their review paper, they presented the main data mining techniques and their applications and development from year 2000 to 2011. The main findings of their review paper are DMT is an increasing application in many areas; these techniques have the ability to continually change and acquire new understanding of the application of DMT in many fields and allowing many new future applications (Shu et al, 2011).

Studies by agricultural researchers have shown that attempts to increase crop productions have led to the use of pesticides in a dangerous and significant manner. These studies have indicated a positive relationship between the use of pesticides and crop (Antonio et. al., 2009). A study (Abdullah et al., 2004) showed that how to extract integrated data for agricultural data, including pest detection, pesticide use and meteorological recordings, is useful for optimizing the use of pesticides.

Data mining techniques are often used to study soil properties. For example, the use of k-means approach for soil classification integrated with GPS-based technologies (Meyer , et al., 2004), and the k-means algorithm was also used to classify soils and plants (Verheyen, et al., 2001) as well as Support Vector Machines (SVM) technique for crop classification (Camps et al., 2003) .

Rani used a data mining to manage the feed security. The study suggests clustering techniques to classify new feed resource into a particular category that will be useful in determining the extent of usage of the new feed. Clustering the feeds into different groups provides multiple options to the end user (farmer or feed industry) to choose from a wide range of feed resources. The main finding of the study is that clustering can be effectively used for grouping of different feed resources without the aid of experts to an extent of 70% and thus can form a sound basis for efficient feed management. (Rani 2010).

There are many researchers who are engaged in the research of food security early warning and they have many achievements (Liu et al, 2010), (Svyazinska, 2016), (Sakurai et al., 2013).

The study of Raorane and Kulkarini discussed how data mining used as an effective tool for yield estimation in the agricultural sector. In general, crop production depends on different factors like climatic, geographical, biological, political and economic factors. In their research, data mining used to extract knowledge from the raw data and estimate the amount of crops production.

WEKA is abbreviation for Waikato's

Knowledge Analysis Environment. It is a Java based open source tool created by researchers at the University of Waikato in New Zealand. It is a tool that provides support for data mining. WEKA is a collection of open source of many data mining and machine learning algorithms, including preprocessing on data, classification, clustering and association rule extraction. Other main features of WEKA include (Alka, et al. 2017):

- Data Processing tools.
- User interface provides portability, rapid prototyping and graphical interface with explorer, experimenter and knowledge flow features.
- Statistical, regression, classification and clustering algorithms.
- The WEKA data mining tool displays better results in terms of true positive, false positive, precision and f-measure (Sharma et al., 2014) .
- Machine learning tools, WEKA is being expanded to support a variety of tools such as statistical analysis programs and spreadsheets, to allow the user to perform additional analysis, verification and data manipulation (McQueen, 1995).

WEKA is used as a data analysis process model for data mining in agriculture (Ayman et. al., 2015), (Cunningham and. Holmes, 2005). The data extraction algorithm in the agricultural domain model can then be easily integrated into the software application. The analysis and application of WEKA based on a case study in the agricultural field, ie in the mushroom classification, has been demonstrated.

Data Mining in Agriculture

In nowadays, a lot of information related to agricultural activities becomes available in electronic format. This information needs further analysis to obtain important information to support people working or related to the field of agriculture. Using Data mining techniques in agriculture is useful in many areas like predicting agricultural problems, increasing agricultural production, detecting diseases, classification of agricultural products and plants, identifying and classification of soil problems and optimizing the use of pesticides (Manjula and Djodiltachoum, 2016) (Mirjankar and Hiremath, 2016).

Data mining in agriculture is one of the most recent research topics (Jyotshna and Yusuf, 2015). It is based on applying data mining techniques and clustering algorithms to agriculture (Milovi and Radojevi , 2015). In the field of agriculture, many researches applied the k-means algorithm for (Mucherino et al., 2009; Antonio et al. 2009) :

- Predicting yield production (Ramesh. and Vishnu, 2013), (Aishwarya, 2016), (Ramesh and Vardhan, 2013).
- Classifying plant, soil (Meyer GE et. al., 2004);
- Classifying soils in combination with GPS-based technologies (Verheyen et al., 2009; Meyer GE et al., 2004);
- Forecasting pollution in the atmosphere (Jorquera, 2001) ;
- Analyze color images of fruits as they run on conveyor belts (Leemans and Destain, 2004);
- Classify eggs as fertility and Patel VC (Das and Evans , 1992);
- detecting weed and nitrogen stress in corn (Karimi et al. , 2006);
- Monitoring water quality changes (Klise. and McKenna, 2006);
- Grading apples before marketing (Leemans and Destain, 2004);
- Detecting weeds in precision agriculture (Tellaeche et. Al., 2008).
- Predicting wine fermentation problems (Urtubia et. Al., 2007);

Clustering

Data collection is the process of placing data in similar clusters. The aggregation algorithm divides a data set into several clusters as shown in Figure 1. In clustering, the similarity between points within a particular pool is greater than the similarity between points within two different clusters. The idea of data collection is simple in nature and very close to human in its way of thinking (Berson et al., 1999). Whenever we deal with a large amount of data, we tend to summarize the vast amount of data into few groups or categories in order to facilitate the analysis process.

