JOURNAL OF AGRICULTURAL INFORMATICS

AGRÁRINFORMATIKA FOLYÓIRAT

Publisher
Hungarian Association of Agricultural Informatics (HAAI)

The publishing of the Journal is supported by the European Federation for Information Technology in Agriculture, Food and the Environment (EFITA)

2019.
Vol 10, No.2
ISSN 2061-862X

http://journal.magisz.org

The creation of the journal has supported by the New Hungary Development

TÁMOP 4.2.3-08/1-2009-0004

"Dissemination of research result on innovative information technologies in agriculture"

The project is financed by the European Union, with the co-financing of the European Social Fund.

Nemzeti Fejlesztési Ügynökség
ÚMFT infovonal: 06 40 638 638
nfu@meh.hu • www.nfu.hu

New Hungary Development Plan

Journal of Agricultural Informatics

Scientific Journal

Name: Agricultural Informatics

Language: English

Issues: 2-4 per year

Publisher: Hungarian Association of Agricultural Informatics (HAAI), H-4032 Debrecen, Böszörményi út. 138. Hungary

ISSN 2061-862X

Content

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 10, No. 2
journal.magisz.org

Comparison of Internal Validity Indices for Fuzzy Clustering

Zeynel Cebeci[1]

INFO
Received 30-11-2019
Accepted 16-12-2020
Available online 28-01-2020
Responsible Editor: M.Herdon:

Keywords:
cluster analysis, fuzzy clustering, number of clusters, cluster validation, internal validity index.

ABSTRACT

Partitioning clustering has been one of the key components of data analytics to discover meaningful patterns in agricultural big data, driven by the increasing use of IoT-based technologies in smart farming. In partitioning clustering, the quality of clustering or performances of clustering algorithms are mostly evaluated by using the internal validity indices. In this study, the effectiveness of some widely used internal fuzzy indices are compared using the basic Fuzzy C-Means clustering algorithm. It is especially aimed to investigate changes in the effectiveness of validity indices when fuzzy data points are at different distances from the cluster centers. According to the results obtained on the simulated two-dimensional datasets, Fuzzy Silhouette, Fuzzy Hypervolume and Kwon are the most successful indices in validation of fuzzy clustering results.

1. Introduction

In data mining for knowledge discovery, clustering is one of the most widely used unsupervised learning techniques to explore the meaningful substructures or patterns in examined datasets. Clustering as an exploratory data analysis is frequently applied in almost every area of agriculture, food, environment and the other life sciences subjects ranging from genomics to biomedical image segmentation. In recent years, clustering has gained an increasing importance in knowledge discovery from the big data collected via agricultural data acquisition systems and sensors networks based on the systems using IoT with special reference to precision agriculture (Vendrusculo & Kaleita 2011; Cao *et al* 2012; Tian & Li 2015; Bangui *et al* 2018, Marcu *et al* 2019). Recently, Majumdar *et al* (2017) analyzed the agricultural big data for finding optimal parameters to maximize the crop production using clustering based data mining techniques. Since the agricultural data frequently exhibit fuzzy characteristics, the use of the fuzzy algorithms and validation techniques are required for clustering applications on agricultural datasets.

Cluster analysis aims to divide a dataset into c (or k), numbers of groups by using the hierarchical or non-hierarchical clustering algorithms. As a result of clustering, similar set of data points are brought together to form subgroups which are called as clusters. In a partitioning cluster analysis, partitioning with the number of clusters that are actually present in an examined dataset or at least with an approximation of it results in good quality of clustering. For this reason, the quality of a clustering is checked via a process called as cluster validation. It is mainly carried out for three purposes:

a) To search the number of clusters giving the optimal clustering result for a dataset,
b) To understand which of the two or more clustering algorithms applied to the same dataset is better,
c) To decide which levels of the parameters, i.e. the amount of fuzziness, perform as the best with an examined algorithm.

Usually the validity indices are classified as 'external indices', 'internal indices' and 'relative indices' (Kovács *et al* 2005; Rendón *et a*l 2011). The external indices compare the obtained classes from a clustering session with the previously known classes (Dudoit & Fridlyand 2002). In this case, it is already known which data points belong to which clusters, and this information is used as a reference for validation of clustering quality. These indices are very useful in evaluating the success of a clustering

[1] Zeynel Cebeci
Çukurova University, Adana - Turkey
zcebeci@cukurova.edu.tr

algorithm because the real pattern of clusters is known (Liu *et a*l 2010). On the other hand, the internal indices do not require any external information, and determine the validity of the clustering results using the analyzed data only (Thalamuthu *et al* 2005). Finally, the relative indices compare the results from the runs of one or more clustering algorithms with different input parameters on the same dataset.

The internal indices are often used to assess the clustering quality because clustering is an unsupervised learning technique, that is, it is used to determine the clustering tendencies on a dataset when its structure is unknown. In the literature, various internal validity indices have been proposed for validating partitioning clustering results. Many of them have been listed and reviewed in detail in several studies (Milligan & Cooper 1985; Halkidi *et al* 2001; Liu *et al* 2010; Rendón *et al* 2011; Charrad *et al* 2012). As given in some comparative studies (Arbelaitz *et al* 2013; Van Craenendonck & Blockeel 2015; Hämäläinen *et al* 2017), most of the existing internal indices are available to use with traditional K-means and its derivative hard clustering algorithms. Hence, they cannot be efficient in assessing the results of fuzzy clustering algorithms such as Fuzzy C-means (FCM) and its successors. In this regard, a taxonomy of the internal validity indices for hard and soft clustering can be viewed in the related literature (Bensaid *et al* 1996, Halkidi *et al* 2002).

The indices such as partition coefficient and partition entropy have been originally introduced with FCM. Later, the more efficient indices have been developed to improve the performance in finding fuzzy partitions in datasets (Wang & Zhang 2007). However, each of the cluster validation indices has a number of pros and cons because the performance can be varied depending on different factors such as shape, volume, orientation and number of the clusters in the examined datasets. Although the above mentioned factors were carried out in the most of the comparison works (Bataineh *et al* 2011; Zhou *et a*l 2014; Zhu *et al* 2019), the effect of the distances between fuzzy data points and cluster centers has not been taken into account yet. But our intuition is that the effectiveness of the fuzzy internal validity indices can be also influenced by the distances of fuzzy points to the cluster centers. So, this study aims to compare the performances of fuzzy internal validity indices using the results from FCM clustering algorithm on some simulated datasets containing different number of clusters with different distances between a fuzzy data point and the cluster centers.

2. Related Works

Let $\boldsymbol{X}$ be a numeric dataset of n data objects in the p-dimensional data space R.

$$\boldsymbol{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\} \subseteq \mathrm{R}^p \tag{1}$$

In Equation 1:

n is the number of data objects in the dataset $\boldsymbol{X}$, $1 \leq n \leq \infty$

p is the number of features (or variables) which describe the data objects,

$\boldsymbol{x}_j$ is the feature vector of p-length for the data object j.

The probabilistic and possibilistic clustering algorithms partition a given dataset $\boldsymbol{X}$ into c, a predefined number of clusters through the minimization of their related objective functions with some probabilistic or possibilistic constraints. In the clustering context, clusters are mostly represented by their prototypes. A prototype is generally the center of a cluster which can be either centroids or medoids. The prototypes of clusters are provided in the prototypes matrix, $\boldsymbol{V}$.

$$\boldsymbol{V}=\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_c\} \subseteq \mathrm{R}^{\mathrm{n}} \tag{2}$$

In Equation 2:

c is the number of clusters, $1 \leq c \leq n$

$\boldsymbol{v}_i$ is the prototype vector of p-length for the cluster i.

The probabilistic and possibilistic partitioning clustering algorithms start with the initialization of a cluster prototype matrix $\boldsymbol{V}$, and updates it through the iteration steps until it is stabilized. The clustering algorithms compute the membership degrees of data objects by using some distance metrics for calculation of their proximities to the cluster centers. A distance measure, $d(\boldsymbol{x}_j,\boldsymbol{v}_i)$, represents the distance between the data object $\boldsymbol{x}_j$ and cluster prototype $\boldsymbol{v}_i$. In general, the squared Euclidean distance metric is used in most of the applications:

$$d_{euc.sq}(\boldsymbol{x}_j,\boldsymbol{v}_i) \rightarrow d^2(\boldsymbol{x}_j,\boldsymbol{v}_i) = \| \boldsymbol{x}_j\text{-}\boldsymbol{v}_i \|^2 = (\boldsymbol{x}_j\text{-}\boldsymbol{v}_i)^T (\boldsymbol{x}_j\text{-}\boldsymbol{v}_i) \quad (3)$$

The clustering algorithms are usually run with the squared Euclidean distance norm, which induces hyper-spherical clusters. Therefore they are able find the clusters with the same shape and orientation because the norm inducing matrix is an identity matrix: $\boldsymbol{A}$=$\boldsymbol{I}$. On the other hand, the distance metrics can be also employed with a diagonal norm inducing matrix $\boldsymbol{A}$=$\boldsymbol{I}$ $1/\sigma_j^2$ of $n \times n$ size. This norm matrix modifies the distances depending on the direction in which the distance is measured (Timm et al, 2004; Balasko et al 2005). In this case, the squared Euclidean distance with the norm matrix $\boldsymbol{A}$ is formulated as in Equation 4.

$$d_{euc.sq}(\boldsymbol{x}_j,\boldsymbol{v}_i) \rightarrow d^2_A(\boldsymbol{x}_j,\boldsymbol{v}_i) = \| \boldsymbol{x}_j\text{-}\boldsymbol{v}_i \|^2_A = (\boldsymbol{x}_j\text{-}\boldsymbol{v}_i)^T \mathbf{A} (\boldsymbol{x}_j\text{-}\boldsymbol{v}_i) \quad (4)$$

The partitioning clustering is based on the partition of the dataset $\boldsymbol{X}$ by minimizing the objective functions (J) of various clustering algorithms depending on a certain distance norm, cluster prototypes (or cluster centers) and other first-order conditions. Partitioning clustering algorithms are classified into two groups as hard and soft clustering algorithms. In hard clustering, each object in the dataset $\boldsymbol{X}$ can be a member of one and only one cluster. Contrarily, in soft clustering, an object is not only a member of a particular cluster but a member of all clusters with varying degrees of membership. In other words, an object is not forced to be a member of a specific cluster, on the contrary, it becomes a member of all of the clusters with some degrees ranging between 0 and 1. This fuzzification approach solves the membership problems arising due to the data objects close to the boundaries of neighbor clusters in the dataset $\boldsymbol{X}$.

2.1. Fuzzy C-Means Algorithm

Fuzzy C-Means (FCM) clustering algorithm was firstly studied by Dunn (1973) and generalized by Bezdek in 1974 (Bezdek 1981). Unlike K-means algorithm, a data point is not only the member of one cluster but also the member of all clusters with varying degrees of membership between 0 and 1. FCM is an iterative clustering algorithm that partitions the dataset into a predefined c clusters by minimizing the weighted within group sum of squared errors. The objective function of FCM can be expressed in Equation 5.

$$J_{FCM}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \, d^2\left(\boldsymbol{x}_j,\boldsymbol{v}_i\right) \quad (5)$$

In Equation 5:

$\boldsymbol{v}_i$ is the prototype (centers) vector for cluster i,
$\boldsymbol{x}_j$ is the feature vector for data object j,
$d^2(\boldsymbol{x}_j, \boldsymbol{v}_i)$ is the Euclidean distances between prototype $\boldsymbol{v}_i$ and the data object $\boldsymbol{x}_j$,
u_{ij} is the fuzzy membership degree of object j to the cluster i,
m is the parameter of fuzzy exponent.

In the objective function J_{FCM}, the fuzzifier exponent m is usually set to 2. However, it can be any positive real number: $1 \leq m \leq \infty$. The higher values of m result with fuzzier clusters while lower values of it give harder clusters. If it equals to 1, FCM becomes a hard algorithm and produces the same results with K-means clustering.

FCM must satisfy the constraints given in the formulas in 6, 7 and 8.

$$u_{ij} \in [0,1];\ \ 1 \leq i \leq c;\ 1 \leq j \leq n \quad (6)$$

$$0 < \sum_{j=1}^{n} u_{ij} < n;\ \ 1 \leq i \leq c \quad (7)$$

$$\sum_{i=1}^{c} u_{ij} = 1;\ \ 1 \leq j \leq n \quad (8)$$

In FCM, membership degrees and cluster prototypes are minimized by updating them with Equation 9 and 10, respectively.

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d^2(x_j,v_i)}{d^2(x_j,v_k)}\right)^{\frac{1}{m-1}}\right]^{-1} \qquad 1 \leq i \leq c;\ 1 \leq j \leq n \quad (9)$$

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} u_{ij}^{m}\, x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad 1 \le i \le c \tag{10}$$

In fuzzy clustering, there are two sources of the fuzziness in a clustering result. The first one is the overlapping degree of the clusters in the analyzed dataset. With Equation 9, a data point which has equal distance to the cluster centers of two overlapped clusters becomes the member of both in equal degree of membership. Secondly, when the proportions seen in Equation 9 are too high, the value becomes cramped around 1.

FCM has been a workhorse for fuzzy clustering in numerous application. However, it has been found that it is sensitive to noise and outliers in datasets. In order to avoid this problem, Krishnapuram and Keller (1993, 1996) proposed the Possibilistic C-Means (PCM) algorithm that relaxes the probabilistic constraint of FCM but it can generate coincident clusters with poor initializations. Hence, some other versions of FCM and PCM have been developed to eliminate the problem with the original PCM. Fuzzy Possibilistic C-Means (FPCM) algorithm (Pal *et al* 1997) and later the Possibilistic Fuzzy C-Means (PFCM) algorithm (Pal *et al* 2005) were proposed to overcome the noise sensitivity defect of FCM and the coincident clusters problem of PCM, and the row sum constraints problem of FPCM. The Possibilistic Clustering Algorithm (PCA) was proposed to improve FCM and PCM (Yang & Wu 2006). Wu *et al* (2010) introduced the Unsupervised Possibilistic Clustering (UPFC) algorithm as an extension of PCA. UPFC is an algorithm that tries to improve the noise sensitivity problem of FCM and the coincident clusters problem of PCM. It also has the advantage that it does not need an FCM initialization for possibilistic part of the clustering. Although several probabilistic and possibilistic algorithms are available for fuzzy clustering, the basic FCM is used in this study because the problematic factors that may affect the success of FCM are controlled with the simulation of datasets.

2.2. Internal Validity Indices for Fuzzy Clustering

Since clustering aims to maximize intra-class similarity and inter-class difference, the validity indices measure the compactness and separation of clusters after a clustering session. Compactness is a measure of how the data points in a cluster are interrelated or adherent to each other. Separation reveals how much a cluster is separated or distinct from the others. So, the low compactness and high degree of separation indicate a good quality of clustering. The internal validity indices compared in this study are described below.

Partition Coefficient (PC) can be considered as the first validity index proposed by Bezdek who also developed the basic fuzzy clustering (FCM) algorithm (Bezdek 1974a). Because it is calculated only from fuzzy membership values, PC is a computationally low-cost index as formulated in Equation 11. Although it is simple, its effectiveness is comparable to the indices PE and XB when the clusters are spherical in the dataset. It was even concluded that if the number of clusters to start an algorithm is chosen larger than the actual one, PC may be better than the index XB (Cebeci & Yildiz 2015).

$$I_{PC}(\boldsymbol{U}) = \frac{1}{n}\left(\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\right) \tag{11}$$

An index value is computed in the range $[1/c, 1]$ using the I_{PC} formula in Equation 11. The index value of $max\{I_{PC}(c_i)\}$; $2 \le i \le c_{max}$ gives the best clustering. Here, the number 2 and c_{max} respectively denote the minimum and the maximum number of clusters to start the FCM runs. c_i is the i^{th} number of clusters in this range. While a lower value of an index which is close to the lower boundary of the range indicates fuzzier clusters, hard clusters are obtained as it approaches to upper boundary, 1. If the index is equal to $1/c$, all members of the clusters have equal membership degrees ($u_{ij} = 1/c$) that indicates that there is no clustering tendency or the applied clustering algorithm fails to find the clusters in the given dataset (Halkidi *et al* 2002).

Modified Partition Coefficient (MPC) was proposed by Dave (1996) in order to reduce the monotonic decreasing tendency of PC depending on the magnitude of *c*.

$$I_{MPC}(\boldsymbol{U}) = 1 - \frac{1}{c-1}(1 - I_{PC}) \tag{12}$$

I_{MPC} index values are in the range [0, 1]. The value of $max\{I_{MPC}(c_i)\}$; $2 \le i \le c_{max}$ indicates the best clustering result.

Partition Entropy (PE) proposed by Bezdek (1974a) is a simple entropy based index which is calculated as in Equation 13.

$$I_{PE}(\boldsymbol{U}) = \frac{1}{n}\left(\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\ log_b(u_{ij})\right) \tag{13}$$

In Equation 13, b represents a logarithm base. I_{PE} values are obtained in the range $[0, log_b(c)]$. Unlike *PC* index, smaller I_{PE} values show the presence of well separated clusters while cluster structures become fuzzy if I_{PE} values approach to the upper boundary of the range. An I_{PE} index value equal to $log_b(c)$ indicates that there is no clustering tendency in the dataset or the used algorithm fails to partition data completely. Therefore, $min\{I_{PE}(c_i)\}$; $2 \leq i \leq c_{max}$ is the index value giving the best quality of clustering.

The indices PC and PE show monotonic dependence on c, the number of clusters used to start FCM runs. By the number of clusters, while a peak is searched in the PC graph a pit point is searched by the number of clusters in the PE graph. Both indices are sensitive to the fuzziness parameter m. Thus, as m goes to 1, the indices give the same values for all c's. However, when m goes to ∞, both indices show a significant peak or valley at $c = 2$. Another disadvantage of both indices is that they do not take the structural information and shapes of clusters into account because the dataset $\boldsymbol{X}$ is not used in calculation of these indices.

Fukuyama & Sugeno (1989) suggested an internal index to fix the problems with the indices PE and PC. As it is seen in Equation 14, the first term of the Fukuyama-Sugeno Index (FS) is the compactness of clusters while the second term is the separation measure that indicates the distances of the cluster representatives from each other.

$$I_{FS}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2 - \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|\boldsymbol{v}_i - \frac{1}{c}\sum_{k=1}^{c}\boldsymbol{v}_k\right\|^2 \tag{14}$$

In Equation 14, the first term takes into account geometry in the representation of $\boldsymbol{X}$ with respect to the prototypes in $\boldsymbol{V}$, and fuzziness provided with $\boldsymbol{U}$. The second term adds the distance of the prototypes from the overall mean and the fuzziness in each row of $\boldsymbol{U}$. Since smaller I_{FS} values indicate the presence of compact and well-separated clusters, $min\{I_{FS}(c_i)\}$; $2 \leq i \leq c_{max}$ gives the most successful clustering result. Pal and Bezdek (1995) reported that the index FS is sensitive to both low and high values of parameter m.

Xie and Beni (1991) developed the fuzzy validity index which is known as the Xie-Beni Index (XB) in Equation 15 when the parameter m is set to 2. The numerator of the equation considers the distance of objects in a cluster from their cluster centers and measures the compactness of fuzzy clustering. The denominator term represents the strength of the separation of clusters with the distance between cluster centers.

$$I_{XB}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2}\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2}{n\left(\min\limits_{i,k=1,..,c;\ i\neq k}\{\|\boldsymbol{v}_i - \boldsymbol{v}_k\|^2\}\right)} \tag{15}$$

Smaller values of the index XB indicate more compact and well-separated clusters. However, the XB index decreases monotonically as c approaches to n. In order to eliminate this tendency, a c_{max} value is determined as the starting point of monotonic behavior, and then $min\{I_{XB}(c_i)\}$; $2 \leq i \leq c_{max}$ is used to find the best clustering result. Another disadvantage of the index XB is that it goes to infinity as m also goes to infinity.

Kwon's validity index (K) eliminates the problem of monotonic decreasing tendency that occurs due to the increase in the number of clusters for the index XB. For this purpose, Kwon (1998) added a second term to the nominator of the index XB in order to penalize high cluster numbers as seen in Equation 16.

$$I_{K}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2}\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\left\|v_i - \frac{1}{n}\sum_{l=1}^{n}\boldsymbol{x}_l\right\|^2}{\min\limits_{i\neq k}\{\|\boldsymbol{v}_i - \boldsymbol{v}_k\|^2\}} \tag{16}$$

The optimal clustering for the index K is investigated with $min\{I_K(c_i)\}$; $2 \leq i \leq c_{max}$.

Chen & Linkens (2004) proposed the Chen-Linkens index (CL) that consists of two terms. As it is seen in Equation 17, the first term of CL formula reflects the compactness within a cluster. The second term indicates the separation degree between the clusters. The optimal clustering is obtained at the maximum value of $max\{I_{CL}(c_i)\}$; $2 \le i \le c_{max}$.

$$I_{CL}(\boldsymbol{U}) = \frac{1}{n}\sum_{j=1}^{n}\max_{j}\left(u_{ij}\right) - \frac{1}{\sum_{i=1}^{c-1} i}\sum_{i=1}^{c-1}\sum_{l=i+1}^{c}\left(\frac{1}{n}\sum_{l=1}^{n}\min\left(u_{ij}, u_{lj}\right)\right) \qquad (17)$$

Fuzzy Hypervolume (FHV) or Gath-Geva clustering validity index is based on the hypervolume and density of clusters in a given dataset. The index FHV is formulated as in Equation 18 (Gath & Geva 1989).

$$I_{FHV}(\boldsymbol{X}; \boldsymbol{V}, \boldsymbol{U}) = \left(\sum_{i=1}^{c}[det(F_i)]\right)^{1/2} \qquad (18)$$

In Equation 18, F_i is the fuzzy covariance matrix of the cluster i. It is calculated as seen in Equation 19.

$$F_i = \frac{\sum_{j=1}^{n} u_{ij}^{m}(x_j - \boldsymbol{v}_i)^T(x_j - \boldsymbol{v}_i)}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (19)$$

F_i is used as a measure of compactness. If the clusters are soft, a fuzzy clustering with lower I_{FHV} is expected. Hence, the value of $min\{I_{FHV}(c_i)\}$; $2 \le i \le c_{max}$ indicates the best clustering.

Pakhira et al. (2004, 2005) proposed the validity indices that can be used in both hard and soft clustering. In order to differentiate these, Pakhira-Bandyopadhyay-Maulik (PBM) Index for fuzzy clustering is denoted as PBMF.

$$I_{PBMF}(\boldsymbol{X}; \boldsymbol{V}, \boldsymbol{U}) = \left(\frac{1}{c}\cdot\frac{E_1}{E_c}\cdot D_c\right)^p \qquad (20)$$

In Equation 20:

$$E_1 = \sum_{j=1}^{n} u_{1j}\left\|\boldsymbol{x}_j - \boldsymbol{v}_1\right\| \qquad (21)$$

$$E_c = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\| \qquad (22)$$

$$D_c = \max_{i,k=1,\ldots,c;\ i\neq k}\left\|\boldsymbol{v}_i - \boldsymbol{v}_k\right\| \qquad (23)$$

The authors of the index PBMF argue that the first term in Equation 20 reduces the value of index as c is increases. The second term is the ratio of E_1 to E_c. As formulated in Equation 21, E_1 is a constant value for a given dataset. Since the ratio decreases with an increase in c, the value of index value increases as E_c decreases which indicates more numbers of compact clusters. The third term, D_c with the formula in Equation 23, measures the maximum separation between two clusters over all possible pairs of clusters. It increases with the value of c. The power p controls the contrast between the different cluster configurations, and is assigned as 2 in general. The optimal clustering with I_{PBMF} is obtained at the maximum value of $max\{I_{PBMF}(c_i)\}$; $2 \le i \le c_{max}$.

The soft version of the Silhouette index (FSIL) can also be used in the assessment of fuzzy clustering results with the formula given in Equation 24.

$$I_{FSIL}(\boldsymbol{X}; \boldsymbol{V}, \boldsymbol{U}) = \frac{\sum_{i=1}^{n}\left(u_{ij} - u_{ij'}\right)^{\alpha} s_i(c)}{\sum_{i=1}^{n}\left(u_{ij} - u_{ij'}\right)^{\alpha}} \qquad (24)$$

In Equation 24, $s_i(c)$ is the silhouette value for the data point i. It is calculated with the formula shown in Equation 25.

$$s_i(c) = \frac{b_i - a_i}{\max(b_i, a_i)} \qquad (25)$$

In Equation 25, a_i is the average dissimilarity between the data point i and all of the remaining data points in the same cluster. b_i is the least mean dissimilarity between data point i and the data points in other clusters. The membership degrees u_{ij} and $u_{ij'}$ are the first and second largest values in the i^{th} row of the membership degrees matrix $\boldsymbol{U}$, respectively. α is a weighing coefficient, generally set to 1. The optimal clustering is proposed at $max\{I_{SILF}(c_i)\}$; $2 \le i \le c_{max}$.

The Average Partition Density (APD) index, proposed by Gath and Geva (1989) is formulated in Equation 26. In the formula, $\boldsymbol{x}_j$ is the set of data points within a predefined region around the center of cluster i, which is the sum of the central members of cluster i. The best clustering is obtained with $max\{I_{APD}(c_i)\}$; $2 \leq i \leq c_{max}$.

$$I_{APD}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \frac{1}{c}\sum_{i=1}^{c}\left(\frac{\sum_{x\in X_j} u_{ij}}{\boldsymbol{v}_i}\right) \quad (26)$$

The validity-guided (re)clustering (VGC) algorithm uses cluster-validity information to guide a fuzzy (re)clustering process toward better solutions. It starts with a partition generated by a fuzzy clustering algorithm and then iteratively alters the partition by applying split-and-merge operations to form the clusters (Bensaid *et al* 1996). The authors proposed the Compactness / Separation (CS) ratio index in order to validate the results of VGC. The formula of CS index is shown in Equation 27. The value of $min\{I_{CS}(c_i)\}$; $2 \leq i \leq c_{max}$ gives the best clustering result.

$$I_{CS}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{n} u_{ij}^{m} d^2(\boldsymbol{x}_j,\boldsymbol{v}_i)}{\sum_{j=1}^{n} u_{ij}\sum_{l=1}^{c}\|\boldsymbol{v}_i-\boldsymbol{v}_l\|^2} \quad (27)$$

3. Experimental Works on the Simulated Datasets

For experimental testing of the above mentioned research question, three artificial datasets having two, three and four clusters have been generated in order to validate the performances of the studied validity indices. In each of the clusters in all of the simulated datasets, five data points are created, one of which is the centroid. The non-centroid data points of the clusters are defined one unit away from the centroid. In other words, the radius of the clusters are 1. As illustrated with a red circle in Figure 1, a fuzzy data point positioned at an equal distance from the centroids of all the clusters has been added in each generated dataset.

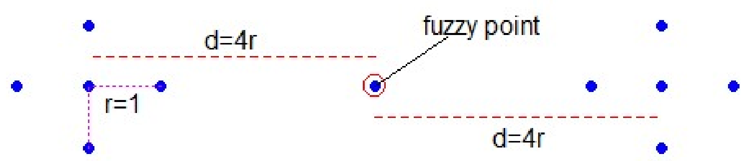


Figure 1. Pattern of the clusters and the location of fuzzy point in the simulated dataset c2r4

Five different versions of each dataset are generated by placing the fuzzy object 2 to 6 units away from the centroids of clusters. As seen in the rows of Figure 2, these datasets are named as c2r2, c3r4, etc. For instance, c3r5 stands for the dataset having three clusters in which the fuzzy point is 5 units (d = 5 × radius) away from the centroids of clusters.