For the K-means clustering algorithm, given

the data $<x_1, x_2,\ldots,x_n>$ and K, assign each x_i to one K clusters, $C_1 \ldots C_k$, minimizing equation 1 (Khedr et. al, 2014)

$$SSE= \sum^{K}_{j=1}\sum_{xi\in Cj}\|x_i-\mu_j\|^2 \quad (1)$$

Where

K is the number of desired clusters

SSE is Sum of Squared Error (SSE)

μ_j is mean over all points in cluster C_j.

K-Means algorithm uses the following steps:

1. Initialize randomly $\mu_j \ldots \mu_k$ randomly
2. Repeat until convergence:

2.1. Assign each point xi to the cluster with closest mean μ_j

2.2. Calculate the new mean for each cluster (equation 2)

$$\mu j= 1/|C_j| \sum_{xi\in Cj}x_i \quad (2)$$

Another machine-learning algorithm used in data mining is EM. This algorithm uses two iterative steps called E-step and M-step:

• the E-step, where each object is assigned to the centroid such that it is assigned to the most likely cluster.

• the M-step, where the model (=centroids) are recomputed (= least squares optimization).

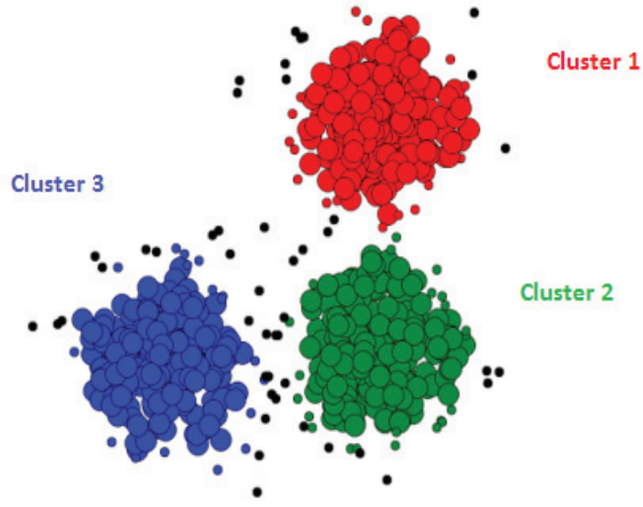

Figure 1 Clustering

The Kohonen Self-Organizing Map (KSOM) is unsupervised learning neural network algorithm. We can use KSOM in solving problems in various areas, especially in clustering complex data sets. Despite its advantages, the KSOM algorithm has a few drawbacks; such as overlapped cluster and non-linear separable problems. (Azlin and Rubiyah, 2016) (Fernando, 2015).

MATERIALS AND METHODS

Agricultural production in the world and in Palestine is of great importance in terms of satisfaction and high quality production, especially in terms of meeting farmers and population needs. Increasing agricultural products has become a necessary need in our time. To increase production, we need water for irrigation, land, tools, equipment, fertilizers, pesticides, new management methods, etc.

This research applies predictive data mining techniques in agriculture to predict the factors that this goal: we used WEKA tool and different algorithms such as K-Means, Kohonen's Self Organizing Map (KSOM) and EM. This section discusses the necessary details of the datasets and methods used in this study.

The data source for this research is the dataset obtained from the Palestinian Statistics Bureau website and we select the dataset for the vegetable for our research.

The dataset related to the recent 11 harvest years. The analysis of the data set contains 12 attributes among them, year, city, area, fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, speed and crop. The whole dataset consists of approximately 2832 complete records. In this paper, we analyzed the estimation of the crop with respect to the previous parameters. In our method in this research, to deal with the challenge of predicting the crop of vegetable crop, we divided the dataset into five major groups. Table 1 shows the parameters and their configuration in different groups.

The challenge is identifying key attributes that affect vegetable crop, such as region, equipment, fertilizers, rain etc. In this paper, we used data mining techniques to help farmers find the combination of factors required to increase crop and identify high performance vegetable crop.

WEKA is a program to deal with artificial

intelligence algorithms, which is an open source software package containing a set of algorithms that help to analyze and extract data (Data mining). We can apply these algorithms easily to a set of data directly through the WEKA software interface. It also contains tools that are capable of dealing with the following: Pre-processing, Classification, Clustering, Association rules, Select attributes and Visualization

To address the clustering issues, we utilized the machine - learning algorithm of simple K-Means, KSOM and EM. Figure 2 presents a schematic illustration of prediction using these three algorithms.

Table 1

Different configuration of attributes

Complete list of variables	Group A without year, city	Group B without year, city, area	Group C without year, city, fertilizers	Group D without year, city, rain	Group E without year, city, speed
year, city, area, fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, speed and yield	area, fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, speed and yield.	fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, speed and yield	area, pesticides, temperature, rain, equipment, humidity, evaporation, speed and yield	area, fertilizers, pesticides, temperature, equipment, humidity, evaporation, speed and yield	area, fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, and yield

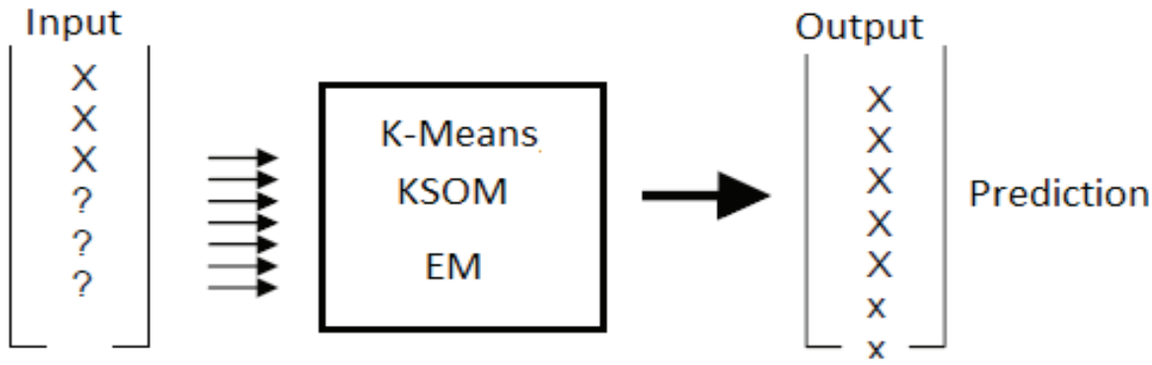

Figure 2

Schematic illustration of prediction, using the K-Means, KSOM and EM Algorithm.

We applied the K-means, EM and KSOM algorithms for clustering using our different datasets.

To test if it is possible to predict the yield and the most attributes that increase the crop production. We apply these algorithmson different groups from our dataset. Then we compared the results with other groups, where the whole set of data is considered.

Results and Discussion

This section discusses the different types of data mining algorithms applied, then compares and evaluates them using the dataset collected and prepared from Bureau of statistic in Palestine (Bureau, 2017).

In general, vegetable crop may vary from one city to another or it may vary within the same city or the same region. There are many factors affecting vegetable crop such as land, rain and various amounts of fertilizers and pesticides, etc. To increase the vegetable crop we have to address these different issues or factors. In this study, we will investigate these factors to expect a better vegetable crop and identify the main factors that affect the vegetable crop.

As mentioned earlier, dataset for group A just

consider 10 variables or attributes (see table 1) including area, fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, speed and crop. Whereas the other groups B, C, D and E related to different configuration with different parameters.

The k-means algorithm is applied to the datasets, by using k = 2, 3,4,5,6,8 and 10 and considering the data related to certain group.

Groups set A, B and C shows similar results that are crop is essentially linked to the shared attributes area, equipment, evaporation. Whereas, A, D and E have in common area, fertilizers. Table 2 shows all the results and other information.

Table 2

Screen dumps of the Results For K-Means (with K= 5)

Group	Results For K-Means (with K= 5)						
Original	Year	2001.5263	2000.9333	2005.6	2002.2941	2001.5556	1999.2727
	City	Jenin	Jerico	Jenin	Gaza	Hebron	Jenin
	Area	27592.9649	26800.2667	44405.6	19659.1176	42952.6667	20726.1818
	Fertilizers	3204.6491	2774.2667	6738.2	1357.4706	2891.1111	5296.6364
	Pesticides	2405.193	1451.6667	3845	901.3529	5173.7778	3109.9091
	Temperature	25.4987	29.0749	25.42	24.8176	20.4778	25.8182
	Rain	436.2535	231.47	451.22	514.7529	488.1667	544.9091
	Equipment	3903.093	1767.4	4650.4	1030.2941	9824.2222	6070.9364
	Humidity	60.2807	54.0667	65.6	62.8235	60	62.6364
	Evaporation	1888.3667	2046.9067	1991.82	1717.0765	1870.2	1904.7364
	Speed	8.1649	7.4067	4.38	10.0471	10	6.5091
	Yield	34144.2456	35020	87249.2	19886.4118	23441.4444	39603.0909
A	Area	27592.9649	29072	68854.8889	7561.4545	9754.4286	18998.8571
	Fertilizers	3204.6491	2223.6522	6219.5556	1777.2727	2676.8571	5322.4286
	Pesticides	2405.193	1188.2174	4486	1189.3636	5794.4286	2249.8571
	Temperature	25.4987	27.875	24.4556	24.1727	20.9	25.7143
	Rain	436.2535	375.2022	422.9556	382.8545	500.3429	673.7714
	Equipment	3903.093	2149.7826	7014.2222	532.2727	8253.6143	6610.4286
	Humidity	60.2807	54.9565	66	67.0909	57.8571	62.1429
	Evaporation	1888.3667	1993.7304	1964.3889	1598.4818	1848.4286	1939.9
	Speed	8.1649	8.2652	6.9556	9.4455	10.2286	5.3143
	Yield	34144.2456	27524.6522	70298.1111	26620	17803.7143	37575.1429
B	Fertilizers	3204.6491	2235.6364	7164.7	1777.2727	2752.75	3377
	Pesticides	2405.193	1215.7273	3488.3	1189.3636	6432.375	1820.8333
	Temperature	25.4987	27.7602	25.33	24.1727	21.225	25.6167
	Rain	436.2535	401.8591	509.74	382.8545	503.1375	448.6083
	Equipment	3903.093	1953.7727	7896.6	532.2727	9256.5375	3436.6667
	Humidity	60.2807	54.2727	64.1	67.0909	60.75	62.8333
	Evaporation	1888.3667	1977.0727	1973.09	1598.4818	1883.575	1959.75
	Speed	8.1649	8.7955	6.56	9.4455	9.825	3.9667
	Yield	34144.2456	28822.5909	66624	26620	25074.75	25411.1667