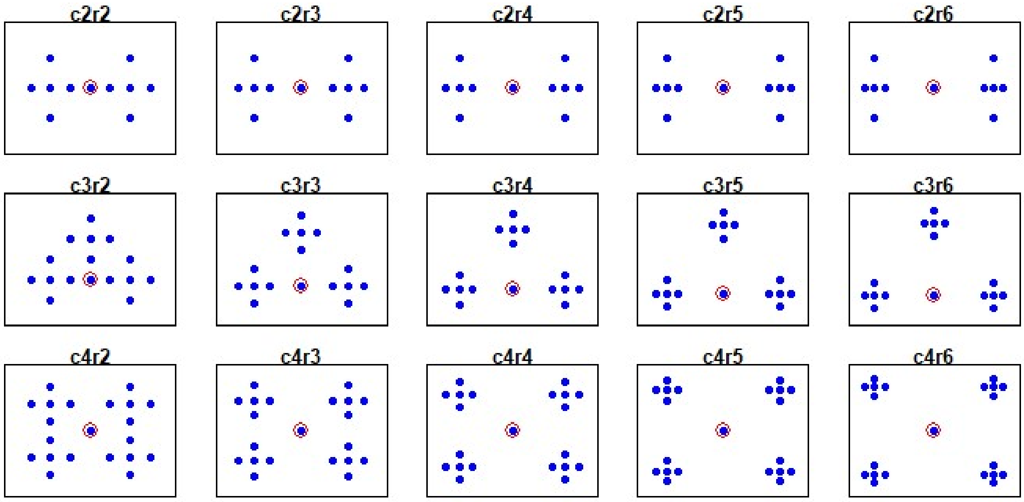


Figure 1. Scatter plots of the cluster structures in the artifical datasets

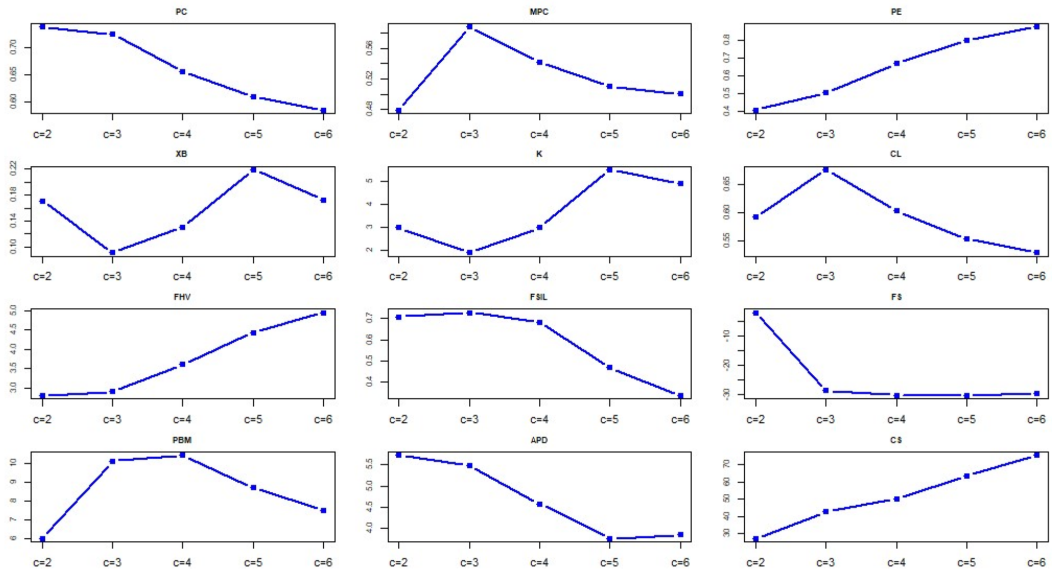


Figure 2. Changes in the values of indices by the number of clusters on the dataset c3r3

In this study, the data analysis has been performed in R environment (R Core Team, 2019). The function `fcm` of the package `ppclust` (Cebeci *et al* 2017) has been run for FCM clustering. In order to decide an optimal value of number of clusters for an examined dataset, cluster analysis should be repeated for a range of number of clusters (c). For this reason, the function `fcm` has been run for five levels of c (from 2 to 6). For each dataset, the function `fcm` has been started with the k-means++ method (Arthur & Vassilvitskii, 2007) to initialize the cluster prototype matrix $\boldsymbol{V}$. The membership matrix $\boldsymbol{U}$ has been initialized using a novel fast initialization method proposed by Cebeci (2018). The fuzzification parameter m has been set to 2 in all the FCM runs. The relevant functions of the package `fcvalid` (Cebeci & Cebeci 2018), downloaded from GitHub have been used for validation of fuzzy clustering result at the end of each FCM run. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ from these successive runs of FCM, which produce the minimum objective function value have been used to evaluate the performance of validity indices. The values of indices returned by the validity functions have been checked to obtain the proposed number of clusters for each dataset, as exemplified in Figure 2 and Table 1 for the dataset containing 3 clusters in which the cluster centroids are 3 units away from the fuzzy point.

Table 1. Proposed number of clusters for the dataset c3r3

	Number of clusters examined in the FCM runs				
Index	c=2	c=3	c=4	c=5	c=6
I_{PC}	**0.740**	0.7253	0.657	0.608	0.584
I_{MPC}	0.480	**0.5880**	0.542	0.510	0.501
I_{PE}	**0.409**	0.5066	0.668	0.799	0.873
I_{XB}	0.171	**0.0908**	0.130	0.219	0.172
I_{K}	2.989	**1.8980**	2.988	5.509	4.924
I_{CL}	0.591	**0.6763**	0.603	0.553	0.528
I_{FHV}	**2.797**	2.8985	3.602	4.435	4.950
I_{FSIL}	0.708	**0.7289**	0.682	0.468	0.336
I_{FS}	-1.966	-28.8361	**-30.330**	-30.570	-29.729
I_{PBMF}	6.023	10.1041	**10.421**	8.718	7.485
I_{APD}	**5.720**	5.4909	4.582	3.761	3.848
I_{CS}	**27.252**	42.6798	50.115	63.393	75.459

As it is seen in Table 2, the indices of PE, XB, K and FSIL propose the actual number of clusters for all the datasets with two clusters. The index APD is also successful to find the actual number of clusters except the dataset c2r4. The indices PC, FHV and MPC overestimate the number of clusters for the dataset c3r6. The indices CL, FS, PBM and CS propose the number of clusters one more than the actual number of clusters in the datasets. The index FS is the worst to detect the fuzzy partitions for the dataset c2r2. The indices PE, XB, K and FSIL are stable regarding the change of distance levels of the cluster centers from the fuzzy point, and totally successful to find the actual number of clusters.

Table 2. Proposed number of clusters for the datasets with two clusters

	Distance of the fuzzy point to the cluster centers (d)				
Index	2	3	4	5	6
I_{PC}	**2**	**2**	**2**	**2**	3
I_{MPC}	**2**	**2**	**2**	3	3
I_{PE}	**2**	**2**	**2**	**2**	**2**
I_{XB}	**2**	**2**	**2**	**2**	**2**
I_{K}	**2**	**2**	**2**	**2**	**2**
I_{CL}	**2**	**2**	3	3	3
I_{FHV}	**2**	**2**	**2**	3	3
I_{FSIL}	**2**	**2**	**2**	**2**	**2**
I_{FS}	6	**2**	3	3	3
I_{PBMF}	**2**	**2**	3	3	3
I_{APD}	**2**	**2**	3	**2**	**2**
I_{CS}	**2**	**2**	3	3	3

According to the results in Table 3, the indices K and FSIL calculate the actual number of clusters at all the distance levels between the cluster centers and the fuzzy point. The indices of PE and FHV find the actual number of clusters except for the result on the dataset c3r2. All the indices except PE, K, FHV, and FSIL overestimate the number of clusters for the dataset having the clusters whose centers are 6 units away from the fuzzy point. The index APD also performs well except the datasets c3r2 and c3r6.

Table 3. Proposed number of clusters for the datasets with three clusters

	Distance of the fuzzy point to the cluster centers (d)				
Index	2	3	4	5	6
I_{PC}	2	**3**	**3**	**3**	4
I_{MPC}	**3**	**3**	**3**	4	4
I_{PE}	2	**3**	**3**	**3**	**3**
I_{XB}	**3**	**3**	**3**	**3**	4
I_{K}	**3**	**3**	**3**	**3**	**3**
I_{CL}	**3**	**3**	**3**	4	4
I_{FHV}	2	**3**	**3**	**3**	**3**
I_{FSIL}	**3**	**3**	**3**	**3**	**3**
I_{FS}	4	**3**	**3**	**4**	4
I_{PBMF}	4	**3**	**3**	4	4
I_{APD}	2	**3**	**3**	**3**	**4**
I_{CS}	2	2	**3**	4	4

For the experimental datasets having four clusters, the indices FSIL and FS have superior performance to obtain the actual number of clusters at all the levels of distances between the fuzzy point and the cluster centers in all the datasets. The indices K, CL and FHV also perform well for all the distances. The index APD underestimates the number of clusters for the distance level of 2 but overestimates for the distance levels of 5 and 6. It is also interesting that the index APD overestimates the number of clusters for all the distances in the dataset having 4 clusters when compared to the datasets having 2 and 3 clusters in which it performs well. The index CS is also totally unsuccessful to find the actual number of clusters for all the distance levels.

Table 4. Proposed number of clusters for the datasets with four clusters

	Distance of the fuzzy point to the cluster centers (d)				
Index	2	3	4	5	6
I_{PC}	2	**4**	**4**	5	5
I_{MPC}	**4**	**4**	**4**	5	5
I_{PE}	2	2	**4**	**4**	5
I_{XB}	5	5	5	5	5
I_{K}	**4**	5	**4**	**4**	**4**
I_{CL}	**4**	**4**	**4**	**4**	5
I_{FHV}	2	**4**	**4**	**4**	**4**
I_{FSIL}	**4**	**4**	**4**	**4**	**4**
I_{FS}	**4**	**4**	**4**	**4**	**4**
I_{PBMF}	**4**	**4**	**4**	5	5
I_{APD}	2	**4**	**4**	5	5
I_{CS}	2	2	5	5	5

4. Conclusions

The indices K and FSIL have successfully discovered the actual number of clusters for all levels of the distances between the fuzzy point and the cluster centers. The index FHV also performs well for the cases in which the fuzzy point is neither so close nor so far from the centers of clusters. The indices PC, MPC and PE tend to calculate the higher number of clusters than the actual ones in the cases where the fuzzy point moves away from the cluster centers. The index XB tends to overestimate the number of clusters if a dataset contains more clusters within. This result indicates that the index XB loses its validation ability for greater values of the distances. The indices FS, PBM, APD and CS return accurate results if the fuzzy point is moderately distant (d = 3-4 × avg. radius of clusters) from the cluster centers, otherwise they may not work well. These results show that these indices may not perform well if fuzzy points are too far from cluster centers.

As a general conclusion, when compared to the others, the indices FSIL, FHV and K are more stable in validating fuzzy clustering results. Additionally, the results also show that average distance between fuzzy points and cluster centers should be taken into consideration to keep the effectiveness of fuzzy validity indices more stable. A future work can enhance the information on the efficiencies of internal validity indices for fuzzy clustering results on the real datasets, sourced from various IoT- based agricultural activities. So, we plan to test the performances of the fuzzy internal validity indices on real agricultural datasets. In addition, the future work will also consider to compare the performances of the examined indices for different orientations and volumes of clusters in simulated and real datasets.

Acknowledgement

This research was supported by the grant FBA-2019-10285 from the Unit of Scientific Research Projects of Çukurova University, Adana – Turkey.

References

Arbelaitz, O, Gurrutxaga, I, Muguerza, J, Pérez, JM & Perona, I 2013 'An Extensive Comparative Study of Cluster Validity Indices'. *Pattern Recognition*, vol. 46, no. 1, pp. 243-256. doi: 10.1016/j.patcog.2012.07.021

Arthur, D & Vassilvitskii, S 2007 'K-means++: The Advantages of Careful Seeding', in Proc. of *the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035.

Balasko, B, Abonyi, J & Feil, B 2005. Fuzzy clustering and data analysis toolbox. Department of Process Engineering, University of Veszprem, Veszprem. Accessed online at https://pdfs.semanticscholar.org/72f6/b22f6db1c2c0c47d8e6ead009b8c4c42bad9.pdf

Bangui, H, Ge, M, & Buhnova, B 2018 'Exploring Big Data Clustering Algorithms for Internet of Things Applications'. In Proc. of the 3rd *Int. Conf. on Internet of Things, Big Data and Security (IoTBDS 2018)*, pp. 269-276. https://doi.org/10.5220/0006773402690276

Bataineh, KM, Naji, M & Saqer, M 2011 'A Comparison Study between Various Fuzzy Clustering Algorithms'. *Jordan J. of Mechanical & Industrial Engineering*, vol. *5*, no. 4, pp. 335-343.

Bensaid, AM, Hall, L., Bezdek, JC, Clarke CP, Silbiger, ML, Arrington, JA & Murtagh, RF 1996. 'Validity-Guided (Re) Clustering with Applications to Image Segmentation'. *IEEE Transactions on Fuzzy Systems,* vol. 4, no. 2, pp. 112–123. doi: 10.1109/91.493905

Bezdek, JC 1974a 'Cluster Validity with Fuzzy Sets'. *J. Cybernet*ics, vol. 3, pp. 58–73. https://doi.org/10.1080/01969727308546047

Bezdek, JC 1974b 'Numerical Taxonomy with Fuzzy Sets'. *J. Math. Biol.,* vol. 1, pp. 57–71. https://doi.org/10.1007/BF02339490

Bezdek, JC 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York. https://doi.org/10.1007/978-1-4757-0450-1

Borgelt, C & Kruse, R 2005 'Fuzzy and Probabilistic Clustering with Shape and Size Constraints'. *Proc. of the 11th Int. Fuzzy Systems Association World Congress*, Beijing, China, pp. 945-950.

Cao, L, Zhang, X, San, X, Zhao, Y & Chen, G 2012 'Application of Fuzzy Clustering Algorithm in Precision Agriculture'. In *World Automation Congress 2012*, pp. 1-4. IEEE.

Cebeci Z & Yildiz F 2015 'Bulanık C-Ortalamalar Algoritmasının Farklı Küme Büyüklükleri için Hesaplama Performansı ve Kümeleme Geçerliliğinin Karşılaştırılması', *9. Ulusal Zootekni Bilim Kongresi,* 3-5 Eylül 2015, Konya - Türkiye. Bildiriler Kitabı, pp. 227-239. doi: https://doi.org/10.13140/RG.2.1.2909.9288

Cebeci, Z, Kavlak, AT & Yildiz, F 2017 'Validation of fuzzy and possibilistic clustering results', in Proc. of *2017 Int. Artificial Intelligence & Data Processing Symposium*, IEEE. pp. 1-7. doi: 10.1109/IDAP.2017.8090183

Cebeci, Z 2018 'Initialization of Membership Degree Matrix for Fast Convergence of Fuzzy C-Means Clustering' in Proc. of *2018 Int.Conf.on Artificial Intelligence & Data Processing,* pp. 1-5. IEEE. doi: 10.1109/IDAP.2018.8620920

Cebeci, Z & Cebeci, C 2018 'kpeaks: an R package for quick selection of k for cluster analysis' in Proc. of *2018 Int. Conf. on Artificial Intelligence & Data Processing*. pp. 1-7, IEEE. doi: 10.1109/IDAP.2018.8620896.

Charrad, M, Ghazzali, N, Boiteux, V & Niknafs, A 2014 'NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set'. *J. of Statistical Software*, vol. 61, no. 6. doi: 10.18637/jss.v061.i06

Chen, MY & Linkens, DA 2004 'Rule-Base Self-Generation and Simplification for Data-Driven Fuzzy Models'. *Fuzzy Sets and Systems*, vol. 142, no. 2, pp. 243-265. doi: 10.1016/S0165-0114(03)00160-X

Chiang, M-T. 2009 'Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads'. *J.of Classification* vol. 27, no. 3-4, pp. 3-40. doi: 10.1007/s00357-010-9049-5

Chou, C-H, Su, M-C & Lai, E 2003 'A New Cluster Validity Measure for Clusters with Different Densities', In *Proc. IASTED Int. Conf. Intell. Syst. Control*, pp. 276 -281.

Chou, C-H, Su, M-C & Lai, E 2004 'A New Cluster Validity Measure and its Application to Image Compression', *Pattern Anal. App.*vol. 7, pp. 205–220. doi: 10.1007/s10044-004-0218-1

Correa C, Valero, C, Barreiro P, Diago, MP & Tardaguila, J 2011 'A Comparison of Fuzzy Clustering Algorithms Applied to Feature Extraction on Vineyard'. In Proc. of the Conference of the Spanish Association for Artificial Intelligence, Spain, Nov 11, 2011. Accessed online at http://oa.upm.es/9246.

Dave, RN 1996, 'Validating Fuzzy Partition Obtained Through C-Shells Clustering', *Pattern Recognition Lett.* vol. 17, no. 6, pp. 613–623. doi: 10.1016/0167-8655(96)00026-8

Di Martino, F & Sessa, S 2009 'Implementation of the Extended Fuzzy C-Means Algorithm in Geographic Information Systems'. *J. of Uncertain Systems*, vol. 3, no. 4, pp. 298-306.

Dudoit, S & Fridlyand, J 2002 'A Prediction-based Resampling Method for Estimating the Number of Clusters in a Dataset'. *Genome Biology*, vol. 3, no. 7, pp. 1-21. doi: 10.1186/gb-2002-3-7-research0036

Dunn, JC 1973 'A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters'. *J. of Cybernetics*, vol.3, no.3, pp. 32-57. doi: 10.1080/01969727308546046

Fukuyama, F & Sugeno, M 1989 'A New Method of Choosing the Number of Clusters for the Fuzzy C-Means Method', In *Proc. of 5th Fuzzy Systems Symposium*, 1989, pp. 247–250.

Gath, I & Geva, AB 1989 'Unsupervised Optimal Fuzzy Clustering', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 11, no. 7, pp. 773–781. doi: 10.1109/34.192473

Halkidi, M, Batistakis, Y & Vazirgiannis, M 2002 'Clustering Validity Checking Methods: Part II', *ACM SIGMOD Record*, vol. 31, no. 3, pp. 19-27. doi: 10.1145/601858.601862.

Hämäläinen, J, Jauhiainen, S & Kärkkäinen, T 2017 'Comparison of Internal Clustering Validation Indices for Prototype-based Clustering'. *Algorithms*, vol. 10, no. 3, pp. 105. doi: 10.3390/a10030105

Kovács, F, Legány, C & Babos, A 2005 'Cluster Validity Measurement Techniques'. In Proc. of 6th *Int. Symposium of Hungarian Researchers on Computational Intelligence*, Nov 18-19, 2005, Budapest, Hungary.

Krishnapuram, R & Keller, J 1993 'A Possibilistic Approach to Clustering'. *IEEE Trans. Fuzzy Systems,* vol. 1, no. 2, pp. 98-110. doi: 10.1109/91.227387

Krishnapuram, R & Keller, J 1996 'The Possibilistic C-Means Algorithm: Insights and Recommendations'. *IEEE Trans. Fuzzy Systems*, vol. 4, no. 3, pp. 385-393. doi: 10.1109/91.531779

Kwon, SH 1998 'Cluster Validity Index for Fuzzy Clustering'. *Electron. Lett.* vol. 34, no. 22, pp. 2176–2178. doi: 10.1049/el:19981523

Liu, Y, Li, Z, Xiong, H, Gao, X & Wu, J 2010 'Understanding of Internal Clustering Validation Measures'. In *Proc. of 2010 IEEE Int. Conf. on Data Mining*, pp. 911-916. doi: 10.1109/ICDM.2010.35

Majumdar, J, Naraseeyappa, S & Ankalaki, S 2017 'Analysis of Agriculture Data Using Data Mining Techniques: Application of Big Data. *Journal of Big Data*, 4(1), 20. doi: 10.1186/s40537-017-0077-4

Marcu, I, Voicu, C, Drăgulinescu, AMC, Fratu, O, Suciu, G, Balaceanu, C & Andronache, MM 2019 'Overview of IoT Basic Platforms for Precision Agriculture'. In *Proc. of Int. Conf. on Future Access Enablers of Ubiquitous and Intelligent Infrastructures*, Springer: Cham, pp. 124-137.

Milligan, GW & Cooper, MC1985 'An Examination of Procedures for Determining the Number of Clusters in a Data Set', *Psychometrika*, vol. 50, pp. 159-179. doi: 10.1007/BF02294245

Pakhira, MK, Bandyopadhyay, S & Maulik, U 2004 'Validity Index for Crisp and Fuzzy Clusters', *Pattern Recognition* vol. 37, pp. 487–501. doi: 10.1016/j.patcog.2003.06.005

Pakhira, MK, Bandyopadhyay, S & Maulik, U 2005 'A Study of Some Fuzzy Cluster Validity Indices, Genetic Clustering and Application to Pixel Classification', *Fuzzy Sets Syst*. vol. 155, no.2, pp. 191–214. doi: 10.1016/j.fss.2005.04.009

Pal, K & Bezdek, JC 1995 'On Cluster Validity for the Fuzzy C-Means Model'. *IEEE Trans. Fuzzy Syst*., vol. 3, no. 3, pp. 370-379. doi: 10.1109/91.413225

Pal, NR, Pal, K & Bezdek JC 1997 'A Mixed C-Means Clustering Model'. In *Proc. of the Sixth IEEE Int. Conf. on Fuzzy Systems*, vol. 1, pp. 11-21. doi: 10.1109/FUZZY.1997.616338

Pal NR, Pal, K, Keller JM & Bezdek JC 2005 'A Possibilistic Fuzzy C-Means Clustering Algorithm', *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517-530. doi: 10.1109/TFUZZ.2004.840099

R Core Team 2019 'R: A Language and Environment for Statistical Computing'. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.

Ramesh D & Vardhan, BV 2013 'Data Mining Techniques and Applications to Agricultural Yield Data'. *Int. J. of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 9, pp. 3477-3480.

Rendón, E, Abundez, I, Arizmendi, A & Quiroz, EM 2011 'Internal Versus External Cluster Validation Indexes' *Int. J. of Computers and Communications*, vol. 5, no. 1, pp.27-34.

Thalamuthu, A, Mukhopadhyay, I, Zheng, X & Tseng, GC 2006, 'Evaluation and Comparison of Gene Clustering Methods in Microarray Analysis', *Bioinformatics*, vol. 22, no. 19, pp. 2405-2412. doi: 10.1093/bioinformatics/btl406

Tian, Z & Li, B 2015 'An Application of Fuzzy C-Means Based Clustering Technique in Smart Farming'. In *2015 Int. Conf.e on Intell. Systems Resch & Mechatronics Eng.,* Atlantis Press. doi: 10.2991/isrme-15.2015.150

Timm, H, Borgelt, C & Kruse, R 2001 'Fuzzy Cluster Analysis with Cluster Repulsion'. In *Proc. of the European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, Tenerife, Spain. Accessed online at http://fuzzy.cs.ovgu.de/publications/doering/TimBorDoerKru01.pdf

Tiwari, M & Misra, B 2011 'Application of Cluster Analysis in Agriculture - A Review Article'. *Int. J. of Computer Applications*, vol. 36, no.4, pp. 43-47.

Van Craenendonck, T & Blockeel, H 2015 'Using Internal Validity Measures to Compare Clustering Algorithms'. In *Benelearn 2015*, Delft, NL, June 19, 2015, Poster presentations pp. 1-8. Accessed online at https://lirias.kuleuven.be/retrieve/330191

Vendrusculo, LG & Kaleita, AF 2011 'Modeling Zone Management in Precision Agriculture through Fuzzy C-Means Technique at Spatial Database'. In *2011 American Society of Agricultural and Biological Engineers.* Louisville, Kentucky, August 7-10, 2011, pp. 1. doi: 10.13031/2013.38168

Wang, W & Zhang, Y 2007 'On Fuzzy Cluster Validity Indices'. *Fuzzy Sets and Systems*, vol. 158, no. 19, pp. 2095-2117. doi: 10.1016/j.fss.2007.03.004

Wu, X, Wu, B, Sun, J & Fu, H 2010 'Unsupervised Possibilistic Fuzzy Clustering'. *J. of Information & Computational Science*, vol.7, no. 5, pp. 1075-1080.

Xie, XL & Beni, G 1991 'A Validity Measure for Fuzzy Clustering'. *IEEE Trans. Pattern Anal. Mach. Intell*. vol. 13, no. 8, pp. 841–847. doi: 10.1109/34.85677

Yang, MS & Wu, KL 2006 'Unsupervised Possibilistic Clustering'. *Pattern Recognition*, vol. 39, no. 1, pp.15-21. doi: 10.1016/j.patcog.2005.07.005

Zhang, Y, Wang, W, Zhang, X & Li, Y. 2008 'A Cluster Validity Index for Fuzzy Clustering'. *Information Sciences,* vol. 178, pp. 1205–1218. doi: 10.1016/j.patrec.2004.11.022

Zhou, K, Ding, S, Fu, C & Yang, S 2014 'Comparison and Weighted Summation Type of Fuzzy Cluster Validity Indices'. *Int. J. Computers Communications & Control*, vol. *9*, no. 3, pp. 370-378. https://doi.org/10.15837/ijccc.2014.3.237

Zhu, LF, Wang, J.S & Wang, HY 2019 'A Novel Clustering Validity Function of FCM Clustering Algorithm'. *IEEE Access*, vol. *7*, pp. 152289-152315. https://doi.org/10.1109/ACCESS.2019.2946599

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 10, No. 2
journal.magisz.org

Mathematical analysis of drone flight path

Róbert Barna[1], Kristóf Solymosi[2], Eleonóra Stettner[3]

INFO
Received 12-11-2019
Accepted 27-12-2020
Available online 28-01-2020
Responsible Editor: M. Herdon:

Keywords:
drone, overlap, footprint, geometry, GeoGebra

ABSTRACT

In practice, the drone may not always be able to follow a pre-planned flight path. Therefore, producing orthophotos can be problematic. The drone's shooting positions can be read from the captured images, so the actual flight path is known. The recorded route data and camera properties can be used to calculate the photographed area. It is a prerequisite for making an orthophoto that the adjacent images overlap a certain percentage. Calculating the area of an overlapping rectangle is lengthy by specifying the coordinates of the intersecting sides. We have developed a simpler method for this. One of the two overlapping rectangles was fixed and examined to determine the curve of the center of the other rectangle such that the overlapping area was exactly equal to the required overlap. This curve has become a hyperbola arc. This makes the calculation simpler and can be traced back to comparing the lengths of two sections. The method was developed using GeoGebra software.

1. Introduction

Nowadays, aerial photography and more specifically unmanned aerial vehicles (UAVs) extended with multispectral cameras (also called Unmanned Aerial Systems, UASs) are becoming more numerous and affordable each year giving access to temporal and spatial resolutions not seen before (Wendel et al., 2006). UAVs have long been used in various, but mostly in military applications in the past. However, due to the fact that the technology is becoming more available and affordable, and because they do not require high-skill piloting anymore, they are emerging in civilian fields of use. This branch of UAVs according to size categories are the smaller (< 20 kg) ones (Laliberte and Rango,). Aerial photography and videography (both casual and professional) can hardly be thought of without the use of drones (UAVs) nowadays. In agriculture, the mapping of fields is no exception to this. Furthermore, with the help of specialized software (such as DroneDeploy or DJI GS Pro) – and, of course, with the miniaturization of technology, e.g. GPS chips – consumers can tap into the opportunity of automatic vehicle guidance for precision agriculture and smart farming needs (Eisenbeiss et al. 2009). Acquired datasets are of course digital images, associated with GPS position and related flight information. Furthermore, with mentioning only a selected few, unmanned aerial imagery has been shown to be useful for mapping spatial variability within fields (Mulla, 2013), crop monitoring for site-specific application (Shanahan et al., 2008; Merotto et al. 2012; Rasmussen et al. 2016) and so on.