C

Area	27592.9649	32494.4167	57308.6667	8032.5385	9050.1667	24031.2857
Pesticides	2405.193	1695.1667	3904.6667	1359	6510.8333	940.4286
Temperature	25.4987	29.677	24.7333	24.4923	21.35	25.2857
Rain	436.2535	185.2542	474.9	427.2077	485.9333	605.3786
Equipment	3903.093	1841.5	5942.4167	607.4615	8832.8833	4869.6429
Humidity	60.2807	53.9167	65.0833	67	59	55.9286
Evaporation	1888.3667	2080.9417	1971.825	1639.6308	1877.6667	1887.3214
Speed	8.1649	7.0333	6.3333	8.5923	9.9333	9.55
Yield	34144.2456	41649.8333	68866.4167	23289.6154	19257.8333	14408.2143

D

Area	27592.9649	30707.1905	59130.625	7561.4545	28707.1429	17077.7
Fertilizers	3204.6491	2030.5714	6604.875	1777.2727	2715.5714	4862.5
Pesticides	2405.193	1167.8095	4266.25	1189.3636	6472.7143	2005
Temperature	25.4987	27.6154	24.9	24.1727	21.2857	25.94
Equipment	3903.093	1944.5238	6575.75	532.2727	9074.1857	5966.1
Humidity	60.2807	54.2381	66	67.0909	60	61.1
Evaporation	1888.3667	1972.4571	1963.15	1598.4818	1891.4714	1968.65
Speed	8.1649	9.0571	6.525	9.4455	10	4.91
Yield	34144.2456	27475.5238	75643	26620	20440.8571	32818.6

E

Area	27592.9649	29072	68854.8889	8032.5385	9050.1667	20954.1667
Fertilizers	3204.6491	2223.6522	6219.5556	2146.6923	2645.3333	5294.3333
Pesticides	2405.193	1188.2174	4486	1359	6510.8333	2110.1667
Temperature	25.4987	27.875	24.4556	24.4923	21.35	24.2833
Rain	436.2535	375.2022	422.9556	427.2077	485.9333	660.15
Equipment	3903.093	2149.7826	7014.2222	607.4615	8832.8833	8168.1667
Humidity	60.2807	54.9565	66	67	59	58.8333
Evaporation	1888.3667	1993.7304	1964.3889	1639.6308	1877.6667	1920.0667
Yield	34144.2456	27524.6522	70298.1111	23289.6154	19257.8333	43693.3333

Referring to the results in table 3, K-Means algorithm shows that area, fertilizers, equipment and evaporation are the most common factors that affect the crop production. When EM and KSOM is used instead, different factor are considered, such as temperature, speed, humidity (Figure 3). These factors reported to be an important factor in the problem crop prediction using EM and KSOM algorithms. The clustering process in which EM and KSOM are also included the factor temperature. Two groups A and B contain temperature, speed, humidity, other two groups C and D contain temperature, humidity, speed, and the remaining group E is only providing a temperature and humidity to be a good factor for crop production. It seems that EM and KSOM does not provide any different information, they show similar results.

In this analysis, one can conclude that Fertilizers, Pesticides and Evaporation are the main factors that increase the crop. Region is a very important factor that increase the crop. The most affected cities are Jenin and Jericho.

Figure 3 shows the result for both EM and Kohonen's Self Organizing Map. To obtain these

results, the number of clusters (K) is set equal to 5. Figure 3 shows only the results for the main factors that most affect the crop in our case.

Referring to Figure 3 and Table 3, it is clear that both EM and KSOM predict the same factors that affect the vegetables production and there is a big similarity between their results. on the other hand, there is a difference between the K-means and results KSOM and EM. All the results of K-means have different factors compared with the other machine learning algorithms.