Modern practice shows that UAVs have the potential to monitor and assess agricultural and environmental variables (Zhang & Kovács 2012). The main caveat of this, however, is the need to capture a large number of overlapping photos that must be stitched together to produce an orthorectified image (to exclude optical and geometrical errors for the large orthoimage, also for the sake of geo-registering every image) of the desired area (Vericat et al. 2009). Important variables in

[1] Róbert Barna
Kaposvár University, Faculty of Economic Science, Institute of Methodology
barna.robert@ke.hu
[2] Kristóf Solymosi
Kaposvár University, Doctoral School in Animal Science
solymosi.kristof@ke.hu
[3] Eleonóra Stettner
Kaposvár University, Faculty of Economic Science, Institute of Methodology
stettner.eleonora@ke.hu

using drone imagery for mapping weeds are the image spatial resolution and the phenological stage of the crop being photographed (as higher the spatial resolution is, the finer the details are on the conjoined image as well) (Gómez & Candón 2014). It is clearly observable that using other types of remote platforms with coarse spatial resolutions, such as satellite imagery has huge limitations in precision practices.

It is very important to assess the accuracy of orthomosaiced images. Co-registration of them is usually expressed in RMSE and is dependent on the camera internal and external orientation (thus gimbal positions) and the density and distribution of ground control points (GCPs), plus the topographic complexity of the scene. (Vericat et al. 2009)

One of the main factors to have an adequate orthophoto is the amount of vertical and horizontal overlap. The photogrammetry is capable of producing 3D models of the joint area of 2D images if overlapped by at least 60% (Detrekői & Szabó, 2013).

The aircraft movement can be described relative to the aircraft body (body frame). Three perpendicular axes are defined such that the origin is the center of mass, and the X axis is directed through the front of the aircraft and the Y axis through the right of the aircraft, and the Z axis is finally through the bottom of the aircraft (forward, right and downward rotation respectively). When describing rotational movement, the X, Y and Z axes are renamed Roll (X), Pitch (Y) and Yaw (Z), and that is the nomenclature that we are using through the article (Figure 1).

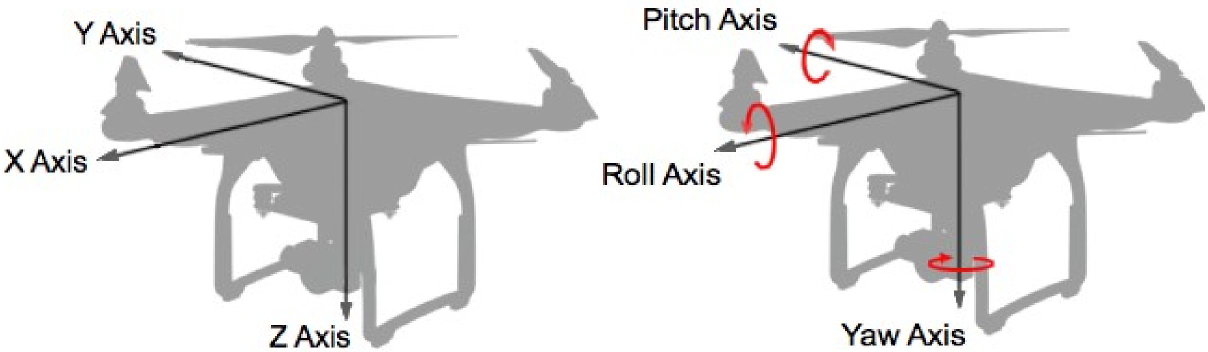


Figure 1. X,Y,Z and Roll, pitch and yaw axes on a drone
(DJI.com 2016)

The orientation of the aircraft and gimbal is known as its altitude and is defined by the rotation around the pitch, roll and yaw axes. Half of the aircraft's propellers spin clockwise, while the other half anti-clockwise. When all spin with equal speed, the aircraft heading will be constant. If one spins faster than the other, the aircraft will rotate around the yaw axis (DJI.com 2016).

1.1. Problem Definition

For surveys, drones either fly a predetermined route with a fixed exposure position (Figure 2) or freehand. The drone may not always be able to follow the planned flight path, or the required coverage may not be achieved during manual control. These may cause errors in the creation of the orthophoto and 3D terrain model. Due to the large number of in-flight shots, it takes a long time to produce an orthophoto. If the error were detected before the orthophoto was produced, it could save much time.

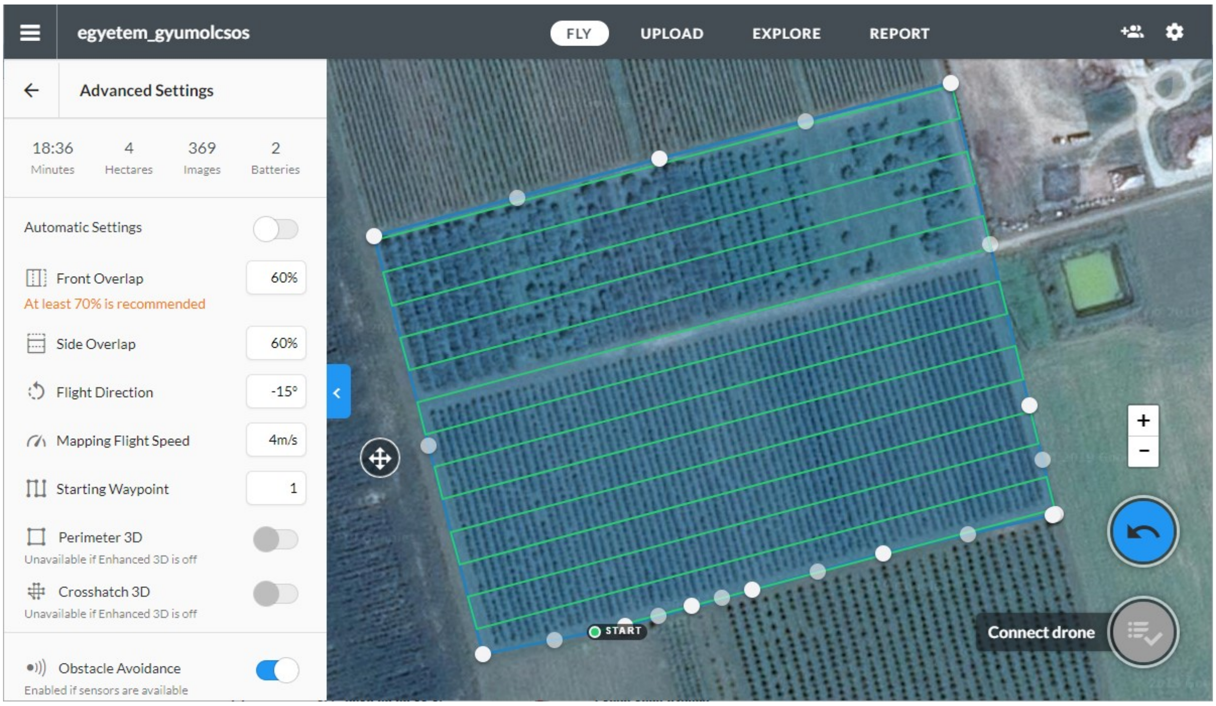


Figure 2. Flight path planned in DroneDeploy

The drone's shooting positions can be read from the captured images, so the actual flight path is known. From the recorded route data and camera characteristics, the area photographed from each drone position, the footprint, and then the coverage map of the entire flight can be calculated.

Based on this map, it can be evaluated whether the orthophoto condition is met, i.e. whether the adjacent images overlap to the required extent. To calculate the overlapping area, you first need to determine the coordinates of the intersections (E, F), and then calculate the area of the resulting overlapping rectangle ($AEGF$). If the area is known, the overlap rate can be calculated (Figure 3).

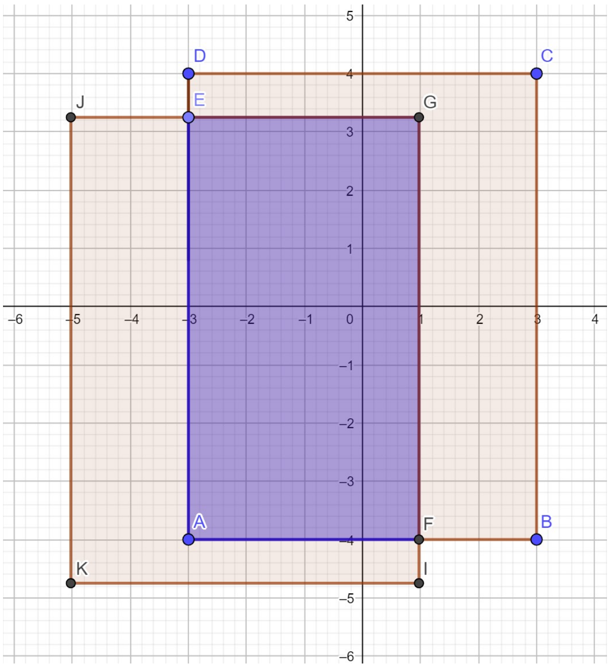


Figure 3. Footprint overlap calculation using intersection points

1.2. Goals

This calculation above can also take a long time, so we wanted to develop a fast computation algorithm that can be used to determine on the spot, after a flight, whether the captured images meet the required overlaps of the flight plan. If the result is negative, we gain the ability to retake the shots from the problematic areas immediately.

In this developed simpler method we estimate the degree of overlap based on the coordinates of the centers of the rectangles and the direction of the vector that connects them.

1.3. Disregarded factors

The effect of the relief on the footprint size was ignored during the study. We did not take into account the dimensional change due to the camera's tilt angle. During flights, the gimbal can hold the camera so well in the desired direction that the deviation from the vertical was only 0.1 degrees so far, which has a negligible effect. In flight, the rectangles of the footprints are not always parallel, but the difference is so small that we ignored it. When setting after turning, the deviation was only 1 degree and during scanning only 0.1 degrees.

2. Material and methods

GeoGebra software was used for the geometric analysis, and the calculations were made using Excel VBA programming language. We used QGIS software for mapping. We used two drones for the study, first a DJI Phantom 4 (with a modified filter on its red channel to be able to sense NIR). Main specifications are 1/2.3" CMOS, with effective pixels of 12.4 M; field of view of 84°, 20 mm (35 mm format equivalent) with f/2.8 focus at infinity; picture size of 4000×3000 pixels. The camera has a 3-axis gimbal that is capable of pitch (-90° to +30°), roll and yaw. The other drone used was a DJI Phantom 4 Pro, which has more infrared sensors (for collision avoidance) and a modified camera: 1'' CMOS with effective pixels of 20M; FOV of 84°, 8.8 mm/24 mm (35 mm format equivalent) at f/2.8, with the other main parameters being the same as the first drone mentioned.

2.1. Modeling in GeoGebra

Why did we use Geogebra for the simulation? On the one hand, Geogebra is freely accessible to anyone on every continent and is widely distributed. Geogebra.org has millions of free materials, simulations, exercises, lessons, games, and maths and other disciplines that everyone can freely browse. Another advantage of the Geogebra program is its dynamism. By changing the parameters we can observe the changes of the results. This feature of the program is heavily utilized by us. Finally, Geogebra is exceptionally applicable to finding loci (Blažek & Pech 2019, Botana & Abánades 2014).

The math problem is as follows. Given are two intersecting rectangles with parallel sides. One (*ABCD*) is fixed with center of origin, the other one is moving (*GIKJ*). Examine the set of points of the moving rectangle centers (the geometry of the centers), while the area of the resulting overlapping rectangle is constant. The sides of the original rectangles should be $2a$, $2b$, their area obviously being $A = 4ab$, the area of the overlapping rectangle should be percentage of the area of the original rectangles (in GeoGebra application $a = 3$, $b = 4$ units, $p = 60\%$). So the area of the overlapping rectangle is $4abp$.

First, E moves on the DM section. The coordinate of M is calculated from the consideration that E only makes sense to move on a given side until the sides of the fixed and moving rectangles parallel to the x axis overlap. So $(a - x(M) \cdot 2b = 4abp$ gives the coordinates of the point M: $M\ (a - 2ap, b)$. While E is the point of the segment DM, the vertex F of the overlapping rectangle is on the BC side of the fixed rectangle. Coordinates are also calculated from the $4abp$ area of the overlapping rectangle, which must give a constant value, so F: $F(a, b - 4abp / (a - x(E)))$. The set of H points can be defined by the GeoGebra GeometryLocation (<Point to Draw Geometric Location>, <Point>) command, where the first point in the command is H and the second is E.

If you also want to get the equation of the set described by H, you can do the following. The coordinates of $x(H) = x(E) + a$, $y(H) = y(F) + b = 2b - 4abp / (a - x(E))$, H will now be briefly denoted by x and y. Express the x-coordinate $x(E)$ of both points from each of the above equations so that we get a relation between the two coordinates of the H-point, which is sorted to give the equation of the curve described by the H-point.

$$y = 2b - \frac{4abp}{2a - x}$$

This is the equation of a hyperbola, or more precisely, a piece of the hyperbola on the condition of $0 \leq x \leq 2a - 2ap$ (Figure 4).

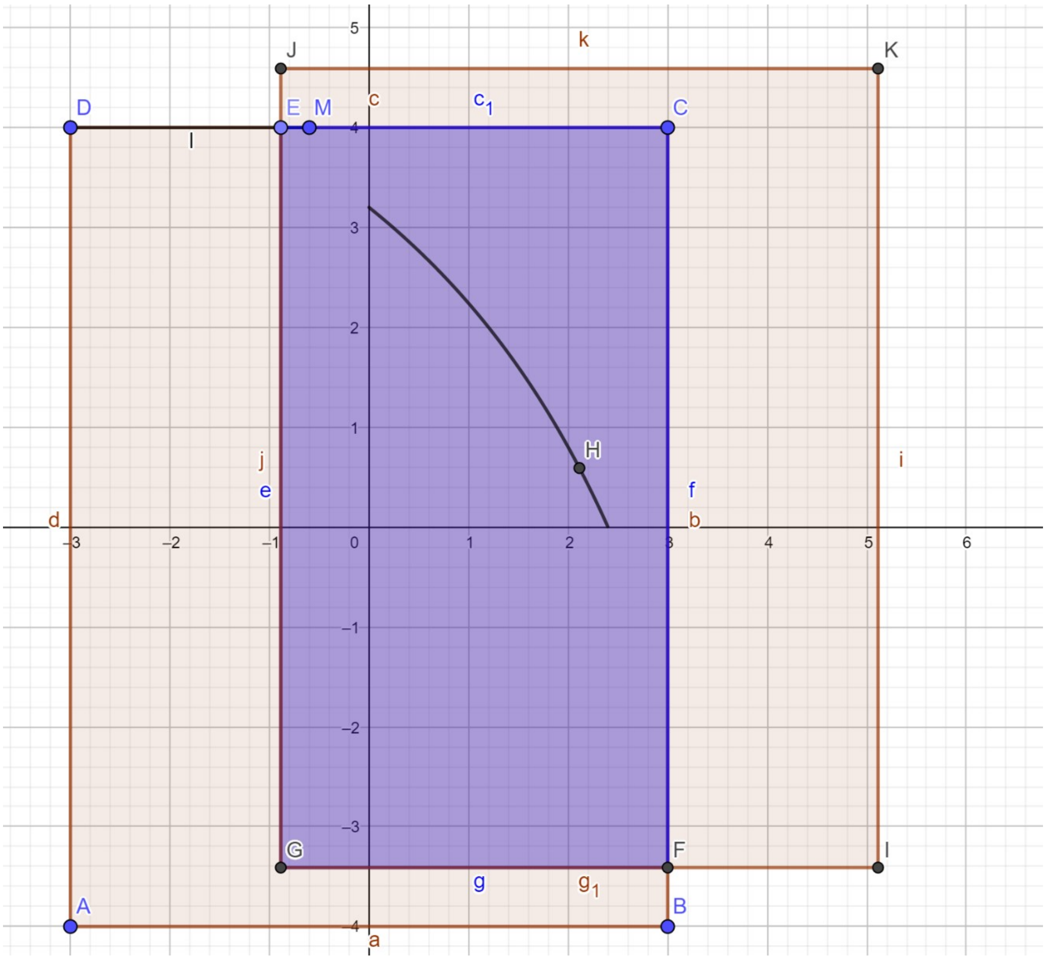


Figure 4. The hyperbola segment described by the center of the overlapping rectangle with the correct

In Figure 5, it follows from a similar idea that if E moves on the DL segment, the center of the moving rectangle H describes the equation $y = 4abp / 2a + x) - 2b$ under the condition $2ap - 2a \leq x \leq 0$.

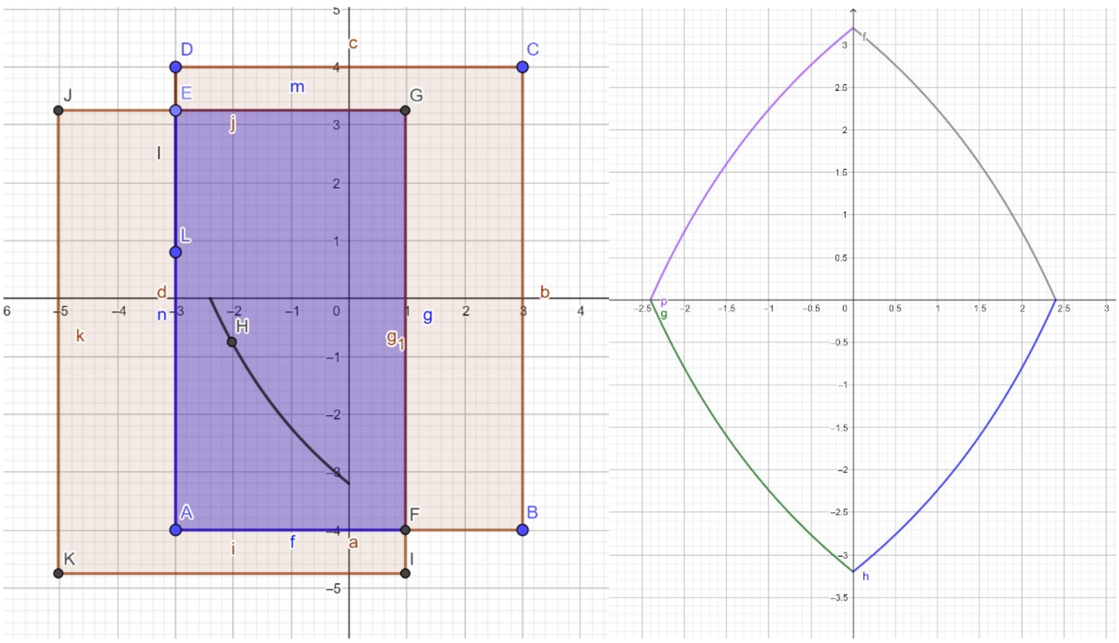


Figure 5. Hyperbola sections in the four plane quarters

For symmetry reasons, the curves shown in Figure 3 are obtained (twice each) from the point E on the other sides of the fixed rectangle to the geometric location in question.

The points of overlap required for drone flight are in the colored area in Figure 6.

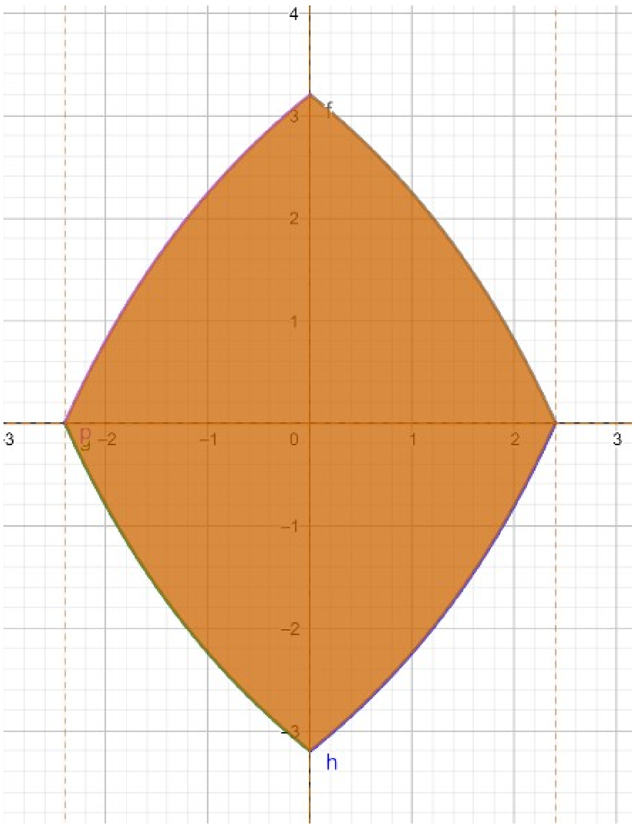


Figure 6. The area within the overlap is sufficient

The completed GeoGebra files can be consulted in the GeoGebra book on the GeoGebra website: https://www.geogebra.org/m/uksrwdw2.

2.2. The VBA program

Why did we choose the VBA programming language? Because VBA is built-in to Excel, it does not require any installation, is easy to use and provides full functionality of Excel. Because of this, it can be used to build and test the program, which, when completed, can be easily coded in a modern programming language.

Steps of the VBA program

First, we extract the EXIF data from the images into a CSV file using BR's EXIFextracter freeware software.

The image files contain EXIF (Exchangeable Image File Format) data, which holds the technical information of the images. This data is diverse, such as date and time, shutter speed, aperture, ISO, image size, GPS coordinates, and more.

Additional XMP (Extensible Metadata Platform) ISO metadata stored in the image is extracted with the VBA program. XMP includes, for example, the absolute (ellipsoidal) and relative (surface) height of the camera, the speed of the drone, the angle of the gimbal Yaw, Pitch, and Roll, etc.

The next step is that the coordinates of center of the images are converted from the WGS84 ellipsoid coordinate system used by GPS to the meter-based EOV (Uniform National Projection) coordinates used in Hungary. The calculation is based on Timár and Molnár (2002). Further calculations are also done with EOV coordinates, since this allows us to calculate in the usual Cartesian coordinate system.

To calculate the size of the photographed area, it is necessary to know the camera field of view (FOV) and the aspect ratio (aspect ratio) of the image, which can be obtained from the specification. The aspect ratio can also be calculated from EXIF data based on the image size. If the camera can capture multiple aspect ratio images, this is the better solution.

Calculating the footprint

The principle of footprint calculation is illustrated in Figure 7, and the calculation is described in detail by S. Srinivasan (2019). From the known FOV data of the drone camera and the altitude, the diameter of the footprint (DC section) can be calculated from which the known side lengths (CE and DE sections) can be calculated. Because of the center symmetry, both DE' and CE' are obtained.

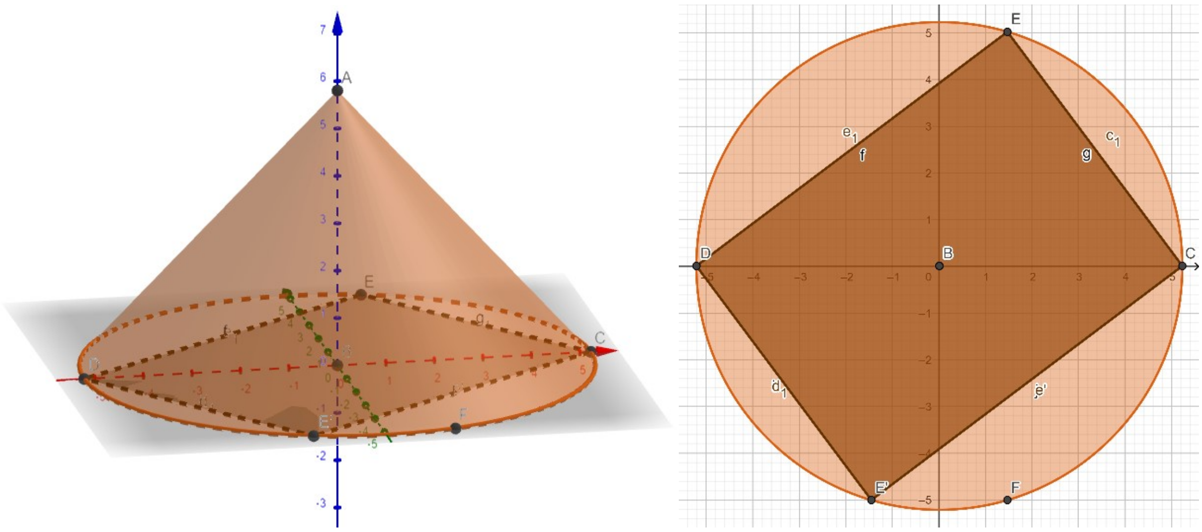


Figure 7. Principle of footprint calculation

VBA code snippet for footprint calculation based on the above:

```
Footprint_Diameter = 2 * altitude * Tan(FOV / 2 * PI / 180)
Footprint_width = Footprint_Diameter / ((1 + Picture_ratio ^ 2) ^ (1 / 2))
```

```
Footprint_height = Footprint_width * Picture_ratio
```

The rectangle (footprint) seen by the camera is rotated according to gimbal data and the coordinates of the corner points are saved.

The overlap calculation is then performed, the principle of which is shown in Figure 8. We select the rectangles in order and calculate the distance between the centers of the other rectangles (*O*, *O1*) with the Pythagorean theorem relative to the selected rectangle. The test is continued with rectangles having a center distance less than the shorter side length (Footprint_height).

Next, we calculate the distance (*OP*) between the center of the selected rectangle and the point P on the hyperbolane. The angle between the center-line segment ($\alpha + \beta$) and the shorter side of the rectangle (α) can be determined by knowing the coordinates of the corresponding points, and then calculating the angle β, which gives the equation of the line through *O* and *P*. The solution of the system of equations of the line and the hyperbola gives the coordinates of the P point. Then, using the Pythagoras theorem, we can calculate the distance of the point P of the hyperbola from the center *O* of the rectangle. If this is greater than the distance between the centers (*OP* > *OO1*), then the two rectangles overlap sufficiently.

The program counts how many other rectangles are covered to the extent needed.