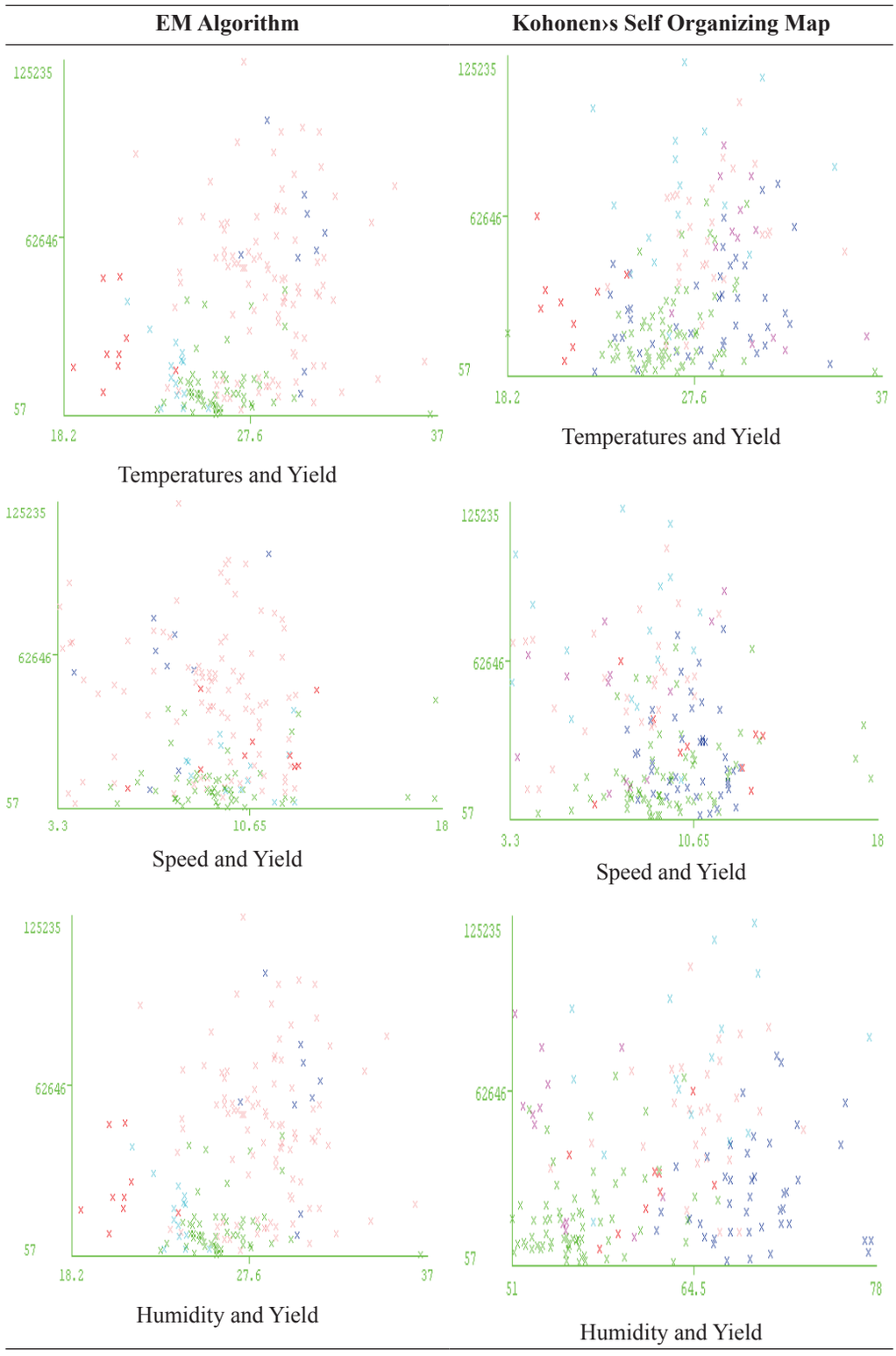

Figure 3

Graphs represent the results for EM and KSOM Algorithms

Table 3

Different Configurations Of Attributes and Attributes Predictions

Groups	Attributes used	K-means	EM	KSOM
Group A (-year, city)	area, fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, speed and yield.	area, fertilizers, equipment, evaporation	temperature, speed, humidity	temperature, speed
Group B (-year, city, area)	fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, speed and yield	fertilizers, equipment , evaporation, pesticides	temperature, speed, humidity	temperature, speed, humidity
Group C (-year, city, fertilizers)	area, pesticides, temperature, rain, equipment, humidity, evaporation, speed and yield	Area, equipment, pesticides, evaporation	temperature, humidity, speed	temperature, humidity
Group D (-year, city,rain)	area, fertilizers, pesticides, temperature, equipment, humidity, evaporation, speed and yield	area, fertilizers,	temperature, humidity, speed	temperature, humidity, speed
Group E (-year, city, speed)	area, fertilizers, pesticides, temperature, rain, equipment, humidity, evaporation, and yield	area, , equipment, fertilizers	temperature, humidity	temperature, humidity

Table 4 represents seven different experiments with different number of clusters, which is formed by the K-Means clustering. The estimation of average production using K-Means algorithm with respect to 12 parameters is computed using WEKA Tool and compare to the actual average production that is 34144.2456. The results in Table 4 and Figure 4 show that the Sum of Standard Errors (SSE) is decreasing as the number of clusters is increasing. From all the results available in the Table 4, we can see clearly the accuracy of prediction using K-Means algorithm lies within the range of the Sum of Squared Errors (SSE) 24 to 70. We used Sum of standard errors (SSE) to describe how well a K-Means algorithm performs on a certain K in our data set.

Another result available from the comparison in table 4, it's clearly seen that different value for K in K-Mean algorithm provide different crop predictions and the best result for the crop predictions is gained when K number of cluster is increased.

Table 4

A Relation Between No. Of Cluster, Sum of Squared Errors and Yield Prediction Using K-Means

No. of Clusters	Sum Of Squared Errors (SSE)	Number Of Iterations	Yield Prediction (Average Actual Data is 34144.2456)
2	70.65	5	51455.15
3	56.56	7	48175.5238
4	45.23	4	65832.1667
5	42.60	6	87249.2
6	31.82	7	65832.1667
8	32.25	4	67890.8333
10	24.58	5	72997.1667