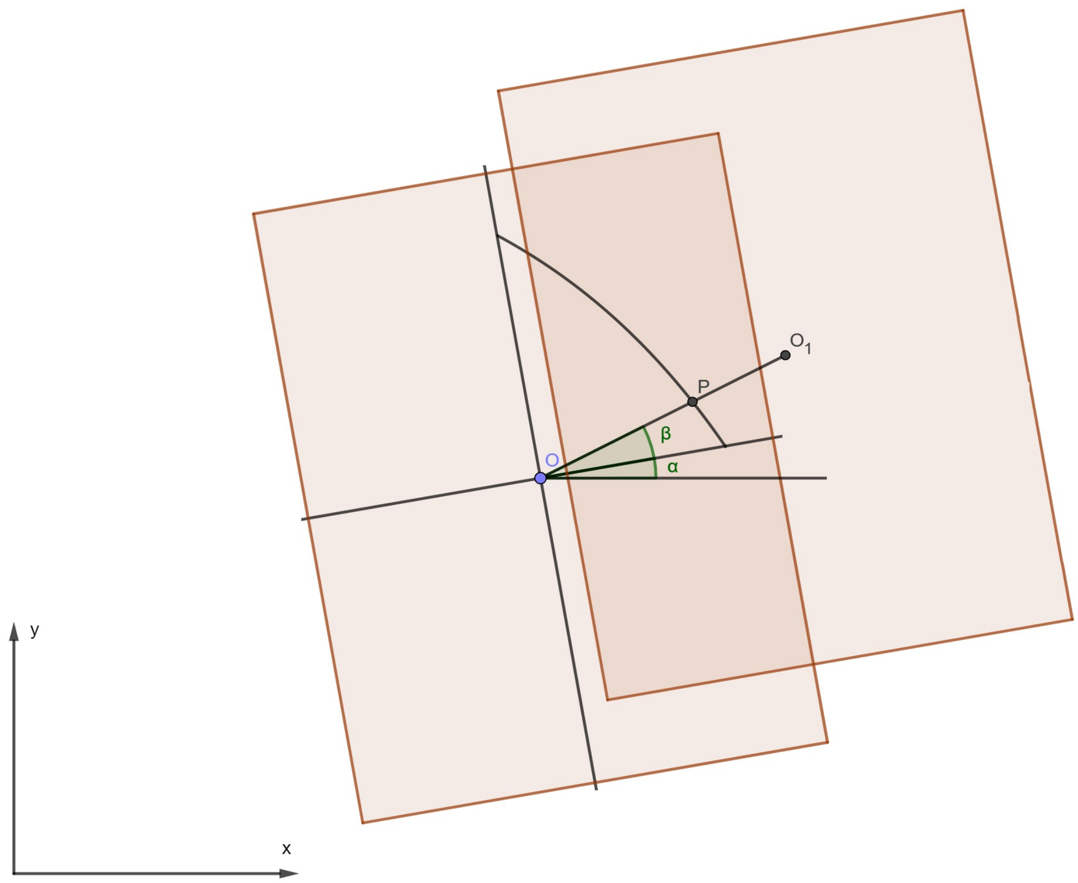


Figure 8. Scheme of overlap calculation

From the available data, the coordinates of the images, the lines connecting them, and the rectangles of the recorded areas are saved in a CSV file and plotted using QGIS.

The runtime of the VBA program depends on the number of recordings, but it produces results in hundreds of images in minutes.

3. Results

Table 1 contains the characteristics of the two flights which are illustrated.

Table 1. Attributes of the two highlighted flights

	Flight 1	Flight 2
Camera	DJI FC330	DJI FC6310
Picture size	4000x3000	4864x3648
Aspect ratio	0.75	0.75
Field of view (FOV)	84	84
Images taken	526	146
First image	DJI_0394.JPG	DJI_0240.JPG
Date	2019.03.20	2019.04.04
Time	10:51:04	11:27:39
Footprint width (m)	28.81	15.13
Footprint height (m)	21.61	11.35
Pixel size (m)	0.0072	0.0031
Front overlap	60%	60%
Side overlap	60%	60%
Relative Altitude (m)	20	10.5

First, we show the layers produced in QGIS from an adequately photographed area (Figure 9). The GPS coordinates of the recordings are indicated by circles and numbered by the flight track. The GPS points are linked by arrows to show the time course of the flight. Footprints were categorized according to the number of adjacent footprints that have a front overlap of at least 60%. As you might expect, there is the least amount of overlap at the edges, but there are at least two 60% overlapping neighbors everywhere. There is also a little overlapping footprint in the middle of Track 1, where for some unknown reason there was less footage. The large number of overlaps occurred because the drone once again flew the previously recorded area after the required battery replacement.

Even in the case of lateral overlap, the edges have the smallest matching neighbors and the most in the double-photographed area (Figure 10).

Images from the area were made into a proper orthomosaic and Digital Elevation Model (Figure 11).

Figure 12 shows the footprints of manual drone control during windy conditions. The GPS points and arrows that connect them indicate the disturbances in drone path. Because of the inappropriate condition, there are several footprints that have only two neighbors and two that have just one overlapping neighbor. The problem is that one of these is not on the edge of the track.

The orthomosaic from the images did not work as expected, and a hole was created due to lack of overlap (Figure 13).

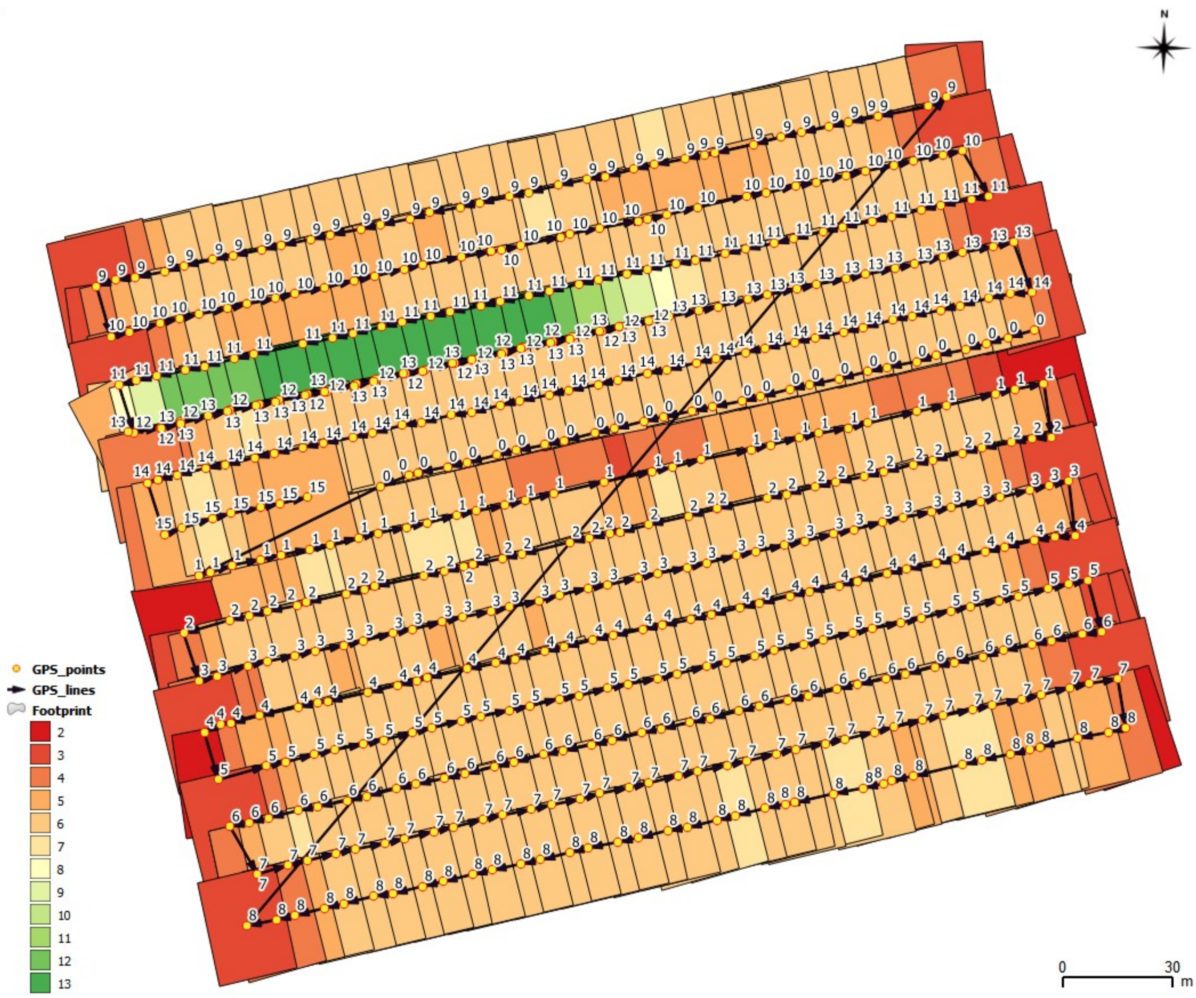


Figure 9. GPS points and lines and footprints categorized by front overlap on an adequately photographed area

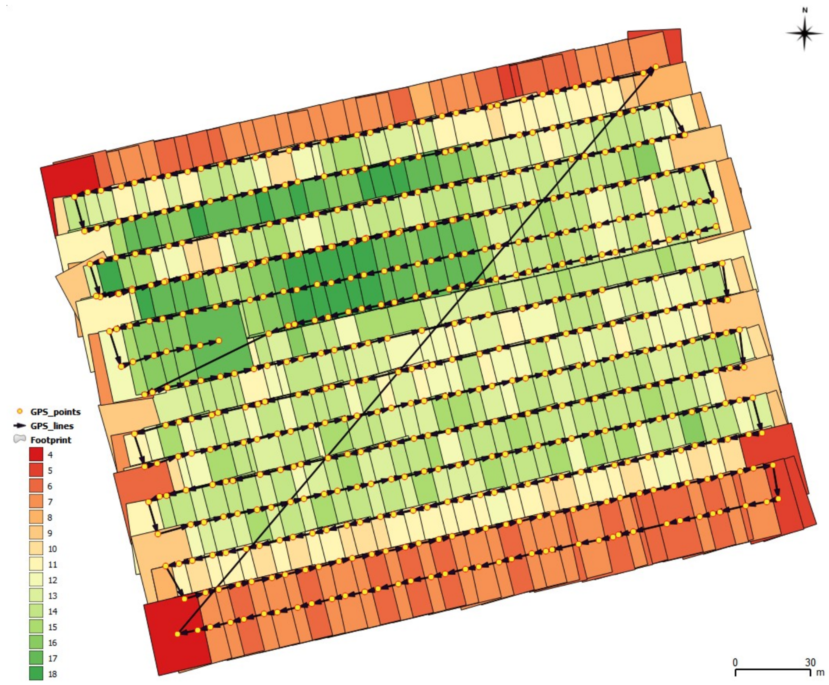


Figure 10. GPS points and lines and footprints categorized by side overlap on an adequately photographed area

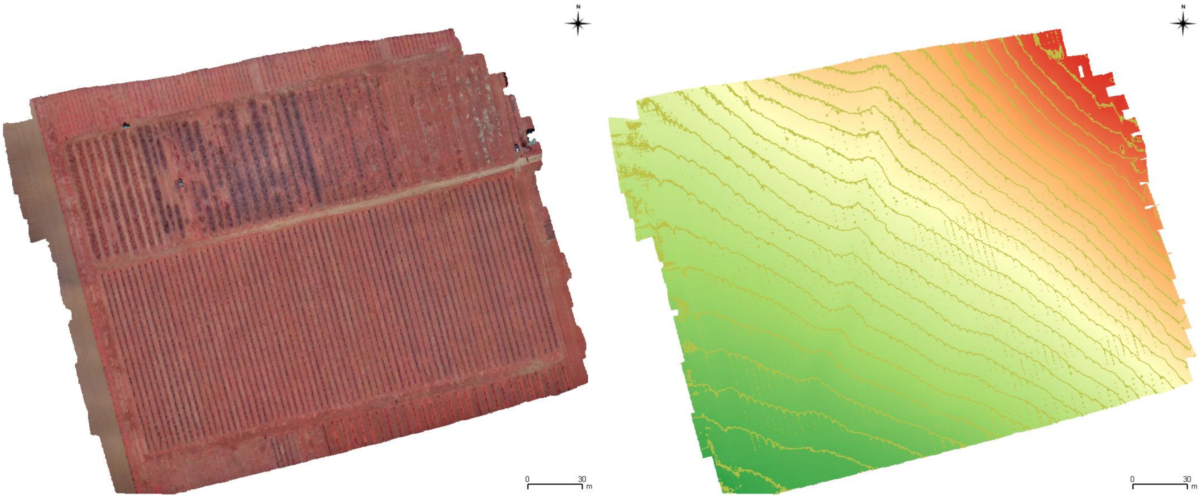


Figure 11. Orthomosaic and Digital Elevation Model from the Flight 1

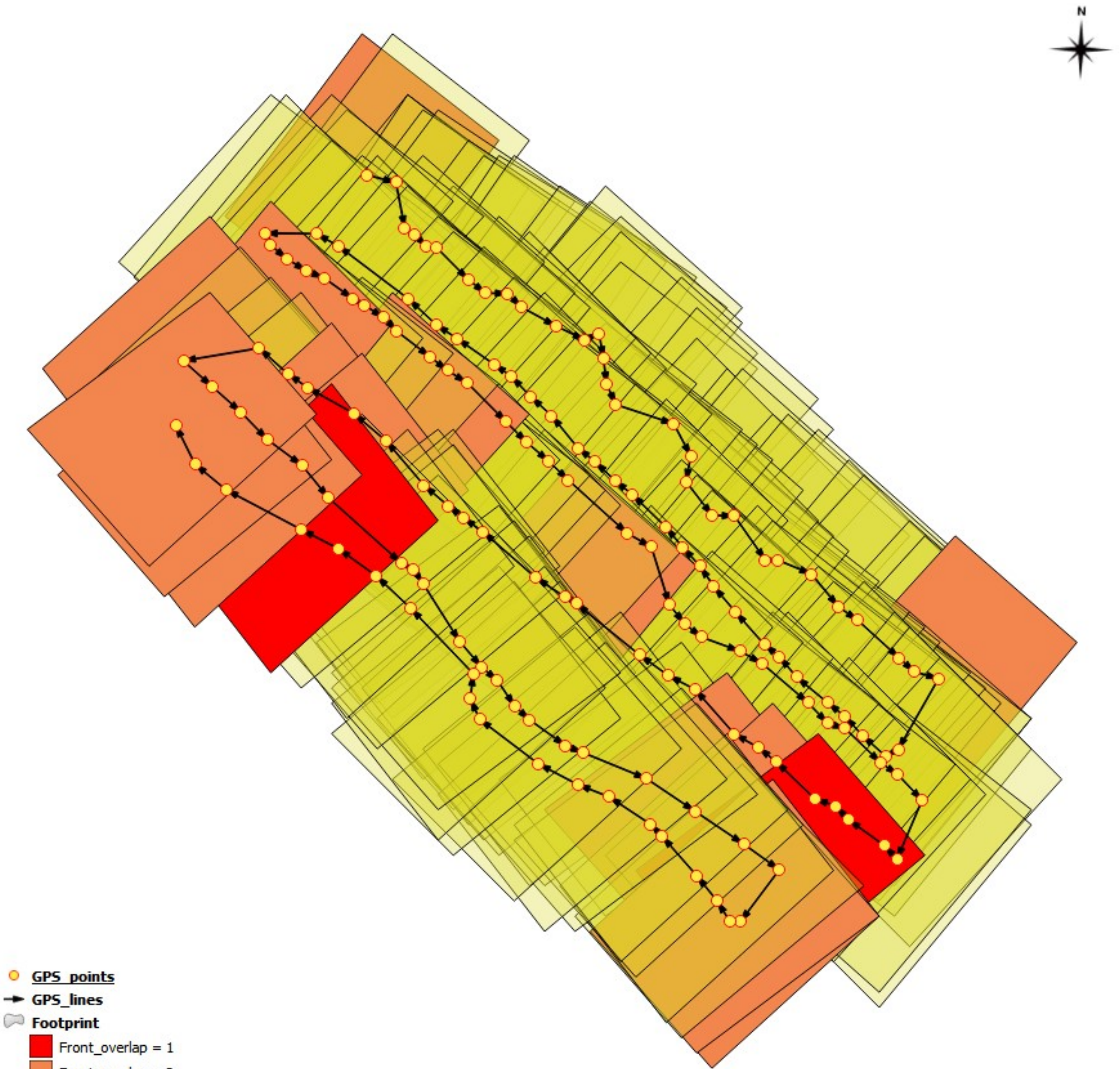


Figure 12. Footprints highlights the inadequate front overlap

Figure 13. Orthomosaic from the Flight 2

4. Conclusions

The presented method is suitable for the examination of aerial flight recordings, even on the spot, thus allowing the re-flight of areas with poor coverage. Developed software delivers results within minutes, so it is worth considering creating a user-friendly application written in a modern programming language. Installing the application on a mobile device (smart phone, tablet) makes it easy to check for the necessary overlaps in the field.

The developed method should be checked to see if the results of the traditional geometrical methods with the usage of intersecting points (Figure 3) are different from the results outlined here. It is also worth comparing the runtime of the two versions in case of different numbers of image files.

Acknowledgements

This work is supported by the EFOP-3.6.3-VEKOP-16-2017-00008 project. The project is co-financed by the European Union and the European Social Fund.

References

Detrekői, Á. & Szabó, Gy. 2013 'Térinformatika – Elmélet és alkalmazások' Typotex Kiadó ISBN: 978-963-4930-16-7. 292. p.

Blažek, J. & Pech, P. 2019 'Locus Computation in Dynamic Geometry Environment' *Mathematics in Computer Science*, vol 13. no. 1-2. pp. 31-40. https://doi.org/10.1007/s11786-018-0355-3

Botana, F. & Abánades, M.A. 2014 'Automatic deduction in (dynamic) geometry: Loci computation' *Computational Geometry*, vol 47. no. 1. pp.75-89. doi: 10.1016/j.comgeo.2013.07.001

DJI.com 2016 'Flight Control' (Online) https://developer.dji.com/mobile-sdk/documentation/introduction/flightController_concepts.html. (Accessed: 17 June, 2019)

Eisenbeiss, H., Stempfhuber, W. & Kolb, M. 2009 'Genauigkeitsanalyse der 3D-Trajektorie von Mini-UAVs' Zukunft mit Tradition "29. Wissenschaftlich-Technische Jahrestagung der DGPF", Eds. Seyfert, E. & Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation (DGPF) e.V., Potsdam, pp. 407-417.

Gómez-Candón, D. De Castro, A.I. & López-Granados F. 2014 'Assessing the accuracy of mosaics from unmanned aerial vehicle (UAV) imagery for precision agriculture purposes in wheat' *Precision Agric,* vol 15 pp. 44–56. doi: 10.1007/s11119-013-9335-4

Laliberte, A.S. & Rango, A. 2011. 'Image processing and classification procedures for analysis of subdecimeter imagery acquired with an unmanned aircraft over arid rangelands' *GIScience & Remote Sensing*, vol. 48. no. 1. pp. 4–23. doi: 10.2747/1548-1603.48.1.4

Merotto, A., Bredemeier, C., Vidal, R.A., Goulart, I.C.G.R., Bortoli, E.D. & Anderson, N.L. 2012 'Reflectance indices as a diagnostic tool for weed control performed by multipurpose equipment in precision agriculture' *Planta Daninha*, vol. 30. no. 2. pp. 437–447. doi: 10.1590/S0100-83582012000200024

Mulla, D.J. 2013 'Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps' *Biosystems Engineering*, vol. 114. no. 4. pp. 358–371. doi: 10.1016/j.biosystemseng.2012.08.009

Rasmussen, J., Ntakos, G., Nielsen, J., Svensgaard, J., Poulsen, R.N. & Christensen, S. 2016 'Are vegetation indices derived from consumer-grade cameras mounted on UAVs sufficiently reliable for assessing experimental plots?' *European Journal of Agronomy*, vol. 74. pp. 75–92. doi: 10.1016/j.eja.2015.11.026

Shanahan, J.F., Kitchen, N.R., Raun, W.R. & Schepers, J.S. 2008. 'Responsive in-season nitrogen management for cereals' *Computers and Electronics in Agriculture*, vol. 61. no. 1. pp. 51–62. doi: 10.1016/j.compag.2007.06.006

Srinivasan S. 2019 'Calculating a drone camera's image footprint OR How I learned to love tan again' (Online) https://www.techforwildlife.com/blog/2019/1/29/calculating-a-drone-cameras-image-footprint (Accessed: 10 June, 2019)

Timár G. & Molnár G. 2002 'A HD72 -> ETRS89 transzformáció szabványosítási problémái' *Geodézia és Kartográfia,* vol. 54. no. 12. pp. 28-30. ISSN 0016-7118

Vericat, D., Brasington, J., Wheaton, J., & Cowie, M. 2009 'Accuracy assessment of aerial photographs acquired using lighter-than-air blimps: Low-cost tools for mapping river corridors' *River Research and Applications,* vol. 25. no. 8. pp. 985-1000. doi: 10.1002/rra.1198

Wendel, J., Meister, O., Schlaile, C. & Trommer, G.F. 2006 'An integrated GPS/MEMS-IMU navigation system for an autonomous helicopter' *Aerospace Science and Technology*, vol. 10. no. 6. pp. 527-533. doi: 10.1016/j.ast.2006.04.002

Zhang, C. & Kovacs, J.M. 2012 'The application of small unmanned aerial systems for precision agriculture: A review' *Precision Agriculture* vol. 13. pp. 693–712. doi: 10.1007/s11119-012-9274-5

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 10, No. 2
journal.magisz.org

Beyond NDVI-Spectral indexing of biomass

János Tamás[1], Attila Nagy[2], Bernadett Gálya Farkasné[3]

INFO
Received 08-01-2020
Accepted 28-01-2020
Available on-line 28-01-2020
Responsible Editor: M. Herdon

Keywords:
vegetation index, remote sensing.precision agriculture

A B S T R A C T

The use of both terrestrial and satellite sensors often lacks the specific knowledge that makes them more effective. There are many reasons for this, one of them is agricultural digitalization to facilitate data availability and shortened automatic spectral indexing, but interpretation requires special remote sensing and agricultural knowledge. The absence of this complex knowledge, will result in misleading or incorrect solutions and on the other hand the existing technical possibilities cannot be exploited by the farmers. The application and calculation of the wide and narrow band channel VIs were presented on two agricultural test areas. The purpose of this publication is to briefly introduce the technical capabilities of vegetation indices based on a Hungarian test site and thereby reduce potential misinterpretation.

1. Introduction

Crop monitoring is an equally important task and tool for all crop growers. Remote sensing-based crop monitoring is a universal tool for all the players of the agro-business (farmers, traders, insurers) as they can find the field health monitoring, climate impact analysis, fertility management and crop yield modeling tools at a one-for-all Platform. The satellite or near-field spectral vegetation index is most often used in this process.

There are more than two hundred other spectral indices besides NDVI (Normalized Difference Vegetation Index), that are widely used to analyze vegetation pattern and generally agro-environmental conditions (Nagy et al., 2018).

Every index is basically a special spectral combination of the sensor-measured band reflectance properties (chlorophyll content, water content, pigment, etc.) at 2 or more wavelengths that reveals particular characteristics of vegetation.

The optimal applicability of vegetation indices highly depends on their developing methodologies. Most indices have been developed for Earth observation, so they have been tested and validated on the sensors of a given satellite (e.g. Landsat, SPOT, Sentinel family). The selection of satellite sensor validation areas was fundamentally dependent on the interests of the project participating countries (Haboudanea et al., 2004). Therefore, the small countries in Central Europe have never been included in the Carpathian Basin area in initial sensor testing programs. However, in recent decades, huge amounts of data have been collected and several successful projects have been completed to investigate the agricultural applicability of remote sensing data in this Central European region, but the original development opportunities of space technologies, were always missed due to lack of financial resources. The first Hungarian aerial hyperspectral imaging programme took place within the framework of the HYSENS project in 2002 (Kardeván et al., 2003). Water quality and vegetation were

[1] János Tamás
University of Debrecen
e-mail address: tamas@agr.unideb.hu
[2] Attila Nagy
University of Debrecen
e-mail address: attilanagy@agr.unideb.hu
[3] Bernadett Gálya Farkasné
University of Debrecen
e-mail address: bernadett.galya@agr.unideb.hu

analysed based on early results of the flight campaign (Jung, Kardeván & Gläßer 2007). From 2007, hyperspectral images were taken by an AISA DUAL (Eagle and Hawk) airborne hyperspectral camera system installed and operated in joint venture by Mechanization Institute of Agricultural Ministry and the University of Debrecen. The Eagle camera takes images in the visible and near-infrared range (400- 970 nm), while Hawk operates in the middle infrared range (970-2500 nm) with 498 spectral channels. Airborne hyperspectral imagery provides the potential for more accurate and detailed information extraction than it is possible with any other types of broadband, remotely sensed data. Precision agriculture requires spatially correct image data to control different field technologies. In this case, an important point was to develop the effective flight campaign to produce highly accurate high ground and spectral resolution data. Hyperspectral narrow-band indices are more sophisticated measures of general quantity than the traditional satellite broadband indices. Many of these indices are currently unknown in agricultural practice or under-used. In our study, the potential of AISA EAGLE airborne hyperspectral sensor data to create a narrowband vegetation indices distribution map of agricultural fields was evaluated.

Beyond the general aims, the purpose of this publication is to briefly introduce the technical capabilities of vegetation indices based on a Hungarian test site and thereby reduce potential misinterpretation.

2. Material and methods

2.1. Data acquisition and processing

Hyperspectral VI test site was a winter wheat site, close to Siófok city (UL Geo 18º 0’ 0.26” E; 46º 54’ 3.14” N); (spectral interval 400-970 nm, average bandwidth 2.9 nm, spatial resolution: 0.5 m). The total area was 20 ha (Figure 1). The hyperspectral images were taken by an AISA EAGLE airborne hyperspectral imaging spectrometer. AISA is a dual sensor system, which provides seamless hyperspectral data in the full range from 400 to 970 nm (Burai, & Tamás 2004).

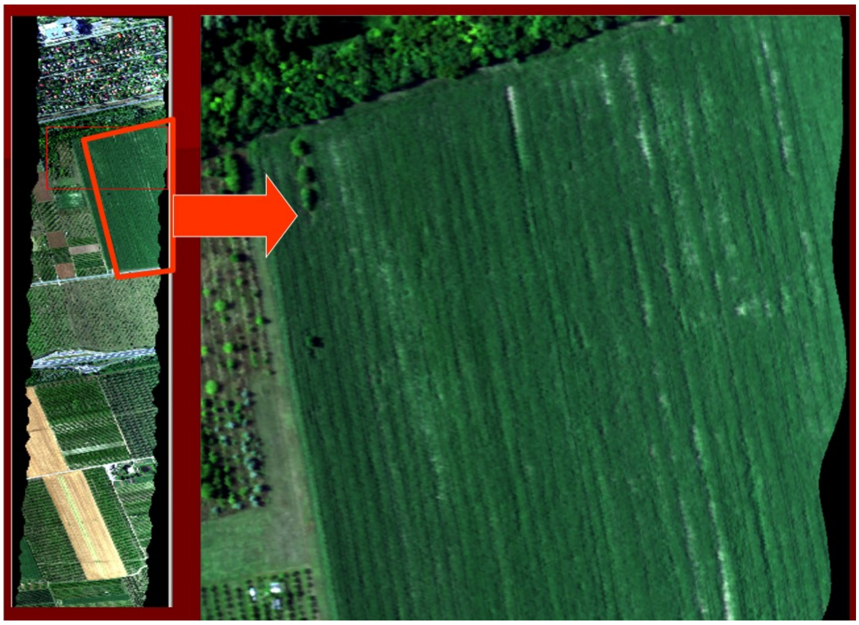


Figure 1. The test site was a winter wheat field near Siófok

The schematic steps of the hyperspectral image processing were the following: 1) Aerial and land image taking. 2) Radiometrical and geometrical corrections. 3) Noise filtering and data decrease. 4) Choosing the objective spectrum. 5) Classification. 6) Interpretation. 7) Checking (Burai & Tamás, 2004). Steps 1 and 2 were made with the CaliGeo (radiometric and geometric corrections), while for steps 3 to 6, ENVI 5.3 raster based remote sensing software and ArcGIS 10.5 GIS environment were applied. Below, steps 2, 5 and 6 as the most crucial ones of the whole process will be reported.

2.2. Vegetation indices

Widely used Vegetation Indices (VIs) are combinations of surface reflectance at two or more wavelengths designed to highlight a particular property of vegetation. Earth observation satellites use wide spectral channels so the vegetation indices use these channel combinations. Hyperspectral narrowband indices are more sophisticated measures of general quantity than the traditional satellite broadband indices. In this study, calculation and application of broad band VIs and narrow band VIs were evaluated. Indices under mentioned were calculated:

NDVI- Normalized Difference Vegetation Index, SAVI- Soil Adjusted Vegetation Index, MSAVI- Modified Soil Adjusted Vegetation Index, EVI- Enhanced Vegetation Index, GCI- Green Chlorophyll Index, SIPI- Structure Insensitive Pigment Index, NBR - Normalized Burn Ratio and ΔNBR-Delta Normalized Burn Ratio, mSR 705- Modified Red Edge Simple Ratio, VOG-1,2- Vogelmann Red Edge Index 1, 2, PRI- Photochemical Reflectance Index, CRI1,2- Carotenoid Reflectance Index, ARI1,2- Anthocyanin Reflectance Index 1 and WBI-Water Band Index. The formulas of calculated vegetation indices are summarized in Table 1.

Table 1. The formulas of applied spectral vegetation indices

Broad band VIs	**Narrow band VIs**
NDVI= (NIR-RED)/(NIR+RED)	$mSR_{705} = \frac{\rho_{750} - \rho_{445}}{\rho_{7050} + \rho_{445}}$
SAVI = $((\lambda_{NIR} - \lambda_{Red}) / (\lambda_{NIR} + \lambda_{Red} + L)) \times (1 + L)$	$mNDVI_{705} = \frac{\rho_{750} - \rho_{705}}{\rho_{750} - \rho_{705} - 2\rho_{445}}$
$MSAVI = \frac{1}{2}\left[2\lambda_{800} + 1 - \sqrt{(2\lambda_{800} + 1)^2} - \sqrt{8(\lambda_{800} - \lambda_{670})}\right]$	$VOG1 = \frac{\rho_{740}}{\rho_{720}}$ $VOG2 = \frac{\rho_{734} - \rho_{747}}{\rho_{715} + \rho_{726}}$
ARVI = $(\lambda_{NIR} - (2 * \lambda_{Red}) + \lambda_{Blue}) / (\lambda_{NIR} + (2 * \lambda_{Red}) + \lambda_{Blue})$	$PRI = \frac{\rho_{531} - \rho_{570}}{\rho_{531} + \rho_{570}}$
EVI = $2.5 * ((\lambda_{NIR} - \lambda_{Red}) / ((\lambda_{NIR}) + (C1 * \lambda_{Red}) - (C2 * \lambda_{Blue}) + L)$	$CR1 = \left(\frac{1}{\rho_{510}}\right) - \left(\frac{1}{\rho_{550}}\right)$
GCI = $(\lambda_{NIR}) / (\lambda_{Green}) - 1$	$CR2 = \left(\frac{1}{\rho_{510}}\right) - \left(\frac{1}{\rho_{700}}\right)$
SIPI = (λNIR – λBlue) / (λNIR – λRed)	$ARI1 = \left[\left(\frac{1}{\rho_{550}}\right) - \left(\frac{1}{\rho_{700}}\right)\right]$
NBR = $(\lambda_{NIR} - \lambda_{SWIR}) / (\lambda_{NIR} + \lambda_{SWIR})$	$ARI2 = \rho_{800}\left[\left(\frac{1}{\rho_{550}}\right) - \left(\frac{1}{\rho_{700}}\right)\right]$
ΔNBR= PrefireNBR-PostfireNBR	$WBI = \frac{\rho_{900}}{\rho_{970}}$

3. Results

3.1. Application of broad band VIs

Nowadays the use of the broadband NDVI index is one of the most widespread vegetation index out of broadband VIs. During the last decades, substantial efforts were taken in improving the Normalized Difference Vegetation Index (NDVI) to reduce saturation impacts when dense canopy is closed and in developing new indices aiming to compensate for soil background influences (Figure 2).

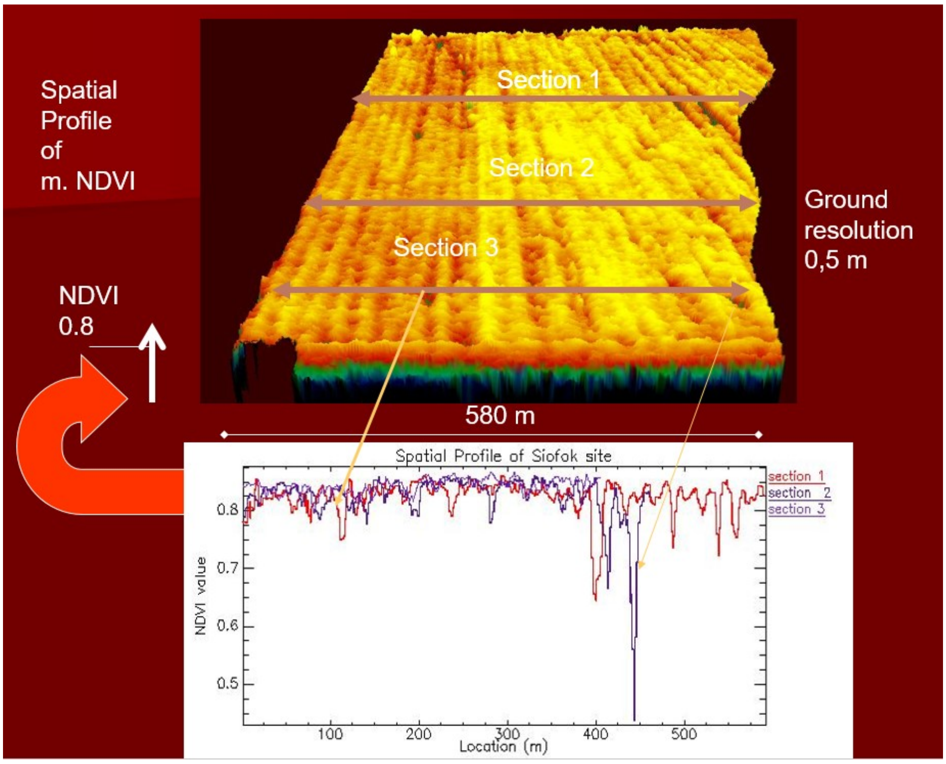


Figure 2. Cross section of spectral and physical space of winter wheat

The Soil Adjusted Vegetation Index (SAVI) was designed to minimize soil brightness influences. Huete (1988) added a soil adjustment factor L to the equation of NDVI in order to correct for soil noise effects (soil color, soil moisture, soil variability across region, etc.), which tend to impact the results. L values range between -1 and 1, depending on the amount of green vegetation present in the area. To run the remote sensing analysis of areas with high green vegetation, L is set to zero (in which case SAVI index data will be equal to NDVI); whereas low green vegetation regions require L=1. to use: for analysis of young crops; for arid regions with sparse vegetation (less than 15% of total area) and exposed soil surfaces (Figure 3). Huete (1988) suggested an optimal value of L= 0.5 to account for first-order soil background variations.

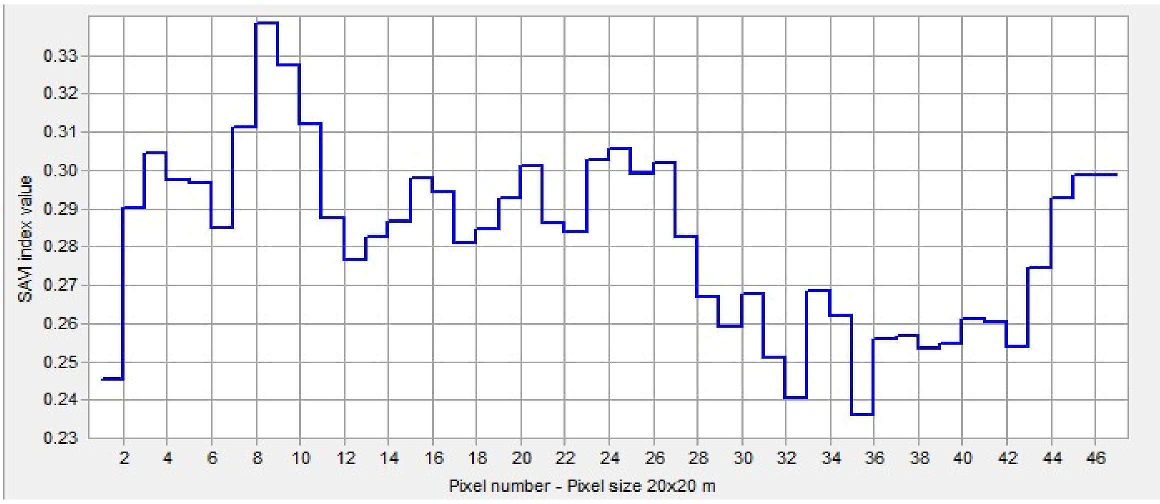


Figure 3. Cross section of Soil Adjusted Vegetation Index Index of winter wheat

Attempting to improve SAVI with regard to the differences in soil background, (Qi, et al. 1994) developed an improved SAVI (MSAVI) with a self-adjustment factor L that does not appear in the formulation of MSAVI.

The Atmospherically Resistant Vegetation Index (ARVI) is the first vegetation index, which is relatively prone to atmospheric factors (such as aerosol). The formula of ARVI index invented by Kaufman & Tanre (1992) is basically NDVI corrected for atmospheric scattering effects in the red reflectance spectrum by using the measurements in blue wavelengths. Compared to other indices, ARVI agriculture index is also more robust to topographic effects, which makes it a highly effective monitoring tool for tropical mountainous regions often polluted by soot coming from slash-and-burn agriculture. An atmospherically resistant vegetation index (ARVI) is proposed and developed for remote sensing of vegetation from the Earth Observing System (EOS) MODIS sensor. The same index can be used for remote sensing from Landsat TM and the EOS-HIRIS sensor. The index takes advantage of the presence of the blue channel (0.47.+or-0.01 μm) in the MODIS sensor, in addition to the red (0.66+or-0.025 μm) and the near-IR (0.865+or-0.02 μm) channels that compose the present normalized difference vegetation index (NDVI) (Kaufman & Tanre 1992). The data source was used to evaluate regions with high content of atmospheric aerosol (e.g. rain, fog, dust, smoke, air pollution).

The Enhanced Vegetation Index (EVI) was invented by Liu & Huete (1995) to simultaneously correct NDVI results for atmospheric influences and soil background signals, especially in areas of dense canopy. The value range for EVI is -1 to 1, and for healthy vegetation it varies between 0.2 and 0.8. EVI contains coefficient C1 and C2 to correct for aerosol scattering present in the atmosphere, and L to adjust for soil and canopy background. Traditionally, for NASA's MODIS sensor (which EVI index was developed for) C1=6, C2=7.5, and L=1 (Figure 4). When to use: for analyzing areas of agricultural region with large amounts of chlorophyll (such as virgin forests), and preferably with minimum topographic effects (not mountainous regions).

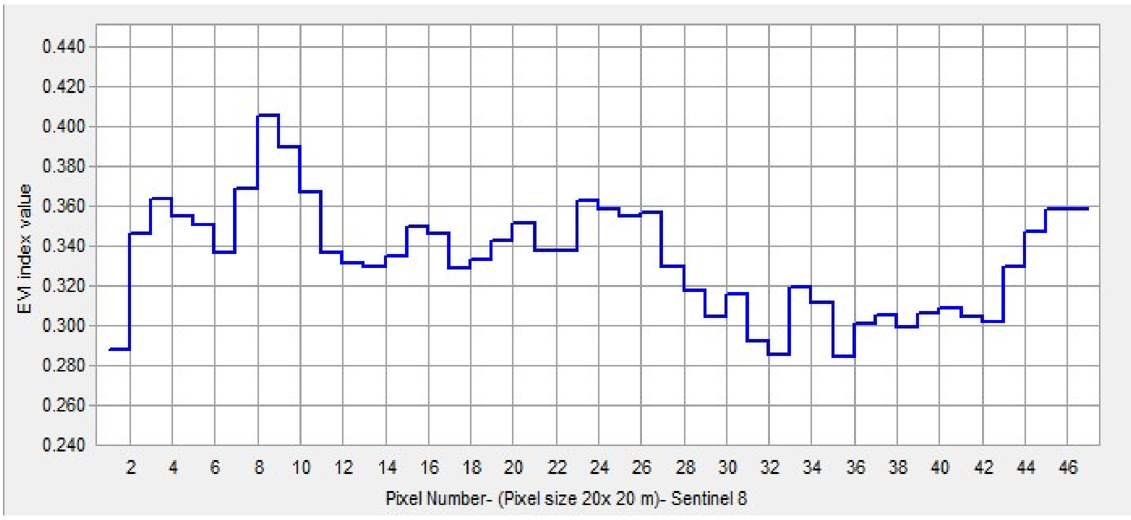


Figure 4. Cross section of Enhanced Vegetation Index of winter wheat

In remote sensing, the Green Chlorophyll Index (GCI) is used to estimate the content of leaf chlorophyll in various species of plants. The chlorophyll content reflects the physiological state of vegetation; it decreases in stressed plants and can therefore be used as a measurement of plant health.

Better prediction of chlorophyll amount with the GCI vegetation index can be achieved by using satellite sensors that have broad NIR and green wavelengths. The main fields of applications are monitoring the impact of seasonality, environmental stresses,and applied pesticides on plant health (Xue & Su, 2017).

The Structure Insensitive Pigment Index (SIPI) is good for analysis of vegetation with the variable canopy structure. It estimates the ratio of carotenoids to chlorophyll: the increased value signals of stressed vegetation.

Most researchers used high SIPI values (increased carotenoids and decreased chlorophyll) which are often applied as an indicator of plant disease, which is associated with loss of chlorophyll in plants.

SIPI can be used for monitoring plant health in regions with high variability in canopy structure or leaf area index, for early detection of plant disease or other causes of stress (Bannari, Morin, Bonn & Huete 1995).

By definition, it is the Normalized Burn Ratio (NBR) that is used to highlight burned areas following fire. The equation of NBR vegetation index includes measurements at both NIR and SWIR wavelengths: healthy vegetation shows high reflectance in NIR spectrum, whereas the recently burned areas of vegetation highly reflect in the SWIR spectrum. NBR fire index has become especially important in the past years as extreme weather conditions (regional drought) cause significant increase in wildfires destroying forest biomass (Escuin, Navarro & Fernández, 2008).

To perform NBR vegetation index calculation, one needs a raster image containing the near infrared and shortwave infrared bands that may be a satellite image collected by Landsat 7, Landsat 8, MODIS, etc. The range of values is between 1 and -1.

It is a common practice to assess burn extent and severity with the relative differenced NBR (delta Normalized Burn Ratio), which has shown the highest response to landscape changes caused by fire. It is a difference between the NBR calculated from an image of an area before the fire and NBR calculated from an image taken immediately after the burn. The range of values < -0.25 - > 0.66 from High post-fire regrowth to High-severity burn categories.

Additionally, there’s the NBR Thermal 1 index, which includes the Thermal band to enhance NBR and provide more accurate differentiation between the burned and unburned land. Although the typical use of NBR index for forestry is detection of active fires and analysis of burn severity, it can also be used for monitoring of agricultural vegetation survival after the burn.

3.2. Calculation and application of narrow band VIs

Narrowband greenness VIs more sophisticated measures of general and vigor of green biomass than the broadband VIs. Hyperspectral data or special LIDAR data give the opportunity to the utilization of vegetation indices based on narrow spectral channel data.

The Modified Red Edge Simple Ratio (mSR 705) index is a modification of the traditional broadband SR index. It differs from the standard SR because it uses bands in the red edge and incorporates a correction for leaf specular reflection. Applications include precision agriculture, forest monitoring, and vegetation stress detection. The value of this index ranges from 0 to 30. The common range for green vegetation is 2 to 8 (Datt 1999). The actual numbers of the winter wheat test site were: mean =11,14; max. = 18,7; min. = 1,55.

The Red Edge Normalized Difference Vegetation Index (NDVI705) is intended for use with very high spectral resolution reflectance data. Applications include precision agriculture, forest monitoring, and vegetation stress detection. This VI differs from the NDVI by using bands along the red edge, instead of the main absorption and reflectance peaks. The NDVI 705 capitalizes on the sensitivity of the vegetation red edge to small changes in canopy foliage content, gap fraction, and senescence. The value of this index ranges from -1 to 1. The common range for green vegetation is 0.2 to 0.9 (Sims & Gamon 2002). The actual numbers of the Winter wheat test site were: mean =0,83; max. = 0.87; min. = 0,17.

The Vogelmann Red Edge Index (VOG1 and VOG2) is a narrowband reflectance measurement that is sensitive to the combined effects of foliage chlorophyll concentration, canopy leaf area, and water content. Applications include vegetation phenology (growth) studies, precision agriculture, and vegetation productivity modeling (Vogelmann, Rock, & Moss, 1993). The value of this index ranges from 0 to 20. The common range for green vegetation is 4 to 8. The actual numbers of the winter wheat test site were: mean =2,59; max. = 2,84; min. = 1,33.

The Photochemical Reflectance Index (PRI) is a reflectance measurement that is sensitive to changes in carotenoid pigments (particularly xanthophyll pigments) in live foliage (Gamon, Penuelas, & Field, 1992). The actual numbers of the Winter wheat test site were: mean =0.016; max. = 0,07; min. = -0,05.

Carotenoid pigments are indicative of photosynthetic light use efficiency, or the rate of carbon dioxide uptake by foliage per unit energy absorbed. As such, it is used in studies of vegetation productivity and stress. Applications include vegetation health in forests, and agricultural crops prior to senescence. The value of this index ranges from -1 to 1. The common range for green vegetation is -0.2 to 0.2. Stress-related pigments include carotenoids and anthocyanins, which are present in higher concentrations in weakened vegetation. Carotenoids function in light absorption processes in plants, as well as in protecting plants from the harmful effects of high light conditions. The Carotenoid Reflectance Index 1-2 (CRI1-2) is a reflectance measurement that is sensitive to carotenoid pigments in plant foliage. Higher CRI1 values mean greater caratenoid concentration relative to chlorophyll (Gitelson, Zur, Chivkunova & Merzlyak, 2002). The common range for green vegetation is from 1 to 11. CR2 provides better results in areas of high carotenoid concentration. The actual numbers of the winter wheat test site were: mean =6,99; max. = 8.7; min. = 4.58.

Anthocyanins are water-soluble pigments abundant in newly forming leaves and leaves undergoing senescence. The Anthocyanin Reflectance Index 1 (ARI1) is a reflectance measurement that is sensitive to anthocynanins in plant foliage. Increases in ARI1 indicate canopy changes in foliage via new growth or death (Gitelson, Merzlyak & Chivkunova, 2001). The ARI2 is a modification of the ARI1 which detects higher concentrations of anthocynanins in vegetation. The value of these indices ranges from 0 to more than 0.2. The common range for green vegetation is 0.001 to 0.1. The actual numbers of the winter wheat test site were: mean =0.00012; max. = 0,0009; min. = 0,000086.

Water Band Index (WBI) is defined by the following equation (Champagne, Pattey, Bannari & Stratchan, 2001). Water content is an important quantity of vegetation because higher water content indicates healthier vegetation that is likely to grow faster and be more fire-resistant. The Water Band Index (WBI) is a reflectance measurement that is sensitive to changes in canopy water status. As the water content of vegetation canopies increases, the strength of the absorption around 970 nm increases relative to that of 900 nm. The common range for green vegetation is 0.8 to 1.2. Applications include canopy stress analysis, productivity prediction and modeling, fire hazard condition analysis, cropland management, and studies of ecosystem physiology. Actual numbers of winter wheat test site were: mean =1,032; max. = 1,129; min. = 0,87.

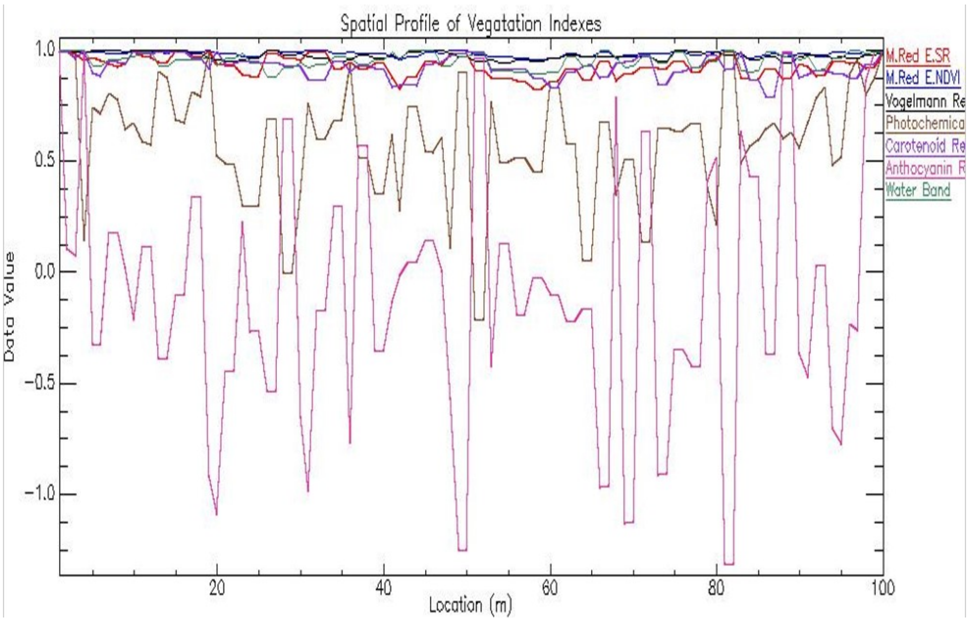


Figure 5. Cross sections of the applied narrow band VIs which was covered by winter wheat

Based on the indices there are several spots with different spectral feauter referring to the heterogeneity of the vegetation which can contribute to a better identification of management zones in precision agriculture.

3. Conclusion

The optimal applicability of vegetation indices highly depends on their developing methodologies. The Earth Observation satellites available today use wide spectral channels that scan large geographic regions. Although the number of satellite sensor channels and ground resolution are increasing, and satellite data are available in better temporal resolution, these properties remain below the spatial and temporal resolution of airborne hyperspectral imaging. However, in precision agricultural applications, the two survey modes are complementary in space and time. In order to compare the values obtained, it is necessary to know how to calculate them, the validation and the applicability of the used EM spectra. Some indices can be spectrally matched, such as NDVI and $NDVI_{705}$, while other indices have completely different purposes, contributing to a better identification of management zones in precision agriculture.

Acknowledgement

This research was partly supported by EFOP-3.6.2-16-2017-00001 project (Research of complex rural economic and sustainable development, elaboration of its service networks in the Carpathian basin). The research was partly financed by the Thematic Excellence Programme of the Ministry for Innovation and Technology in Hungary (ED_18-1-2019-0028), within the framework of the Space Sciences thematic programme of the University of Debrecen.

References

Bannari, A, Morin, D, Bonn, F, & Huete AR 1995 'A review of vegetation indices' Remote Sensing Reviews, 13:1-2, pp. 95-120. https://doi.org/10.1080/02757259509532298

Burai, P, & Tamás, J 2004 'Hyper- and multispectral remote sensing technologies in precision agricultural water management ' In: Proc. III. Alps-Adria Scientific Workshop, (Dubrovnik, Croatia) pp 54–57.

Champagne, C, Pattey, E, Bannari, A, & Stratchan, I.B 2001 'Mapping Crop Water Status: Issues of Scale in the Detection of Crop Water Stress Using Hyperspectral Indices' In: Proceedings of the 8th International Symposium on Physical Measurements and Signatures in Remote Sensing, Aussois, France. pp. 79-84.

Datt, B 1999, 'A New Reflectance Index for Remote Sensing of Chlorophyll Content in Higher Plants: Tests Using Eucalyptus Leaves' Journal of Plant Physiology 154, pp. 30-36. https://doi.org/10.1016/S0176-1617(99)80314-9

Escuin S, Navarro, R & Fernández P 2008 Fire severity assessment by using NBR (Normalized Burn Ratio) and NDVI (Normalized Difference Vegetation Index) derived from LANDSAT TM/ETM images, International Journal of Remote Sensing, 29:4, pp. 1053-1073, https://doi.org/10.1080/01431160701281072

Gamon, JA, Penuelas, J & Field, CB 1992 'A Narrow-Waveband Spectral Index That Tracks Diurnal Changes in Photosynthetic Efficiency' Remote Sensing of Environment 41.pp. 35-44. https://doi.org/10.1016/0034-4257(92)90059-S

Gitelson, AA, Merzlyak, MN & Chivkunova, OB 2001 'Optical Properties and Nondestructive Estimation of Anthocyanin Content in Plant Leaves' Photochemistry and Photobiology 71. pp. 38-45. https://doi.org/10.1562/0031-8655(2002)075<0272:ACCIPL>2.0.CO;2

Gitelson, AA, Zur, Y, Chivkunova, OB, & Merzlyak, MN 2002 'Assessing Carotenoid Content in Plant Leaves with Reflectance Spectroscopy' Photochemistry and Photobiology 75. pp. 272-281. https://doi.org/10.1562/0031-8655(2002)075<0272:ACCIPL>2.0.CO;2

Haboudanea, D, Millera JR, Patteyc, E, Zarco-Tejadad, PJ & Ian B. Strachane 2004 'Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture' Remote Sensing of Environment 90, pp. 337–352. https://doi.org/10.1016/j.rse.2003.12.013

Huete, AR 1988 'A soil vegetation adjusted index (SAVI)' Remote Sens. Environ., 25, pp295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Jung, A, Kardeván, P, & Gläßer, C 2007 'Hyperspectral technology in urban vegetation analysis' In: Proceedings 5th EARSeL Workshop on Imaging Spectroscopy. Bruges, Belgium, April 23-25. pp. 1-10.

Kardeván, P, Vekerdy, L, Róth, L, Sommer, S, Kemper, T, Jordán, Gy, Tamás, J, Pechman, I, Kovács, E, Hargitai, H, & László, F 2003 'Outline of scientific aims and data processing status of the first Hungarian hyperspectral data acquisition flight campaign' HYSENS 2002 HUNGARY. In: Proc. at the 3rd EARSEL Workshop on Imaging Spectroscopy, Herrsching, Germany, May 13–16) pp. 324–332.