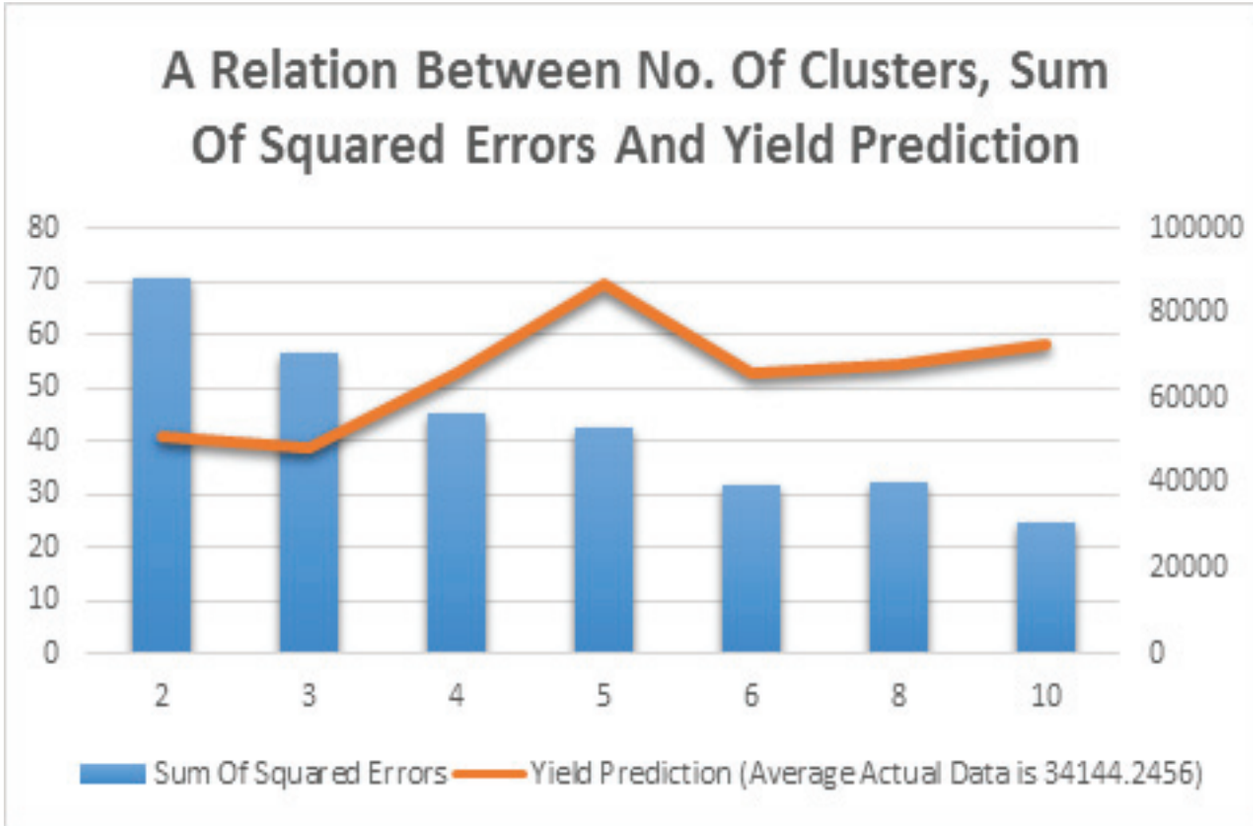

Figure 4

Relation Between No. Of Clusters, Sum Of Squared Errors And Yield Prediction Using K-Means

Table 5 and figure 5 show the prediction of the yield production by using both KSOM and K-Means with respect to actual average yield production (34144.2456). Table 4 gives the computed results by WEKA.

When comparing the results in the two columns for both K-Means with KSOM algorithms, it is clear that there is no intersection or similarity between K-Means and KSOM algorithm. The same result can be concluded to EM. Simple K-Means is an effective clustering model to increase vegetable corp. It gives highest accuracy result for increasing the yield production compared to average actual data is 34144.2456

After comparing the resulting techniques using the WEKA tool and consulting the agricultural engineer, this study concludes that the results of the K-Means algorithms is the most appropriate and accurate model for selecting various factors affecting the production of agricultural vegetable crop. At the same time KSOM is the most suitable and more accurate method to predict the estimated yield productions.

Table 5

Yield Prediction K-Means vs KSOM

No. of Clusters	Yield Prediction (Average Actual Data is 34144.2456)	
	K-Means	KSOM
2	51455.15	13348.8109
3	48175.5238	16894.9268
4	65832.1667	7808.8823
5	87249.2	53650.9932
6	65832.1667	61808.0951
8	67890.8333	60388.4371
10	72997.1667	62904.0615
Yield Prediction Average	65633.17	39543.46