Kaufman, KJ & Tanre, D 1992 'Atmospherically resistant vegetation index (ARVI) for EOS-MODIS' in IEEE Transactions on Geoscience and Remote Sensing, vol. 30, no. 2, pp. 261-270, https://doi.org/10.1109/36.134076

Liu, HQ & Huete, AR 1995 'A feedback based modification of the NDV I to minimize canopy background and atmospheric noise' IEEE Transactions on Geoscience and Remote Sensing, 33, pp. 457–465. https://doi.org/10.1109/TGRS.1995.8746027

Nagy, A, Fehér, J, Tamás, J 2018 'Wheat and maize yield forecasting for the Tisza river catchment using MODIS NDVI time series and reported crop statistics' Computers and Electronics in Agriculture 151 pp. 41-49. , 9 p. https://doi.org/10.1016/j.compag.2018.05.035

Qi, J, Chehbouni, A, Huete, AR, Keer, YH, & Sorooshian, S 1994 'A modified soil vegetation adjusted index' Remote Sens. Environ., 48, pp. 119–126. https://doi.org/10.1016/0034-4257(94)90134-1

Sims, DA & Gamon, JA 2002 'Relationships Between Leaf Pigment Content and Spectral Reflectance Across a Wide Range of Species, Leaf Structures and Developmental Stages' Remote Sensing of Environment 81. pp. 337-354. https://doi.org/10.1016/S0034-4257(02)00010-X

USGS 2017. https://landsat.usgs.gov/landsat-8

Vogelmann, JE, Rock, BN & Moss, DM 1993 'Red Edge Spectral Measurements from Sugar Maple Leaves' International Journal of Remote Sensing 14. pp. 1563-1575. https://doi.org/10.1080/01431169308953986

Xue, J, & Su, B, 2017 'Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications' Journal of Sensors, vol. Article ID 1353691, pp 17 pages, https://doi.org/10.1155/2017/1353691

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 10, No.2
journal.magisz.org

Computer vision in agriculture, application development using open source tools and systems

Mihály Tóth[1], Dániel Dér[2], Szilvia Borbásné Botos[3], Róbert Szilágyi[4]

INFO
Received 19-12-2019
Accepted 22-01-2020
Available on-line 28-01-2020
Responsible Editor: M. Herdon

Keywords:
computer vision, agriculture, artificial intelligence, neural network

ABSTRACT

Nowadays the optimization of agricultural production is a crucial task. The use of computer vision may serve in increasing efficiency as this technology has achieved significant results in many research and practical applications to date. In the following years, we will experience more and more use cases of the technology and its usability in the intensification of agricultural production to meet the demand of the growing population. Computer vision appears as a sub-domain, also in the new and popular concept of Industry 4.0, ensuring an integrated aspect of the technology. Our practical experiment was performed to examine the utility of the currently available open-source toolkits in commuter vision, utilizing OpenCV and Google TensorFlow libraries. In this experiment, the typical processes of computer vision were implemented using various algorithms for each step, including imaging, pre-processing, post-processing and finally, classification. For the experiment, pictures of apples have been used as training data, representing various conditions. The steps including processing, segmentation, and identification of the fruit, were presented. The most commonly used detection algorithms were tested to determine estimated size, shape, and texture properties. Using a convolutional neural network, the identification of the fruit was presented with a recognition accuracy greater than 93%.

1. Introduction

In the paper, the possibilities of open-source computer vision methods will be assessed, in agricultural production. Due to the changing nature of our environment as well as the increasing population, it has become a major challenge to reduce its environmental impact, increase food security and develop sustainable agricultural production (Agathokleous and Calabrese, 2019). Machine vision systems have already been developed to a level where it is possible to assess existing electronic systems, upgrade them, or replace them with more efficient ones, thus supporting clients by automating their routine tasks (Van der Stuyft et al., 1991). For the sensory analysis of agricultural and food products, automatic control systems based on cameras and IT solutions have recently been applied and tested for their effectiveness. The machine vision systems have been successful in objectively measuring various parameters of agricultural and food products, that influences a given aspect of the product. Machine vision involves capturing, processing, and analysing images, making it easier to evaluate the visual properties of a given object. Its potential has long been recognized in the food industry. Recent

[1] Mihály Tóth
University of Debrecen, Faculty of Economics and Business, Hungary
University of Debrecen, Károly Ihrig Doctoral School of Management and Business, Hungary
toth.mihaly@econ.unideb.hu
Mihály Tóth
[2] Dániel Dér
derdanis@gmail.com
[3] Szilvia Borbásné Botos
University of Debrecen, Faculty of Economics and Business, Hungary
botos.szilvia@econ.unideb.hu
[4] Róbert Szilágyi
University of Debrecen, Faculty of Economics and Business, Hungary
szilagyi.robert@econ.unideb.hu

developments in hardware and software have assisted at increasing the use of the system by providing low cost and efficient solutions, which have led to new studies that have contributed to the further development of machine vision. As a result, the methods of automated computer vision continue to grow significantly today in the food industry and other sectors as well due to cost-effectiveness, consistency, excellent speed, and accuracy (Brosnan and Sun, 2004).

1.1 Computer vision

Computer vision is part of the IoT (Internet of Things) concept, Cyber-Physical Systems (CPS) and, at the same time, artificial intelligence (AI), due to the fact that the former two is based on sensor-driven functions when the measured data (in this case, images in form of a matrix) are transmitted to computers for processing where artificial intelligence provides the framework for extracting an aggregated output, in form of a classification. After obtaining the data, the process is followed by a customized intervention that is performed according to the results.

Computer vision is the process by which a machine, usually a computer or embedded systems, automatically processes an image and sends a report of what is in the image considering an attribute (Snyder and Qi, 2017). The subject of machine vision is a crucial area of artificial intelligence. The term should not be understood solely as perceptual images (based on a colour imaging sensor) because it includes all sensors that provide spatially ordered visual information about the environment (IR camera etc.). Computer vision consists of two components, which can be divided further. The measurement must first be performed, thereafter the provided data by the measurement must be processed and interpreted. The complexity of this process varies, depending on the environment (real-life environment or laboratory, eliminating noises), the circumstances of the imaging (lighting parameter, including the intensity, temperature, and quality of the light) and finally, the expected results (determining the presence, quality parameters or damages). Based on the data, it is possible to recognize and follow objects, but also to draw new conclusions from indirect data. The digital image is produced by one or more image sensing sensors, which, in addition to the various imaging sensors, contain range sensors, tomography devices, radars or ultrasound cameras (Sinha, 2012).

The classification is one of the important aspects of computer vision. It can be based on K-Nearest Neighbour (KNN), Support Vector Machine (SVM) or Artificial Neural Network (ANN) (Naik and Patel, 2017), mentioning the well-known alternatives. KNN is one of the simpler methods, working in an unsupervised manner. On the basics, it assign data to the most represented category (Naik and Patel, 2017). as its name suggests, it represents each k number of clusters based on the weighted average of the data, included in the clusters (Chinchuluun et al., 2010). The number of clusters has to be chosen manually, resulting in varying result (Kim et al., 2012). SVM classification, on the other hand, uses the concept of decision hyperplanes. The linear hyperplanes are determined using the training dataset, which is being used to split the dataset. It is capable of linear two-class separation, but it can be extended to one or more class separation, non-linear separation, and non-linear regression problems as well. The support vectors are the samples, that are located close to the hyperplanes (Shastry et al., 2017). ANN is the most commonly used method nowadays, which is a network of interconnected nodes, associated with weights that determine the strength of a particular connection (Patrício and Rieder, 2018). We can distinguish between supervised, semi-supervised and unsupervised neural learning methods (Khan et al., 2018), but in this case, supervised methods are dominant. The structure of the network can be either convolutional (CNN) or recurrent (RNN) (Andreas Kamilaris and Prenafeta-Boldú, 2018) to mention the main structures, however as in computer vision we are working with matrices, convolutional neural networks are typical. A deep learning network is different in a sense, that they are built using multiple, interconnected layers. The disadvantage of ANN is that it needs larger dataset and proper labelling in order to obtain an applicable training set (A. Kamilaris and Prenafeta-Boldú, 2018).

1.2 Practical applications

Computer vision can be used throughout the food chain regardless of the process. Based on researches, categories, including weed identification, land cover classification, plant recognition, fruit counting and corn type classification were determined (Andreas Kamilaris and Prenafeta-Boldú, 2018), however, other use cases can be examined as well, including grading by colour or external quality parameters, ripeness inspection or blemish detection (Bhargava and Bansal, 2018).

During the production process the most frequent method is to segment the specific object considered the required resolution (an object, in case of fruits of animals or an aggregated view of a field in case of arable crop production). In the case of detection, apple detection was performed on the tree, based on decision tree, KNN and SVM (Marzoa Tanco et al., 2018). In an experiment, various fruits could be detected and classified using deep CNN (convolutional neural network) with accuracy more than 82% (Sa et al., 2016). In the case of disease detection, research performed an experiment, where diseases of papaya could be determined using SVM classification with more than 90% accuracy (Habib et al., 2018). Other than classification, we can see examples of prediction, capable of predicting the yield based on colour, representing nutrient content (Shidnal et al., 2019). During food processing, sorting is an important task for quality management. The detection of defects in apples was determined using C4.5 classification with an accuracy rate between 73%-93% based on a research (Sofu et al., 2016). The defects and ripeness of tomato can also be determined based on ANN, with 1005 and 96,47% accuracy accordingly (Arakeri and Lakshmana, 2016). An experiment about dried fig grading explains that classification, based on various indexes can be performed with accuracy between 90% and 100% (Baigvand et al., 2015). Other research describes that soybeans quality evaluation (including the identification of stems, pods and defects) could be performed with accuracy between 75% and 100% (Momin et al., 2017). An experiment with K-means clustering for segmentation and multi-class SVM for classification shows that apple diseases could be determined with more than 93% accuracy (Dubey and Jalal, 2013). It is also possible to build a 3D model to determine length and width, thickness, surface area and volume (Su et al., 2017).

Table 1. Accuracy of the methods

Description	**Processing**	**Classification**	**Accuracy (%)**	
Detection and classification of apple disease	GLCM algorithm	SVM	93	(Dinesh et al., 2017)
Mango recognition	KNN	SVM	92	(Kumari et al., 2019)
Grape leaf disease recognition	KNN	SVM	88,89	(Padol and Yadav, 2016)
Forest recognition	N.a.	SVM	87,9	(Pal, 2005)
Detection of olive infection	Wilks' lambda	PLSDA, SVM	80,85	(Beyaz et al., 2019)
Tree recognition	GLCM algorithm	MLP	72-60	(Tou et al., 2007)
Apple damage detection	I-RELIEF algorithm	RVM	95,63	(Zhang et al., 2015)

In traceability, computer vision and QR codes also serve an important role, because it is necessary to track data starting from raw products, through food processing, transport, warehousing, to retailing and reaching the end consumer (Tarjan et al., 2014). As a consumer, we can also encounter computer vision in food recognition (Kawano and Yanai, 2014) and even food recommendation with balanced nutrition in mind to maintain a healthy diet (Resende Silva and Cui, 2017).

1.3 Material and methods

The experiment has been performed based on open-source libraries, available for Python programming language. OpenCV is an open-source library for machine vision and machine learning. OpenCV is built to provide a common infrastructure for computer vision-based applications. OpenCV is an open-source library for computer vision and machine learning. OpenCV Python is a wrapper for the original C++ library that can be used with Python. OpenCV is built to provide a common infrastructure, with machine vision-based applications and the ability to increase the machine's sensing speed, as an important consideration. OpenCV supports some programming languages, namely Python, Java, C and C ++. It is true that Python-based programs are usually slower than their C++ alternatives, but they also take much less time to develop, since the codes used in Python can be three to five times shorter. It is open-source, unlike MATLAB, which specializes in data analysis, exploration, and visualization. Based on the mentioned advantages, it proved to be a better alternative in the case of a simple experiment.

The applied methods include pre-processing algorithms (exposure, contrast, gamma and white balance correction, scaling, sharpening, colour space conversion), as well as other methods, including Sobel filter, Gaussian blur, erosion and dilation. After pre-processing the training set using the mentioned methods, the following procedure was the segmentation. In this process, several methods were assessed, including adaptive thresholding, binary thresholding, K-means clustering, watershed algorithm and Canny edge detection. As an alternative solution, several feature extraction methods were assessed as well, including ORB detection (Oriented FAST and Rotated BRIEF), Harris corner detection and blob detection. Using the processed training set, size estimation (based on reference data) and classification were applied, using CNN (convolutional neural network), which can be applied for the whole area of the image as well as for extracted details.

2. Results and their evaluation

Figure 1. illustrates the practice test. The attributes and contact information of the captured images were loaded into a simple relational database. During the processing step, the obtained results were processed using various segmentation methods.

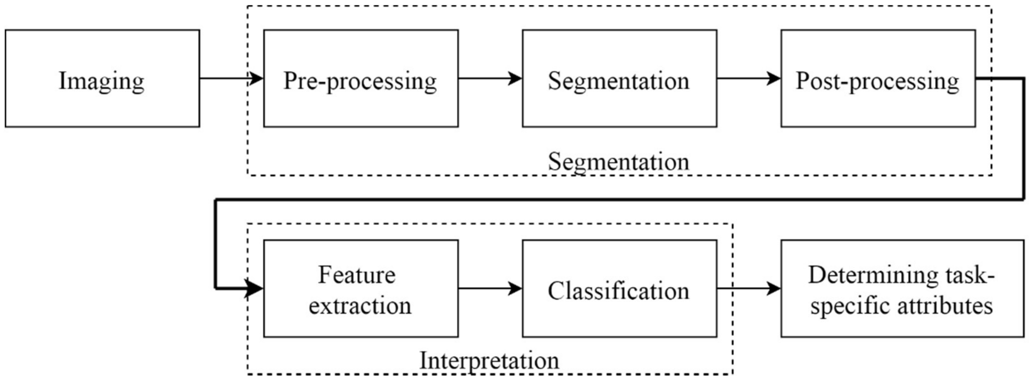


Figure 1. The applied Image Recognition Process (Own-work)

2.1 Imaging

The database is formed by 6 tables. In the database there were uploaded the metadata of 304 images, making it possible to load the file using ID and name, ensuring the usability in iterative functions. The pictures were taken using a mobile device and using a natural light source, pointed from above. The mobile has an aperture of f1.7, which greatly influences the image quality as it determines how much light the sensor can let in at given shutter speed. It also affects the depth of field, however, considering the size of the sensor, the effect is not noticeable. This value cannot be changed on the device being used, as it has a fixed aperture. The exposure and white balance are handled phone itself, which is a significant disadvantage since it results in inconsistent images. The ISO value was set to 125, high enough to work using the ambient light. The focal length is 4.20 mm, which is also a predefined value. The shutter speed is at 1/50s as average, which corresponds to daylight condition and indicates the time required for the exposure. The image metadata is stored in the database. The path to the image is stored here, which can be queried via code into a list, after which the application goes through the list. It also writes the changes made to the image into the database. Each modified image will be given a new ID and may have a unique name.

2.2 Preprocessing

First of all, we need to standardize the images, and then make image corrections to work easier with them. Input images have a resolution of 4032x3024 pixels and contain hundreds of images in the dataset. The pictures have been resized in order to reduce the unnecessary details (as a possible source of noise) and to determine a standard resolution for all images. During scaling, the aspect ratios were kept despite the smaller resolution. By default, the images have red, green, and blue (RGB) colour channels (additive colour model). In the input image, a gamma correction was performed to optimize the contrast values. The next step was to convert the images to grayscale (12 bit), since the process of image segmentation does not handle colours. The colour images are stored in a separate variable in order to perform operations on the grayscale image so that the processing methods can be performed on the colour image as well, taking the modification into account. At the same time, a blur filter was applied to reduce further details. Of the median blur and Gaussian blur, the latter was our choice after testing both methods, since Gaussian uses a technique that assumes that our image is two-dimensional, so it puts the most weight on the middle pixels. The results of these are illustrated in Figure 2.

Figure 2. RGB picture grayscale version and Gaussian Blur (Own-work)

After the Gaussian Blur was performed, there was still a large amount of unnecessary information in the picture. This was removed by contour recognition, which identified the earliest next white pixel on each side. The image was cut while the background of the apple was 40 pixels larger than the background. This reduced the image to a uniform 300x300 pixel. Pre-processing can be performed using the Sobel operation, which allows us to recognize the edges of the image both vertically and horizontally. A similar operation used for edge detection, including Gabor filtering. The Gabor filter is mathematically structured to work with different image shapes, sizes, and applied filters. For example, if an image has diagonal edges, the Gabor filter set will only give a strong answer if its direction is the same as the edges Prateekvjoshi (2014).

Figure 3. shows the result for the Sobel operation in the top row, dl/dx detects the edges of the image horizontally and dl/dy detects vertically. The second line shows the Gabor filter, where the middle kernel image scans the image at a 45-degree angle to get the most detail.

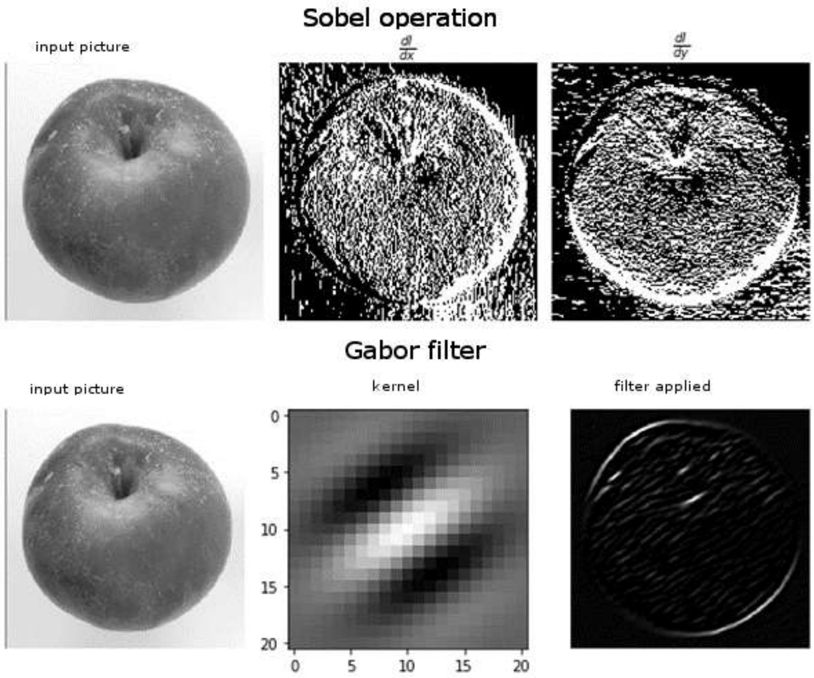


Figure 3. Sobel operation and Gabor filter comparison (Own-work)

The next image processing method, which already plays a role in the segmentation process, is thresholding. This will produce a grayscale image and convert it to a binary image, so it will have only 0 and 1 values. After defining the interval of the intensity range, running the code gives the result shown in Figure 4.

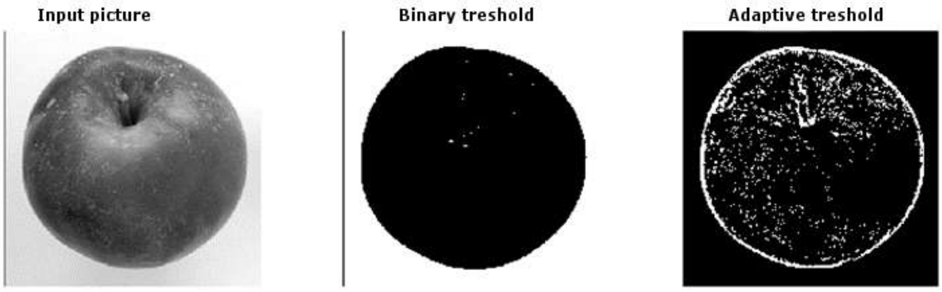


Figure 4. Binary and Adaptive treshold comparison (Own-work)

OpenCV provides various threshold methods. These methods fall into two groups: global, where each pixel is placed at the same threshold, and adaptive, where the threshold depends on the pixels, so the algorithm assigns its own threshold to each pixel.

2.3 Postprocessing

For the first test of the segmentation process, we used the k-means cluster algorithm. At times, the colour of pixels can help define semantically close areas. The algorithm only needs to know how many clusters there are in an image, in other words how many clusters we want our image to appear (Spizhevoy and Rybnikov, 2018). By defining this information, the algorithm will automatically find the best clusters shown in Figure 5.

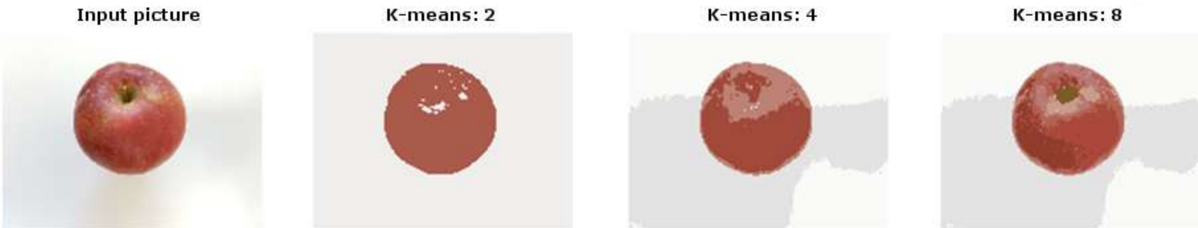


Figure 5. K-means algorithm examples based on different clusters (Own-work)

The figure shows more and more details appear in the images as we progressively increase the number of clusters. While in a two-cluster display the image is completely homogeneous, displaying a total of three colours, the eight-cluster, image is completely reproduced. This allows us to analyse the clusters separately and to classify the separately analysed parts. In segmentation processes, we often hear the expression Watershed. We can use the Watershed algorithm when we have initial segmented points and want to fill automatically the surrounding areas in the same segmentation class. These initial segmented points, called cores, need to be set manually, but in some cases it is possible to automate them (Spizhevoy and Rybnikov, 2018). Figure 6 shows the result.

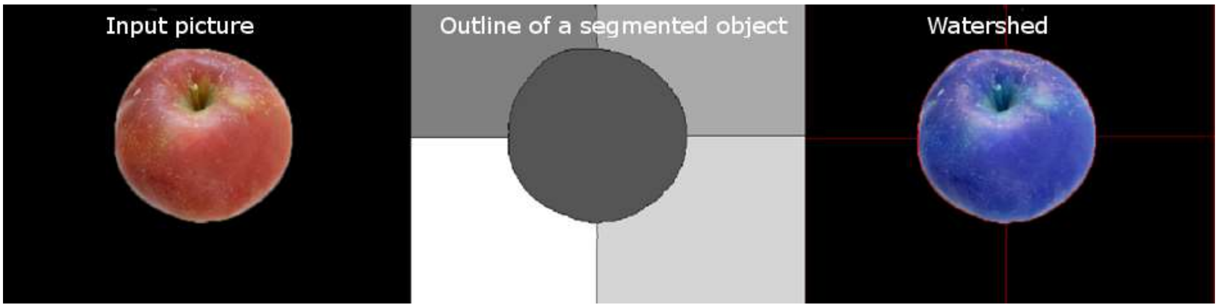


Figure 6. Watershed algorithm (Own-work)

The use of the Watershed algorithm became applicable after several attempts. Variable results were generated with input images with no background removed. The middle picture shows the outline of the segmented apple in 8-bit form, for which we used erosion and dilation, then it was given a specific threshold so that the fruit could be separated from the background. After the background has been successfully separated by the algorithm, it removes the modifications from the segmented object so that we finally get our input image highlighted by its contours (Joe Minichino, 2015).

One of the apples had a noticeable injury showing cuts at right angles to each other. This gave us the opportunity to use Harris's corner detection method. Chris Harris and Mike Stephens developed the algorithm in 1988, and let us not only detect corners, but also detect edges. The algorithm is capable of recognizing edges and corners and cannot be used for texture recognition. Figure 7. illustrates the input image and the result.

Figure 7. Harris's corner detection (Own-work)

This algorithm assists in detecting the corners of an image by scanning the image and detecting locations with the greatest deviation. Once the algorithm recognizes the corners in the image which can be highlighted as shown in the figure (Gollapudi, 2019). The process began by converting the image to grayscale, then it is possible to apply Harris edge detection, based on the built-in methods of the library.

2.4 Classification

One way to extract properties is to determine and recognize the size of the object. In the following test, we tried to determine the size of an apple based on a reference point. The result of this process is illustrated in Figure 8.

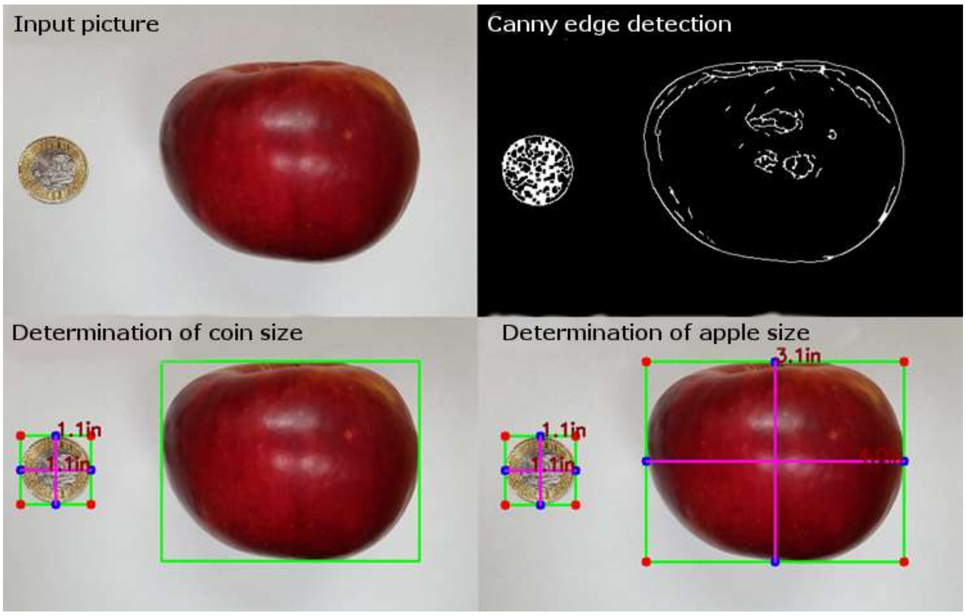


Figure 8. The process of apple size determination (Own-work)

The test requires two objects in one image. In this case, we used a 200 HUF coin with a diameter of 2.8cm or 1.1 inches. To use this method, we also need to know the resolution of the image, which is 600x450 pixels. After converting it to a grayscale image, the input image is subjected to a variety of operations such as erosion, dilation, and Canny edge detection. These play an important role in the mapping of contours, as further operations are based on them. Once the contours are created, you need to know the size of the coin and enter its value in order to determine the size of the apple in the image. Also, it is important that there must be enough space between the reference object and the other object (Rosebrock, 2016). The method has a various limitation, mainly due to the perspective way of imaging (we cannot be sure how far the object is to the reference point in Z-axis), which can be solved by utilizing an alpha channel (using a correspondent sensor and projector), describing the distance from the imaging sensor.

The next step was the classification, which should be done in multiple steps, beginning with the object itself (to determine if it’s really is an apple and, in that case, what kind of apple is that), which can be followed by classifying features, like the extracted damages. However, due to the requirement of the dataset, it was only possible to perform the former analysis. The classification of the apples was performed using a basic convolutional neural network. It was important in this experiment to rely on basic, predefined settings, thus the form of the activation function and the depth of the model is not considered. After 30 epochs, the result was 93.38% based on the test dataset.

3. Conclusions

The main purpose of this article was to present the details of the process and activities of computer vision, from image acquisition to processing through an application-specific presentation. As a development opportunity, a system concept was compiled. To make our machine vision operations more professional and environment-friendly, we recommend using a Raspberry Pi-based system.

The operating system would be a Debian based, standard Raspbian Linux distribution, with Python installed by default. The suggested packages that we need, such as Google Tensorflow and OpenCV, can be installed using commands specific to the console (using packet managers). In order to achieve correct imaging, diffuse illumination is required for homogeneous lighting. The advised method is to use light modifiers with a large surface area (softbox and tripod) to illuminate the object on all sides.

A further area of study could be the segmentation and classification of injuries and texture, which would also enable us to classify apples by quality. It is now possible to create a graphical interface using the PyQt graphical framework. PyQt is a cross-platform library for developing GUIs for Python-based applications. This allows us to create an entire development environment that contains the necessary libraries, packages, and frames. In practice, the user can configure the interface in a widget-like way.

Acknowledgments

The work/publication is supported by the EFOP-3.6.1-16-2016-00022 project. The project is co-financed by the European Union and the European Social Fund.

References

Agathokleous, E., Calabrese, E.J., 2019. Hormesis can enhance agricultural sustainability in a changing world. Glob. Food Sec. 20, 150–155. https://doi.org/10.1016/j.gfs.2019.02.005

Arakeri, M.P., Lakshmana, 2016. Computer Vision Based Fruit Grading System for Quality Evaluation of Tomato in Agriculture industry. Procedia Comput. Sci. 79, 426–433. https://doi.org/10.1016/j.procs.2016.03.055

Baigvand, M., Banakar, A., Minaei, S., Khodaei, J., Behroozi-Khazaei, N., 2015. Machine vision system for grading of dried figs. Comput. Electron. Agric. 119, 158–165. https://doi.org/10.1016/j.compag.2015.10.019

Beyaz, A., Martínez Gila, D.M., Gómez Ortega, J., Gámez García, J., 2019. Olive fly sting detection based on computer vision. Postharvest Biol. Technol. 150, 129–136. https://doi.org/10.1016/j.postharvbio.2019.01.003

Bhargava, A., Bansal, A., 2018. Fruits and vegetables quality evaluation using computer vision: A review. J. King Saud Univ. - Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2018.06.002

Brosnan, T., Sun, D.-W., 2004. Improving quality inspection of food products by computer vision––a review. J. Food Eng. 61, 3–16. https://doi.org/10.1016/S0260-8774(03)00183-3

Chinchuluun, A., Xanthopoulos, P., Tomaino, V., Pardalos, P.M., 2010. Data Mining Techniques in Agricultural and Environmental Sciences. Int. J. Agric. Environ. Inf. Syst. 1, 26–40. https://doi.org/10.4018/jaeis.2010101302

Dinesh, S., Ramya, G., Menaga, S., 2017. A Computer Vision Based Diseases Detection and Classification in Apple Fruits. Int. J. Eng. Res. V6, 161–165. https://doi.org/10.17577/IJERTV6IS100073

Dubey, S.R., Jalal, A.S., 2013. Adapted approach for fruit disease identification using images. Image Process. Concepts, Methodol. Tools, Appl. 3–3, 1395–1409. https://doi.org/10.4018/978-1-4666-3994-2.ch069

Gollapudi, S., 2019. Learn computer vision using OpenCV : with deep learning CNNs and RNNs.

Habib, M.T., Majumder, A., Jakaria, A.Z.M., Akter, M., Uddin, M.S., Ahmed, F., 2018. Machine vision based papaya disease recognition. J. King Saud Univ. - Comput. Inf. Sci. 0–9. https://doi.org/10.1016/j.jksuci.2018.06.006

Joe Minichino, J.H., 2015. Learning OpenCV 3 Computer Visionwith Python Second Edition.

Kamilaris, Andreas, Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. Comput. Electron. Agric. 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Kamilaris, A., Prenafeta-Boldú, F.X., 2018. A review of the use of convolutional neural networks in agriculture. J. Agric. Sci. 156, 312–322. https://doi.org/10.1017/S0021859618000436

Kawano, Y., Yanai, K., 2014. Food image recognition with deep convolutional features. UbiComp 2014 - Adjun. Proc. 2014 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. 589–593. https://doi.org/10.1145/2638728.2641339

Khan, S., Rahmani, H., Shah, S.A.A., Bennamoun, M., 2018. A Guide to Convolutional Neural Networks for Computer Vision, Synthesis Lectures on Computer Vision. https://doi.org/10.2200/S00822ED1V01Y201712COV015

Kim, J., Kim, B.-S., Savarese, S., 2012. Comparing Image Classification Methods: K-Nearest-Neighbor and Support-Vector-Machines. Appl. Math. Electr. Comput. Eng. 133–138.

Kumari, N., Bhatt, D.K., Dwivedi, R.K., Belwal, R., 2019. Performance Analysis of Support Vector Machine in Defective and Non Defective Mangoes Classification 1563–1572.

Marzoa Tanco, M., Tejera, G., Di Martino, M., 2018. Computer Vision based System for Apple Detection in Crops, in: Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS - Science and Technology Publications, pp. 239–249. https://doi.org/10.5220/0006535002390249

Momin, M.A., Yamamoto, K., Miyamoto, M., Kondo, N., Grift, T., 2017. Machine vision based soybean quality evaluation. Comput. Electron. Agric. 140, 452–460. https://doi.org/10.1016/j.compag.2017.06.023

Naik, S., Patel, B., 2017. Machine Vision based Fruit Classification and Grading - A Review. Int. J. Comput. Appl. 170, 22–34. https://doi.org/10.5120/ijca2017914937

Padol, P.B., Yadav, A.A., 2016. SVM classifier based grape leaf disease detection, in: Conference on Advances in Signal Processing, CASP 2016. pp. 175–179. https://doi.org/10.1109/CASP.2016.7746160

Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222. https://doi.org/10.1080/01431160412331269698

Patrício, D.I., Rieder, R., 2018. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. Comput. Electron. Agric. 153, 69–81. https://doi.org/10.1016/j.compag.2018.08.001

Prateekvjoshi, 2014. Understanding Gabor Filters [WWW Document].

Resende Silva, B.V., Cui, J., 2017. A Survey on Automated Food Monitoring and Dietary Management Systems. J. Heal. Med. Informatics 08. https://doi.org/10.4172/2157-7420.1000272

Rosebrock, A., 2016. Measuring size of objects in an image with OpenCV [WWW Document].

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. Deepfruits: A fruit detection system using deep neural networks. Sensors (Switzerland) 16. https://doi.org/10.3390/s16081222

Shastry, K.A., Sanjay, H.A., Deexith, G., 2017. Quadratic-radial-basis-function-kernel for classifying multi-class agricultural datasets with continuous attributes. Appl. Soft Comput. J. 58, 65–74. https://doi.org/10.1016/j.asoc.2017.04.049

Shidnal, S., Latte, M. V., Kapoor, A., 2019. Crop yield prediction: two-tiered machine learning model approach. Int. J. Inf. Technol. https://doi.org/10.1007/s41870-019-00375-x

Sinha, P.K., 2012. Image Acquisition and Preprocessing for Machine Vision Systems, Image Acquisition and Preprocessing for Machine Vision Systems. SPIE. https://doi.org/10.1117/3.858360

Snyder, W.E., Qi, H., 2017. Fundamentals of Computer Vision. Cambridge University Press. https://doi.org/10.1017/9781316882641

Sofu, M.M., Er, O., Kayacan, M.C., Cetişli, B., 2016. Design of an automatic apple sorting system using machine vision. Comput. Electron. Agric. 127, 395–405. https://doi.org/10.1016/j.compag.2016.06.030

Spizhevoy, A., Rybnikov, A., 2018. OpenCV3: Computer Vision with Python Cookbook, Packt Publishing. Packt Publishing.

Su, Q., Kondo, N., Li, M., Sun, H., Al Riza, D.F., 2017. Potato feature prediction based on machine vision and 3D model rebuilding. Comput. Electron. Agric. 137, 41–51. https://doi.org/10.1016/j.compag.2017.03.020

Tarjan, L., Šenk, I., Tegeltija, S., Stankovski, S., Ostojic, G., 2014. A readability analysis for QR code application in a traceability system. Comput. Electron. Agric. 109, 1–11. https://doi.org/10.1016/j.compag.2014.08.015

Tou, J.Y., Lau, P.Y., Tay, Y.H., 2007. Computer Vision-based Wood Recognition System. Proc. Int'l Work. Adv. Image Technol.

Van der Stuyft, E., Schofield, C.P., Randall, J.M., Wambacq, P., Goedseels, V., 1991. Development and application of computer vision systems for use in livestock production. Comput. Electron. Agric. 6, 243–265. https://doi.org/10.1016/0168-1699(91)90006-U

Zhang, B., Huang, W., Gong, L., Li, J., Zhao, C., Liu, C., Huang, D., 2015. Computer vision detection of defective apples using automatic lightness correction and weighted RVM classifier. J. Food Eng. 146, 143–151. https://doi.org/10.1016/j.jfoodeng.2014.08.024

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 10, No. 2
journal.magisz.org

Corporate Social Responsibility in the Agri-Food Sector: Evidence from Greece

Konstantinos Demestichas[1], Constantina Costopoulou[2], Maria Ntaliani[2], Konstantina Remoundou[1], Evgenia Adamopoulou[1], and Antonia Papagiannaki[2]

INFO
Received 20-10-2019
Accepted 15-11-2020
Available online 28-01-2020
Responsible Editor: M. Herdon:

Keywords:
Corporate Social Responsibility, Website, Agri-Food Sector.

ABSTRACT

Corporate Social Responsibility (CSR) can benefit business directly, such as by improving the working environment and increasing productivity, and indirectly, such as by increasing customers' interest and sales. The focus of this article lies on the investigation of CSR in the Greek agri-food companies through two perspectives, namely the business and the consumer perspective. The business perspective is based on the investigation of CSR activities promoted on the corporate websites of 222 food companies, SMEs and large sized. The consumer perspective is based on a questionnaire analysis of 200 participants, investigating their willingness to purchase products from companies practicing CSR activities. The results of this research showed that CSR activities are rather low in Greece and only few agri-food companies' have a clear and sound strategy to promote such actions. It seems that, during the economic crisis, most companies, especially SMEs, do not invest in such activities, but their main concern is survival, which -given the economic environment- is challenging. In addition, companies mainly support social activities, give less support to environmental activities, and even less support to human resource activities. Although most consumers seem to be socially aware of the CSR actions performed by the agri-food companies and have a sense of responsibility themselves, a significant part still is unaware of such actions.

1. Introduction

In recent years, Corporate Social Responsibility (CSR) has received a lot of attention in the business and scientific world, due to its significance both in economic and in societal terms. Studies on CSR primarily focus on whether businesses in various countries take up CSR-related initiatives and perform relevant activities (e.g. sponsorship, charity); which business sectors are more active in CSR; and what is the potential for CSR development (Porter & Kramer, 2006, Aguinis & Glavas, 2012, Carroll & Shabana, 2010, Lindgreen & Swaen, 2010).

The concept of CSR is dynamic and constantly evolving. Although several attempts have been made to define it, a widely accepted definition has not yet been established (Secchi, 2007). CSR was firstly defined by Johnson (1971), according to whom "in a socially responsible enterprise, senior management balances a combination of different interests". On the other hand, the European Commission (EC), through its renewed CSR strategy published in 2011, defines CSR as "the responsibility of businesses for their impact on society". However, ten years earlier, in 2001, the relevant Green Paper, which was the first communication of the EC's CSR strategy, defined CSR in more detail as "the notion that companies integrate social and environmental concerns into their business activities on a voluntary basis, and in their contacts with other interested parties". Within these ten years and

[1] Konstantinos Demestichas, Konstantina Remoundou, Evgenia Adamopoulou
National Technical University of Athens
cdemest@cn.ntua.gr, kremoundou@cn.ntua.gr, eadam@cn.ntua.gr,
[2] Maria Ntaliani, Antonia Papagiannaki, Constantina Costopoulou
Agricultural University of Athens
ntaliani@aua.gr, tina@aua.gr

hence, economic and social conditions changes have affected CSR issues and perceptions, among businesses and societies.

The Greek CSR Network, set up in 2000, defines CSR as "the voluntary commitment of enterprises to integrate into business practices social and environmental actions without being imposed by law and to have a direct relationship with them or indirect cooperation with their activities". The Greek strategy for CSR has been based on European and international practice, emphasizing on Greece's economic environment (Skouloudis et. al., 2011). The aim of this effort was to promote a new development model based on a participatory process, including three basic dimensions; policies and standards, rules and tools, and fields of application.

The objective of this paper is to investigate CSR in the Greek agri-food sector through two perspectives, namely the business and the consumer perspective. The business perspective is based on the investigation of CSR activities of 222 food companies, medium and large sized. The consumer perspective is based on a questionnaire analysis of 200 participants, investigating their willingness to purchase products from companies practicing CSR activities. The remainder of this article is structured as follows: Section 2 gives an overview of CSR, its main applications and its relation to marketing. Section 3 describes good practices of CSR by agri-food businesses at global level. Sections 4 and 5 describe CSR activities in the Greek agri-food sector from the business perspective and the consumer perspective respectively. Section 6 concludes the paper.

2. CSR – History and Applications

2.1. History

The implementation of CSR was not actually incorporated globally until the 1950s. By that time, the social actions of companies mostly regarded donations made to various charity institutions. Typically, these actions did not follow any systematic pattern or scheduling. In the following years, Bowen (1953) identified the notion of "business responsibilities towards society" and analyzed the political actions or decisions that companies should follow to be widely accepted from the general public. This approach was the cornerstone for the concept of CSR. Until then, most businesses were forced to raise equity by undertaking activities targeted to increase business profit (Friedman, 2007). From that moment on, companies slowly started to consider additional factors that raise their value, including the clients', government's and society's satisfaction, as well as the employees' engagement (Clarkson, 1995, Donaldson & Preston, 1995), realizing (especially during the 1960s) that they have a responsibility towards their environment that exceeds legal and financial obligations (McGuire, 1963). Over the next decade, Johnson (1971) argued that the company should not have as sole incentive to gain more profits for its shareholders, but should pay more attention to workers, suppliers, traders, local communities and the nation. Wood (1991) defined CSR as a form of corporate self-commitment, embedded in a business model.

A key element of this era is the term "stakeholders"; in principle, a socially responsible business should be governed by the stakeholder theory (Lin, 2018). This was based on Smith (2003), who argued that companies should take account of the results of their actions to all those involved, even if this is not profitable. Shareholders, on the other hand, were opposed to this concept, claiming that the sole responsibility of a company is to generate profits for its shareholders in a legal way. This, however, leads to the adoption of short-term policies; in contrast, CSR is based on actions which build a long-term investment that yields lasting and sustainable benefits. More generally, the 1970s could be described as a period of scientific analysis to define the concept of CSR, rather than a period of application to businesses. In the 2000s, businesses began to be more interested in developing and applying best practices of CSR, while several companies started to publicize and promote social work. In 2001, the EC presented the Green Paper in its effort to promote a "European Corporate Social Responsibility Framework", which was intended to launch a wider public debate on how the European Union can promote CSR at both European and international level. The main aspects of this policy included the following: i) the increase of the number of the European companies adopting and promoting CSR strategies; ii) the rise of companies' standards regarding social development, environmental protection and respect of fundamental rights; iii) the embrace of open governance schemes that reconcile the interests of various stakeholders and are quality-driven and sustainability oriented; and iv) the development of new partnerships and relationships within the companies, especially in terms of social dialogue, skills acquisition, equal opportunities, anticipation and management of change, reinforcement of economic and social cohesion, and health protection.

2.2. Types of CSR Activities

An organisation may carry out a variety of different CSR activities (Zinczuk, 2012). These can generally be distinguished into types of activities that are related to:

- *Human resources*: development of skills and staff training, labor rights protection, health and safety, work-life balance;
- *Business leadership*: fostering CSR business culture, developing trust and transparency relationships with stakeholders;
- *Market*: product and service development incorporating CSR principles;
- *Natural environment*: improvement of energy footprint, integration of environmental criteria; and
- *Social environment*: developing good relationships with local communities, supporting social groups in need.

2.3. CSR and Marketing

A CSR strategy designed on the basis of business operations can comprise a major competitive advantage for a company. Some elements of CSR activities, although observable and measurable, are difficult for consumers and stakeholders to assess. In this context, the most effective policy of promoting CSR is to involve people in these activities, so that they feel an important part of the process. In addition, a company's reputation is a key asset for its profitability. In recent years, a pillar of businesses' decision-making is dedicated to customer satisfaction and customer service concerns, including environmental issues. This creates a business profile that is responsible and ethical, boosting moral values through word of mouth while increasing corporate sales (Schreck, 2011).

There is a straightforward correlation between marketing and CSR (Farina & Burnaz, 2019). Some marketing strategies are specified within the framework of CSR, as follows:

- *Promotions of Social Causes:* This refers to a business that offers funds and resources in order to promote a social subject, to support a money-raising effort or to increase its participation in it.
- *Cause-relating Marketing*: It concerns a business committed to contribute or donate a percentage of its revenue for a particular purpose. Typically, such actions are conducted for a specific time, a specific product and a specific institution.
- *Corporate Social Marketing*: It refers to a company that supports the change of public attitudes on a social issue.

- *Corporate Charity*: This is directly related to the contribution or donation of a business to an institution or organization.
- *Community Voluntary*: This refers to a business that encourages its employees, partners, suppliers and customers to participate in a social purpose or activity. They can also donate working hours to their employees, in order to support a social purpose.
- *Socially Responsible Business Practices:* It includes practices and investments that are implemented on a voluntary basis and are aimed at supporting social goals and environmental protection.

3. Current Status of CSR in the Agri-Food Sector

According to Vaxevanidou (2011), the need to apply CSR to business practice has evolved as a result of a global trend. Globalization has led businesses to operate in accordance with international conventions, such as the United Nations International Human Rights Treaties. The three mains pillars behind the existence and sustainability of a business are economic growth, ecological balance and social progress.

Studies carried out globally show that the objectives, content and intensity of CSR activities vary across companies, regions and countries. For example, Ewing and Windisch (2007) argue that the Western approach to CSR can fail in Asian countries due to cultural differences. Welford (2007) adds that CSR in Asia is characterized by both the cultural context and the economic and political conditions. Alon et al. (2010), after conducting a survey in 105 companies from Brazil, Russia, India and China (BRIC), concluded that most businesses specializing in a wide range of sectors (e.g., finance, materials, fuel) perform CSR activities. Only eight of the 105 companies did not implement any CSR actions.

Regarding the agri-food sector, CSR has emerged as an important issue since the beginning of the 1990s. It was a series of defamatory campaigns that motivated major multinational companies, such good paradigms of CSR are the following: Haagen Danz in order to raise awareness for the decrease of the bee population, developed a website, as well as an associated campaign in social media, in order to support research activities on the matter (Ralston and Maignan, 2002). In 2010, Mondelez published its key environmental strategies and promoted a set of relevant policies. The main concern was the sustainability of the planet, assuring that further actions should be taken to reduce the environmental footprint and improve the dietary profile of the products. This would ensure sustainability in agricultural supplies, companies and the society. Mondelez also presented the Call for Well-Beginning strategy, which started in 2013 and focuses on the health and sustainability of the planet (Mondelez International, 2019).

4. The Case Study of Agri-Food Companies

4.1. Overview of the CSR Case Study

In order to address the research question of this paper, 222 SMEs and large agri-food companies operating in Greece were reviewed in the period of 2016-2017, constituting a significantly representative sample. The main objective was to record CSR activities of these companies, examining their corporate websites. More specifically, the companies were classified into groups based on their products. Hence, 14 agri-food sub-sectors were identified. Companies operate in these sub-sectors as follows: dairy products (13%), confectionery & snacks (10%), poultry, meat & sausage (16%), pasta-rice-legumes (5%), flour (4%), fish farming (6%), general food commerce (6%), fruit & vegetables (16%), olives & olive oil (9%), pastries (10%), canned food (2%), and drinks & beverage (3%).

CSR activities under review were classified into three different types, namely: i) social activities, including charities, local community support programs, and others; ii) environmental activities, such as recycling initiatives, promotion of activities for reduced carbon footprint, etc.; iii) human resources activities, addressing volunteering events, and promotion of social responsibility practices. In this study, each CSR activity is classified into exactly one of these types (i.e. there are no overlaps). However, a company might be performing activities that belong into one, two or all of the three types. Each company performing CSR activities selects the most appropriate methods and tools for communication, in order to publicize their interest and dedication in addressing problems of the local community.

4.2. Analysis and Results

The companies under study were initially classified by subsector. Following that, another important element that was taken into consideration was the size of the company, i.e. SMEs or large enterprises. Subsequently, this study counted the companies that invest in CSR activities or not, in order to determine the type of companies that demonstrate higher interest in CSR. It was found that 29.3% out of the 222 agri-food companies had actually performed any kind of CSR activity. 64.15% of the large companies were applying CSR, while only 18.45% of the SMEs did not. Figure 1 provides an overview of the companies' engagement in the three types of CSR activities. As observed, large companies give more emphasis on CSR. In addition, companies mainly support social activities, give less support to environmental activities, and even less support to human resource activities.

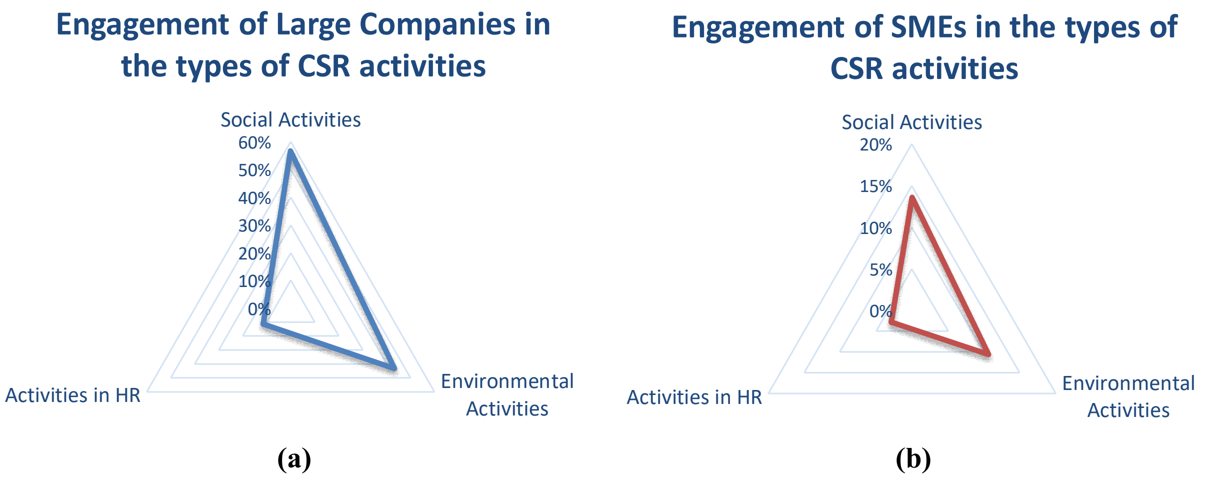


Figure 1. Engagement of (a) large companies and (b) SMEs in CSR activities in the Greek agri-food sector

Table 1 presents the number of the large and SME agri-food companies that perform CSR activities in detail; the activities are categorized according to the three distinct CSR types: social activities, environmental activities and human resource activities.

More specifically, the following conclusions can be extracted:

Dairy products: 4 out of 5 (80%) of the large companies that were studied in this paper, had active CSR activities, while *60% of the large companies were focused on multiple CSR activities*. On the other hand, only 4 out of 24 (16.6%) of the SMEs had active CSR activities, while *4% of the SMEs were focused on multiple CSR activities.*

Confectionery & snacks: 4 out of 9 (44%) large companies that were studied in this paper had active CSR activities, while *33% of the large companies were focused on multiple*

CSR activities. On the other hand, only 4 out of 14 (28.5%) of the SMEs had active CSR activities, while *14.2% of the SMEs were focused on multiple CSR activities.*

Poultry: 2 out of 3 (66%) large companies that were studied in this paper had active CSR activities, and *all of them were focused on multiple CSR activities.* On the other hand, only 4 out of 13 (30%) of the SMEs had active CSR activities, while *15% of the SMEs were focused on multiple CSR activities.*

Meat: 2 out of 5 (40%) large companies that were studied in this paper had active CSR activities, while *all of them were focused on multiple CSR activities*. On the other hand, only 1 out of 14 (7%) of the SMEs had active CSR activities, and this company was actually *focused on multiple CSR activities.*

Pasta/Rice: 1 out of 2 (50%) of the large companies that were studied in this paper had active CSR activities, which was focused on both social activities and environmental activities. On the other hand, only 1 out of 9 (11%) of the SMEs had active CSR activities, and this company was focused only on human resources.

Flour: 1 out of 2 (50%) of the large companies that were studied in this paper had active CSR activities, and this company was focused on both social activities and environmental activities. On the other hand, only 1 out of 6 (16%) of the SMEs had active CSR activities, and this company was focused on both social activities and environmental activities.

Fish farming: 4 out of 5 (80%) of the large companies that were studied in this paper had active CSR activities, while *40% of the large companies were focused on multiple CSR activities*. On the other hand, 4 out of 9 (44%) of the SMEs had active CSR activities, while *22% of the SMEs were focused on multiple CSR activities.*

General goods: 4 out of 4 (100%) of the large companies that were studied in this paper had active CSR activities, while *50% of the large companies were focused on multiple CSR activities*. On the other hand, only 1 small-medium company in this specific sector was studied for this paper, and this company did not have any CSR activities.

Fruit & Vegetables: 3 out of 4 (75%) of the large companies that were studied in this paper had active CSR activities, while *50% of the large companies were focused on multiple CSR activities.* On the other hand, only 6 out of the 31 (19.4%) of the SMEs had active CSR activities, while *13% of the SMEs were focused on multiple CSR activities.*

Olives and Olive oil: 1 out of 2 (50%) of the large companies that were studied in this paper, had active CSR activities, which was focused on social activities and environmental activities. On the other hand, no company out of the 17 SMEs that were studied for this paper had any CSR activities.

Pastries: 5 out of 8 (62.5%) of the large companies that were studied in this paper, had active CSR activities, while *12.5% of the large companies were focused on multiple CSR activities.* On the other hand, only 2 out of 14 (14%) of the SMEs had active CSR activities, while *none of them were focused on multiple CSR activities.*

Canned food: only one large company was studied in this paper, which had active CSR activities on social activities and environmental. On the other hand, 2 out of the 4 (50%) of the SMEs had active CSR activities, while *none of them were focused on multiple CSR activities.*

Animal feed: only SMEs were studied in this paper, from which only 1 out of 9 (11%) in total had any CSR activities focused on society.

Beverage: 2 out of 3 (67%) of the large companies that were studied in this paper had active CSR activities, while *33% of the large companies were focused on multiple CSR activities*. On the other hand, only 2 out of 4 (50%) of the SMEs had active CSR activities, while both of them were focused on social activities only.

Table 1. Number and percentage of Greek agri-food companies that perform CSR activities

Agri-food Sector	Social Activities		Environmental Activities		Human Resources Activities		Total No. of Companies	
	Large	SME	Large	SME	Large	SME	Large	SME
Dairy	4 (80%)	1 (4%)	2 (40%)	3 (12.5%)	1 (25%)	1 (4%)	5 (100%)	24 (100%)
Snacks	4 (44%)	3 (21%)	3 (33%)	2 (14%)	1 (11%)	2 (14%)	9 (100%)	14 (100%)
Poultry	2 (66%)	3 (23%)	2 (66%)	3 (23%)	1 (33%)	0 (0%)	3 (100%)	13 (100%)
Meat	1 (20%)	1 (7%)	2 (40%)	1 (7%)	2 (40%)	0 (0%)	5 (100%)	14 (100%)
Pasta/Rice	1 (50%)	0 (0%)	1 (50%)	0 (0%)	0 (0%)	1 (11%)	2 (100%)	9 (100%)
Flour	1 (50%)	1 (16.6%)	1 (50%)	1 (16.6%)	0 (0%)	0 (0%)	2 (100%)	6 (100%)
Fish Farming	3 (60%)	4 (44%)	2 (40%)	2 (22%)	1 (20%)	0 (0%)	5 (100%)	9 (100%)
General goods	4 (100%)	0 (0%)	2 (50%)	0 (0%)	0 (0%)	0 (0%)	4 (100%)	1 (100%)
Fruits and Vegetables	2 (50%)	5 (16%)	3 (75%)	4 (13%)	0 (0%)	1 (3%)	4 (100%)	31 (100%)
Olives & Olive oil	1 (50%)	0 (0%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)	17 (100%)
Pastries	4 (50%)	1 (7%)	2 (25%)	1 (7%)	0 (0%)	0 (0%)	8 (100%)	14 (100%)
Canned food	1 (100%)	1 (25%)	1 (100%)	1 (25%)	0 (0%)	0 (0%)	1 (100%)	4 (100%)
Animal feed	-	1 (11%)	-	0 (0%)	-	0 (0%)	-	9 (100%)
Beverage	2 (66%)	2 (50%)	1 (33%)	0 (0%)	0 (0%)	0 (0%)	3 (100%)	4 (100%)
Total	**30 (56,6%)**	**23 (13,6%)**	**23 (43,4%)**	**18 (10,6%)**	**6 (11,3%)**	**5 (2,9%)**	**53 (100%)**	**169 (100%)**

5. Consumer Study on CSR

In order to observe how consumers perceive such CSR activities, a survey has been conducted from December 2016 till January 2017. The sample consisted of 200 persons residing in the region of Athens. In the survey, a questionnaire was used, based on relevant work by Liapakis et al. (2017) and included three parts: demographics, knowledge on CSR, and willingness to purchase products coming from companies with active CSR strategies. The questionnaire was developed in Greek and was pre-tested with a small group of volunteers so as to be checked for clarity and that it measures correctly consumers' attitude towards CSR. According to the test, the questions were re-adjusted and corrected by two experts. In total, consumers answered 21 questions regarding CSR awareness, purchasing habits, willingness to buy from companies that conduct CSR and socio-demographics.