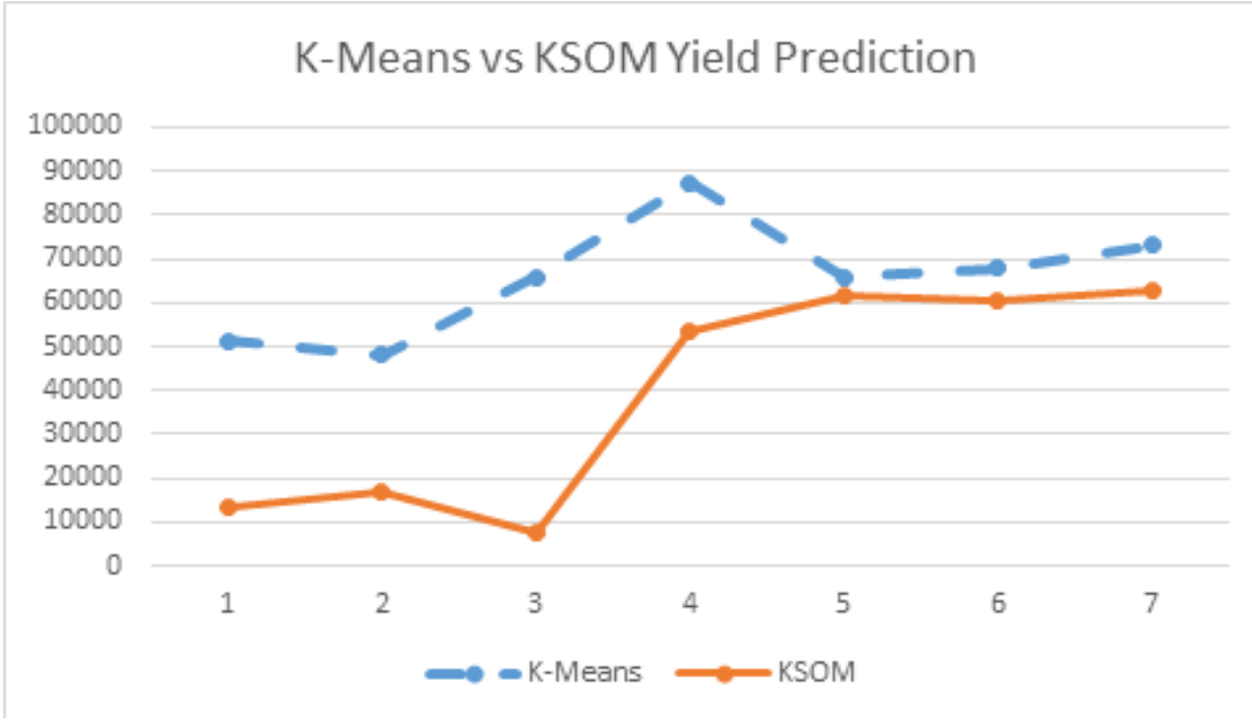

Figure 5

K-Means vs KSOM Yield Prediction

CONCLUSION

Data mining is relatively a novel research field and it is expected to grow in the future. In this emerging, multidisciplinary and interesting research field, there is a lot of work to be done. It will help in forecasting/ managing agricultural crop effectively.

This research focuses on applying data mining techniquesregarding vegetable data to extract knowledge from these data and estimate vegetable crop. Also, it aims to predict the main factors that impact the vegetable' production to satisfy farmers and citizens' needs for the upcoming years using K-means, KMOS and EM algorithms. We used WEKA as a predictive Data Mining Tool. Results showed that, byusing data mining we succeeded to predict the main factors that affect the vegetable' production. The obtained results could help decision makers for achieving food security and the country's productivity for the upcoming years continuously. K-means is useful for predicting the main factors that affect the vegetables production; meanwhile KSOM is the more accurate algorithm to predict the yield production.

This study is concerned with datasets resulting from 11 agricultural seasons sites in Palestine. To reach meaningful outputs and predictions, various experiments conducted through modifications of the attributes and the use of different numbers of these attributes to reach meaningful outputs and prediction.

The experiments conducted in this paper showed that farmers attempt to increase vegetable crop by using pesticides in a dangerous and significant manner. This study has indicated a positive relationship between the use of pesticides and crop.

This study did not deal with other factors like soil type, seasonal and environmental conditions, prediction and the accuracy of the weather forecast because the web site of the Bureau of Statistic in Palestine did not have a clear and complete dataset related to these factors.

In future, it is hoped that more interesting results will be reached by extending this model and considering the previous attributes in order to improve the accuracy of the prediction. The extensions should take into consideration these attributes in advance as well as regional variations in crop.

References

1. Abdullah, A., Brobst, S, Pervaiz.I., Umer M.,A.Nisar. 2004."Learning dynamics of pesticide abuse through data mining". Proc. of Australian Workshop on Data Mining and Web Intelligence, New Zealand, January.

2. Ahmad, A., & Yusof, R. (2016). A modified kohonen self-organizing map (KSOM) clustering for four categorical data. Jurnal Teknologi, 78(6-13), 75-80. DOI: 10.11113/jt.v78.9275.

3. Aishwarya B.R, A Literature Study on Application of Data Mining Tools for Rice Yield Prediction, (IJITR) International Journal Of Innovative Technology And Research, Volume No.4, Issue No.1, December - January 2016, 2757 – 2759. 2320 –5547 @ 2013-2016 http://www.ijitr.com .

4. Alka A., Malhotra P. K., Sudeep M., Anshu B., and Shashi D., Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets, Indian Agricultural Statistics Research Institute, (ICAR), E-book, online: http://www.iasri.res.in/ebook/win_school_aa/

5. Antonio M., Petraq J.Papajorgji, and Panos M.Pardalos, Data mining in agriculture, Springer, 2009.

6. Ayman E. Khedra, Mona Kadryb, Ghada Walidb,Proposed framework for implementing data mining techniques to enhance decisions in agriculture sector Applied case on Food Security Information Center, in Proceedings of International Conference on Communication, Management and Information Technology (ICCMIT2015),Procedia Computer Science 65 (2015) 633 – 642.

7. Berson A., Smith S. J., and Thearling K., An Overview of Data Mining Techniques : Building Data Mining Applications for CRM , McGraw-Hill Companies, December 22, 1999.

8. Brown M., Data mining techniques, IBM DeveloperWork, December 2012, online https://www.ibm.com/developerworks/library/ba-data-mining-techniques/

9. Bureau of Statistic in Palestine, http://www.pcbs.gov.ps/pcbs_2012/Publications.aspx , (2017).

10. Camps-Valls G, Gomez-Chova L, Calpe-Maravilla J, Soria-Olivas E, Martin-Guerrero JD, Moreno J., 2003, "Support vector machines for crop classification using hyperspectral data". Lect Notes Comp Sci 2652: pp. 134–141.

11. Cunningham S.J., G. Holmes. 2005. "Developing innovative applications in agriculture using data mining". Proc. of 3rd International Symposium on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005.

12. Das KC, Evans MD, Detecting fertility of hatching eggs using machine vision II: Neural Network classifiers. Trans ASAE 35(6):2035–2041, 1992.

13. Fernando Bacao, et al., "Self-organizing Maps as Substitutes for KMeans Clustering," Springer Computational Science – ICCS 2015 Lecture Notes in Computer Science, vol 3516,pp 476- 483,2015.

14. Jorquera H., Perez R., Cipriano A., and Acuna G., Short Term Forecasting of Air Pollution Episodes, In: Environmental Modeling 4, P. Zannetti (Ed.), WIT Press, UK, 2001.

15. Jyotshna Solanki and Yusuf Mulge, Different Techniques Used in Data Mining in Agriculture, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 5, May 2015 ISSN: 2277 128X.

16. Karimi Y, Prasher SO, Patel RM, Kim SH Application of support vector machine technology for Weed and nitrogen stress

detection in corn. Comput Electronics Agricult 51:99–109, 2006.

17. Khedr A., El Seddawy A. , and Idrees A. Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS, International Journal of Innovative Research in Computer Science & Technology (IJIRCST), ISSN: 2347-5552, Vol. 2, Issue 6, November 2014, pp. 111- 118.

18. Klise K.A. and McKenna S.A., Water Quality Change Detection: Multivariate Algorithms, Proceedings of SPIE 6203, Optics and Photonics in Global Homeland Security II, T.T. Saito, D. Lehrfeld (Eds.), 2006.

19. Leemans V, Destain MF, A real time grading method of apples based on features extracted from defects. J Food Eng 61:83–89, 2004.

20. Liu Z., Meng L., Zhao W. and Yu F. Application of ANN in food safety early warning, The 2nd International Conference on Future Computer and Communication, Wuhan, Vol. 3, 2010, pp. 677-80.

21. Manjula E. and DjodiltachoumS. y, Analysis of Data Mining Techniques for Agriculture Data, International Journal of Computer Science and Engineering Communications, Vol.4, Issue.2, Page.1311-1313, (2016).

22. McQueen R. J, Garner S.R.,Nevill-Manning C.G. , Ian H. Witten, "Applying machine learning to agricultural data". Comptuers and Electronics in Agriculture. Vol. 12: pp. 275-293, 1995.

23. Meyer GE, Neto JC, Jones DD, Hindman TW, 2004, "Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images". Computer Electronics Agric Vol. 42: pp. 161–180.

24. Milovi B. 1 and Radojevi V., Application of Data Mining in Agriculture, Bulgarian Journal of Agricultural Science, 21 (No 1) 2015, 26-34.

25. Mirjankar N. and Hiremath S., Application of Data Mining In Agriculture Field, International Journal of Computer Engineering and Applications,iCCSTAR-2016,Special Issue,May 2016.

26. Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, May 2016. ISBN: 9780521766333.

27. Mucherino A., Papajorgji P., Pardalos P.M., A Survey of Data Mining Techniques Applied to Agriculture, Operational Research: An International Journal 9(2), 121–140, 2009.

28. Patel Hetal P and Patel Dharmendra, A Brief survey of Data Mining Techniques Applied to Agricultural Data, International Journal of Computer Applications (0975 – 8887), Volume 95– No. 9, June 2014 .

29. Ramesh D B and Vardhan Vishnu, "Data Mining Techniques and Applications to Agricultural Yield Data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013, pp.3477-3480.

30. Rani V. Efficient management of feed resources using data mining techniques, Master of Philosophy, Department of Computer Science, Christ University, Bangalore, 2010, pp. 5, 87, 88.

31. Raorane A. and Kulkarini F., Data Mining: An effective tool for yield estimation in the agricultural sector, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 1, Issue 2, July – August 2012, pp.1-4.

32. Sakurai T., Mochida, K., Yoshida, T., Akiyama, K., Ishitani, M., Seki, M. & Shinozaki, K. Genome-wide discovery and information resource development of DNA polymorphisms in cassava. PLoS ONE 8 e74056 (2013). doi: 10.1371/journal.pone.0074056 http://dx.doi.org/10.1371/journal.pone.0074056 .

33. Sharma N. and Om H., Comparing the Performance of Data Mining Tools: WEKA and DTREG, International Journal of Scientific & Engineering Research, Volume 5, Issue 4, April-2014.

34. Shu-HsienLiao, Pei-HuiChu and Pei-YuanHsiao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications,

Volume 39, Issue 12, 15 September 2012, Pages 11303-11311 , https://doi.org/10.1016/j.eswa.2012.02.063 .

35. Svyazinska N., Data Mining of Agriculture as of Element of Food Security in Ukraine, Master Thesis, Department of the Mathematical Methods of System Analysis, National Technical University of Ukraine (NTUU),2016.

36. Tellaeche A., Burgos-Artizzu X.-P., Pajares G. and Ribeiro A., A Vision-Based Hybrid Classifier for Weeds Detection in Precision Agriculture Through the Bayesian and Fuzzy k-Means Paradigms, Advances in Soft Computing 44, 72–79, 2008.

37. Urtubia A., Perez-Correa J. R., Soto A., Pszczolkowski P., Using Data Mining Techniques to Predict Industrial Wine Problem Fermentations, Food Control 18, 1512–1517, 2007.

38. Verheyen K, Adriaens D, Hermy M, Deckers S., 2001, "High-resolution continuous soil classification using morphological soil profile descriptions". Geoderma Vol. 101: pp. 31–48.