As depicted in Figure 2, the questionnaire was answered equally by 100 male and 100 female consumers, mostly of ages between 36-55 years old (53%). Consumers were mostly high-school graduates (38.5%), followed by those with a university degree (30.5%). For the majority the annual salary ranged from 10k to 20k € (29%). Most of them were family members with spouses and children (40%), and they were usually involved in the purchasing of daily needs of products (58%).

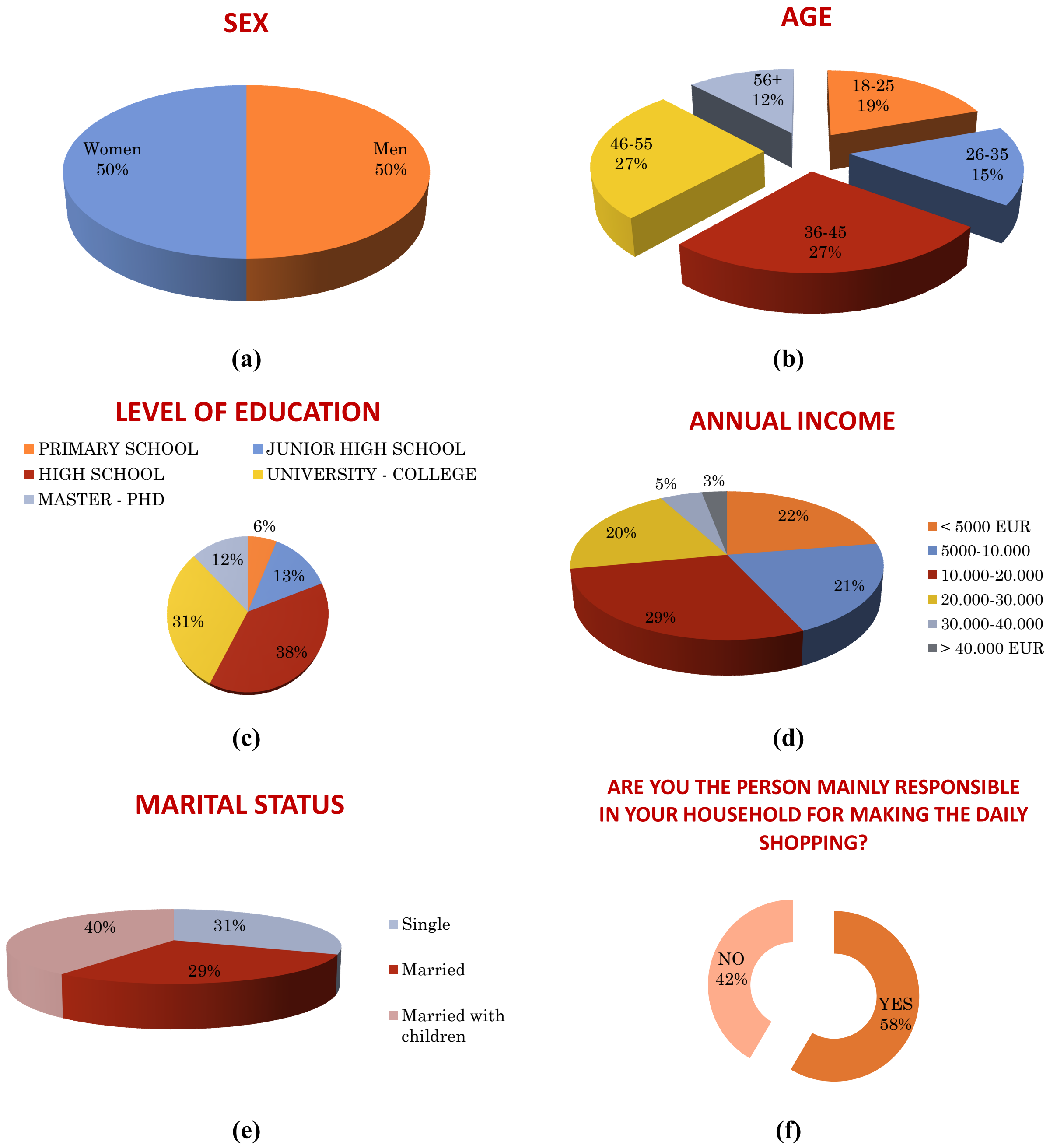
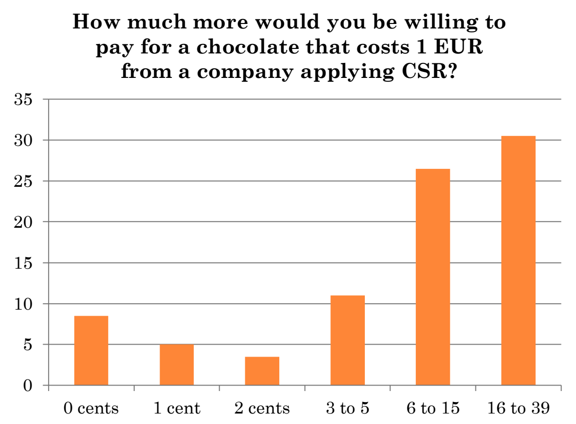


Figure 2. Demographics of the consumer survey (sample size N = 200) in terms of gender (a), age (b), education level (c), annual income (d), marital status (e), and shopping responsibilities (f)

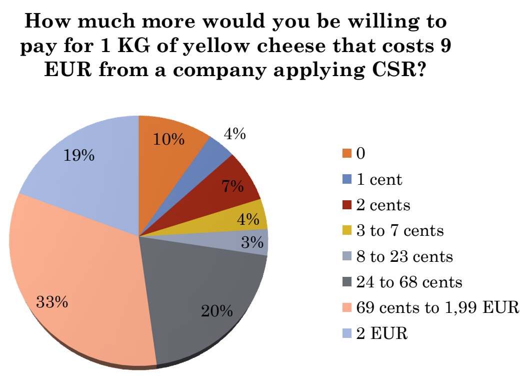
80% of the respondents considered that the social image of the company is very important when choosing a specific product, followed by the quality of the product (76%). The less common response was about their shopping habits, with only 11% of the respondents considering this is an important factor when choosing a product.

Regarding the consumers' knowledge on CSR, most of the respondents had a good knowledge regarding the CSR activities of the companies, but there was also a significant percentage (22%) of respondents that were unfamiliar with such actions. Most consumers who responded to the questionnaire would be purchasing products or services from companies active in CSR, because they feel that providing support to socially responsible strategies is worth the cost. 64% of the respondents feel socially responsible when consuming corporate products with CSR strategies for vulnerable community groups. A high percentage (35.5%) answered that they would probably prefer to buy a product from a company with social responsibility, while 17% did not care and would continue to buy the product from the company they were accustomed to.

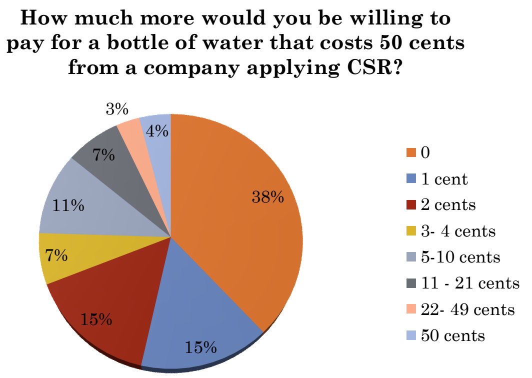
As depicted in Figure 3, the amount of the extra cost the respondents are willing to pay for a product combined with a CSR activity is analogous to the original price of the product. More specifically, for a product costing 1 €, the majority of the consumers (30.5%) was willing to pay around 16 to 39 cents more, while for a product at a cost of 9 €, the majority of the consumers (29%) were willing to pay 0.69 € to almost 2 €. However, for essential goods such as water, which has a fixed (regulated) price in the Greek market, most of the consumers (38%) were not willing to pay any extra cost.



(a)



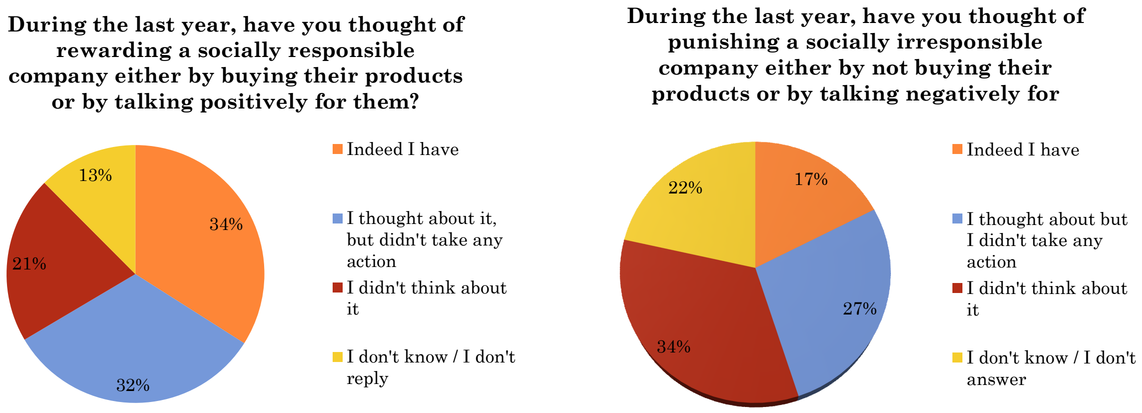
(b)



(c)

Figure 3. Consumer study: Willingness to pay more for a product offered by a company active in CSR

According to the respondents' replies, as illustrated in Figure 4, 34% had thought of rewarding a socially responsible company and actually did reward it by either buying its products or speaking positively to others. Nevertheless, 33.5% stated that, although they were not satisfied with the social behaviour of a company, they had not thought of "punishing" it, i.e. stop buying from it and talk "negatively" about it. Regarding the credibility of companies with active CSR actions in Greece, 41% of the respondents believe they are fairly reliable and 18.5% believes they are very reliable. However, there is also a significant percentage of 40.5% who consider them to be of limited reliability (29.5%) or not at all reliable (11%).



(a) **(b)**

Figure 4. Willingness of consumers to reward (a) or penalize (b) a company with respect to its CSR

6. Conclusions

Businesses strive to ensure not only their profitability but also their sustainability. By adopting a CSR strategy, businesses can increase their benefits in both direct and indirect ways, especially since consumers can more easily trust a company that has a socially responsible profile. However, in the Greek agri-food sector, the degree of correlation between business and social responsibility is low, especially when compared to other European countries that have been implementing such actions for a longer time. The survey regarding the 222 companies from various agri-food sectors, showed that only 29.3% perform CSR actions, i.e. less than 1/3 of the total number of companies, with a clear dominance of the large companies over the SMEs. The majority of implemented actions can be classified as social actions, and most of them concern cases regarding the protection and support of people in need (financial and psychological support). The above facts can be explained through the phenomenon of the financial crisis the Greek society struggles with for almost the last ten years.

The majority of businesses associated with CSR choose their corporate website to promote their actions, even if they have other networking tools such as social media (e.g. Facebook). Although the majority of consumers in Greece is aware of CSR, a significant part of them (40%) tends to be sceptical with regards to the reliability of the performed CSR activities. However, 8 out of 10 of the respondents in the survey consider that the social image of the company is very important in order to choose a specific product, which comes to confirm that CSR and marketing are the two sides of the same coin for the success of a business. In addition, the majority of the respondents mentioned that they are willing to reward a company that follows a specific CSR strategy, by either buying their products, paying an additional cost that is proportionate to the initial product price, and/or recommending them to other people. Future work will focus on reporting CSR activities promotion through social media by Greek agri-food companies.

References

Aguinis, H. & Glavas, A., 2012. What we know and don't know about corporate social responsibility: A review and research agenda. *Journal of management*, 38(4), pp.932-968. https://doi.org/10.1177/0149206311436079

Alon, I., Lattemann, C., Fetscherin, M., Li, S. & Schneider, A.M., 2010. Usage of public corporate communications of social responsibility in Brazil, Russia, India and China (BRIC). *International Journal of Emerging Markets*, 5(1), pp.6-22. https://doi.org/10.1108/17468801011018248

Bowen, R., 1953. *Social Responsibility of Businessman*, New York, Harper & Row.

Carroll, A.B. & Shabana, K.M., 2010. The business case for corporate social responsibility: A review of concepts, research and practice. *International journal of management reviews*, 12(1), pp.85-105. https://doi.org/10.1111/j.1468-2370.2009.00275.x

Clarkson, M.E., 1995. A stakeholder framework for analyzing and evaluating corporate social performance. *Academy of management review*, 20(1), pp.92-117. https://doi.org/10.5465/amr.1995.9503271994

Donaldson, T. & Preston, L.E., 1995. The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of management Review*, 20(1), pp.65-91. https://doi.org/10.5465/amr.1995.9503271992

European Commission, Green Paper - Promoting a European framework for Corporate Social Responsibility, 2001. European Commission, viewed 7 December 2019, <https://ec.europa.eu/commission/presscorner/detail/en/DOC_01_9>.

Ewing, M.T. & Windisch, L.E., 2007. Contemporary corporate social responsibility in China: an extension of confucian philosophy?.

Farina, I., & Burnaz, S. (2019). *Ethics, Social Responsibility and Sustainability in Marketing*. Springer. Springer Singapore.

Friedman, M., 2007. The social responsibility of business is to increase its profits. In *Corporate ethics and corporate governance* (pp.173-178). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70818-6_14

Johnson, H.L., 1971. *Business in contemporary society: Framework and issues*. Wadsworth Publishing.

Liapakis, A., Costopoulou, C., Tsiligiridis, T. & Sideridis, A., 2017. Studying Corporate Social Responsibility Activities in the Agri-Food Sector: The Greek Case. International Journal of Agricultural and Environmental Information Systems (IJAEIS), 8(1), pp.1-13. https://doi.org/10.4018/978-1-5225-6192-7.ch082

Lin, Tom C. W, 2018. Incorporating Social Activism. *Boston University Law Review*, vol. 98, pp. 1535-1605.

Lindgreen, A. & Swaen, V., 2010. Corporate social responsibility. *International Journal of Management Reviews*, 12(1), pp.1-7. https://doi.org/10.1111/j.1468-2370.2009.00277.x

Maignan, I. & Ralston, D. J, 2002. Corporate Social Responsibility in Europe and the U.S.: Insights from Businesses' Self-presentations. *Journal of International Business Studies,* 33(3), pp.497-514. https://doi.org/10.1057/palgrave.jibs.8491028

McGuire, J., 1963. *Business and Society*, McGrow-Hill, New York.

Mondelez International 2019, Well-being Library, Mondelez International, viewed 7 December 2019, <https://www.mondelezinternational.com/impact/well-being-library>.

Schreck, P., 2011. Reviewing the business case for corporate social responsibility: New evidence and analysis. *Journal of Business Ethics*, pp.103-167.

Secchi, D. (2007). Utilitarian, managerial and relational theories of corporate social responsibility. International Journal of Management Reviews, 9(4), pp.347-373. https://doi.org/10.1111/j.1468-2370.2007.00215.x

Smith, N.C., 2003. Corporate social responsibility: whether or how?. *California management review*, *45*(4), pp.52-76. https://doi.org/10.2307/41166188

Skouloudis, A., Evangelinos, K., Nikolaou, I. & Filho, W. (2011). An overview of corporate social responsibility in Greece: Perceptions, developments and barriers to overcome. Business Ethics: A European Review, 20(2), pp.205-226. https://doi.org/10.1111/j.1467-8608.2011.01619.x

Porter, M.E. & Kramer, M.R., 2006. The link between competitive advantage and corporate social responsibility. *Harvard business review*, *84*(12), pp.78-92.

Vaxevanidou, M., 2011. Corporate Social Responsibility. Stamoulis Editions, Athens.

Welford, R., 2007. Corporate governance and corporate social responsibility: issues for Asia. Corporate Social Responsibility and Environmental Management, 14(1), pp.42-51. https://doi.org/10.1002/csr.139

Wood, D., 1991. Corporate Social Performance Revisited. The Academy of Management Review 16(4), 691–717. https://doi.org/10.2307/258977

Zinczuk, B., 2012. Corporate Social Responsibility and its Application in the Enterprise. In Knowledge and Learning: Global Empowerment; Proceedings of the Management, Knowledge and Learning International Conference 2012 (pp. 525-531). International School for Social and Business Studies, Celje, Slovenia.

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 10, No. 2
journal.magisz.org

The opinion of farmers and small and medium-sized enterprises on the importance of ICT in Hajdú-Bihar County, Hungary

Sándor Boros-Papp[1], László Várallyai[2]

INFO
Received 20-12-2019
Accepted 13-01-2020
Available online 28-01-2020
Responsible Editor: M. Herdon

Keywords:
ICT, SME, primary producer

ABSTRACT

The purpose of this study is to gain a better understanding, conceptualization, and measurement about the usage of ICT among SMEs that could enhance and transform their business to be more efficient and resilient. A more specific aim is to explore the state of ICT adoption as well as to examine the drivers and barriers of digital transformation through the lens of SMEs. In the primary (quantitative) phase of our research, we conducted a questionnaire survey among micro-enterprises and primary producers on the use of ICT in Hajdú-Bihar County, where 237 people answered the questionnaire. In all cases, the respondent was listed as a primary producer or in a joint primary producer certificate.

We were investigated which tools are using in communication with suppliers, customers and government organisations, which type of cloud services are used and the safe storage by Internet Service Providers. We get about the same results in the case of the suppliers, customers. In the case of the governmental institutions, Internet usage was bigger, because of a significant number of them require national or EU grants, which can be submitted electronically. The secure data storage is the safest possible for the farmers.

1. Introduction

In Hungary, the concept of a farmer's certificate is provided by the Governmental Decree 436/2015. (XII. 28.). At the end of 2017, there were 1 million 720 thousand registered enterprises in Hungary, which consist of 256.758 primary producers. The number shows that we are talking about a non-negligible economic form. Its peculiarity is we can talk about one or more person's economies. We know many benefits of Information and Communication Technologies (ICT), including financial efficiency and sustainability (Gouvea et al., 2018). At the same time, Information and Communication Technologies (ICT) have become the engine of economic growth today (Fleischer, 2003). It is also important to note that due to the continuous acceleration of technological development, not only new markets can be opened, but also competitors may appear (Hetyei, 2001).

The food sector is increasingly becoming a closely interconnected system with a whole network of complex relationships. Consequently, supply chains are highly complex systems and they are built on a lot of partnerships and forms of cooperation. The length and complexity of the chains have a significant effect on the quality of information flow among chain members and the number of inventory levels. Information and Communication Technologies (ICTs) play an increasingly important role in supply chain management. The use of ICT is unavoidable in business relationships and analyses on the scientific level can only determine their effect on business activities. The present paper aims to overview the level of ICT infrastructure of the Hungarian SME sector and the primary producer sector at a regional level. At this stage of our investigation, we rely on primary data form Hajdú-Bihar county. We used data for 2018 and 2019, on the other hand, our investigation we rely on seconder data (HCSO - Hungarian Central Statistical Office - and Eurostat) in case of the cloud services. We used data for 2015 and 2016 as these are the latest data available.

[1] Sándor Boros-Papp
University of Debrecen
papps@hotmail.hu
[2] László Várallyai
University of Debrecen
varallyai.laszlo@econ.unideb.hu

Small and medium-sized enterprises (SMEs) are the engines of the European and Hungarian economies (Billing, 2017). The adoption of the European Charter also demonstrates this for Small Enterprises in 2000 (European Commission, 2004). Their specificity is greater flexibility, better innovation capacity, lower fixed costs, smaller market power, and less capital characterize them (Motwani et al., 1998). There are a strong presence of SMEs in Europe and their contribution to employment and growth. According to the Annual Report on SMEs produced by DG Enterprise and Industry in 2014, SMEs account for 99 percent of all businesses, provide 67 percent of all employment, and almost 60 percent of the value-added in the EU (Billing, 2017).

These enterprises have intuitive and unstructured decision-making style (Gilmore et al., 2001). One of the main objectives of SMEs, like any business, is to enhance their competitiveness. If an enterprise wants to achieve long-term competitiveness, it must be cooperative in order to obtain knowledge and technological devices (Nyíri & Szakály, 2010). In 2014, 99.8 percent of Hungarian enterprises were SMEs, which employed 69.8 percent of the employees (Nemzetgazdasági Minisztérium, 2016). According to Holicza (2016), the significant problems the SMEs are faced with are the followings:

- education does not prepare for starting a business,
- insufficient financing,
- if no success is achieved, losses may be significant, complicated administration procedures.

It is to be noted that the Master course of Business development has been launched in 2009 at the Faculty of Economics and Business of University of Debrecen and one of its main purposes is to reduce these failures as mentioned above. To reduce administrative burdens „The Small and Medium Enterprises Support Strategy 2014–2020" supports the „Think small first" principle (EC Vállalkozáspolitikai és Ipari Főigazgatóság, 2016). The lack of further training and IT knowledge are among the problems of the SMEs (Chikán et al., 2014).

The use of ICT remains a key challenge for most companies. The major part of the enterprises lacks the appropriate knowledge about the opportunities of ICTs. It is a quite complex system and one of its crucial points is the quality of the Internet access (availability, speed, and bandwidth).

Meanwhile, the European Commission cites that SME is among the catalyst that spurs economic growth, innovation and social integration in the EU (European Commission, 2019). Hungary is among the member of the EU that receive a positive impact on their economy as EU policy helps make the industry and business more competitive since joined in on May 1st, 2004.

Towlson (1993) said Hungary is one of the fastest moving central European countries in the shift towards the market economy. Their SMEs sector was dominated by micro firms and makes an important contribution to the domestic labour market when created an additional 25.000 jobs from 2002 to 2008 (Azam, 2014).

In the past few years, the growth in ICT usage has increased tremendously across industries around the world. Various studies (Moghavvemi et al., 2011) found that SMEs have a high potential to develop their businesses by taking proper advantage of Information and communications technology (ICT). ICT has an enormous role in enhancing the growth of SMEs not only through an increase in efficiency and productivity but also in expanding their market reach.

According to Azam (2014), the previous study revealed that ICT would enhance their performance. However, the usage of ICT in helping to facilitate one or more strategic choice is debatable. Harindranath et al., 2008 have expressed the opposing views to the value of IT in achieving superior business performance. Thus, the adoption is widely seen as critical for the competitiveness of SMEs in the emerging global market.

2. Materials methods

There is no doubt about the relevance of this topic, as solutions enabling traceability in supply chains are becoming more widespread today. This is especially true for the food industry. A good example of this is the "From the Ground to the Table" directive, which aims to track the lifecycle of the product purchased fully. Among the primary producers and micro-enterprises on this topic, focusing on ICT.

The purpose of this study is to gain a better understanding, conceptualization, and measurement about the usage of ICT among SMEs that could enhance and transform their business to be more efficient and resilient. A more specific aim is to explore the state of ICT adoption as well as to examine the drivers and barriers of digital transformation through the lens of SMEs.

Two main objectives had been identified in this study. The objectives are:

- To identify how the SMEs evaluate the ICT tools that enable SMEs to grow their business and enhance their organizational performance.
- To examine the factors that contribute to the adoption of ICT at regional level (Northern Great Plain region)

In the primary (quantitative) phase of our research, we conducted a questionnaire survey among micro-enterprises and primary producers on the use of ICT in Hajdú-Bihar County. Data collection from the other two counties in the region (Jász-Nagykun-Szolnok and Szabolcs-Szatmár-Bereg) is ongoing. In our questionnaire, closed-ended questions were chosen in the hope of getting an accurate answer.

When designing the questionnaire, we have considered the following:

- Exact defining questions (subjects of research, hypotheses).
- Choose how to ask questions (in groups or individually).
- Choosing the right structure
- Formulation of questions (closed or open)
- Questionnaire format (striving for a short, clear questionnaire)
- Small sample testing (small group pre-testing, filtering out possible misunderstandings)

The data of the questionnaire published in this article, which collected in the fourth quarter of 2018 and in the first quarter of 2019 with the help of the Farmers' Network of the Hungarian Chamber of Agriculture.

The questionnaire was filled by SMEs and primary producers from the listed 3 counties, which contain 3 question groups with 21 questions. The question groups focused on the main objective of the study.

The sampling was random, and the farmers who came to the extension with problems answered their current questions. 237 people answered the questionnaire. In all cases, the respondent was listed as a primary producer or in a joint primary producer certificate. Answers came from all 10 districts of the county. Taking into account the economic activity, data were provided by producers, processors, and persons carrying out both activities.

The main objective of the article is how the surveyed primary producers evaluate the importance of internet interfaces, mobile phones, wired phones in corporate communication (suppliers, customers-clients, government institutions). Respondents scored on a 5-point Likert scale.

3. Results

3.1. The importance of tools in communicating with suppliers

Figure 1, shows the results of the communication with suppliers, the wired phone is the worst (value 1 on Likert scale: 39%); the most important thing is to communicate with the mobile phone (value 5 on Likert scale: 70%), but the internet also plays an important role (value 5 on Likert scale: 45%). It is can be understood because most of the chain operators mainly use mobile devices for communication instead of the wired phone. The other possible way is Internet services (which are working via the Internet – e-mail, information system, etc.) because in more cases the processors requested the usage of his system on the Internet.

Figure 1. The importance of tools in communicating with suppliers (Own edition 2019)

3.2. The importance of tools in communicating with customers

The next step was to investigate the importance of tools in communicating with customers, where results are illustrated in Figure 2. The communication with customers, the wired phone is the worst (value 1 on Likert scale: 36%); the most important thing is to communicate with the mobile phone (value 5 on Likert scale: 70%), but the internet also plays an important role (value 5 on Likert scale: 49%). So we can conclude, the same tendency can be seen in customer relations as in the case of the suppliers. The percentage values are about the same as was in the case of the suppliers.

Figure 2. The importance of tools in communicating with customers and clients (Own edition 2019)

3.3. The importance of tools in communicating with governmental institutions

The next step was to determine the communicating with the governmental institutions. The results shown in Figure 3. The communication with suppliers, the wired phone is the worst (value 1 on Likert scale: 32%); the most important thing is to communicate with the mobile phone (value 5 on Likert scale: 61%), but the internet also plays an important role (value 5 on Likert scale: 62%), here is the highest value. Here are about the same percentage use of mobile phones and the Internet. The usage of a wired phone is about the same as was in Figures 1 and 2.

Not all primary producers, but a significant number of them require national or EU grants, which can be submitted electronically. Therefore, the usage of the Internet is probably higher here than in the other case of communication of customers or suppliers, because not only took the request on the Internet but also one of the methods of changing the data in various registers and the most up-to-date information can be found here.

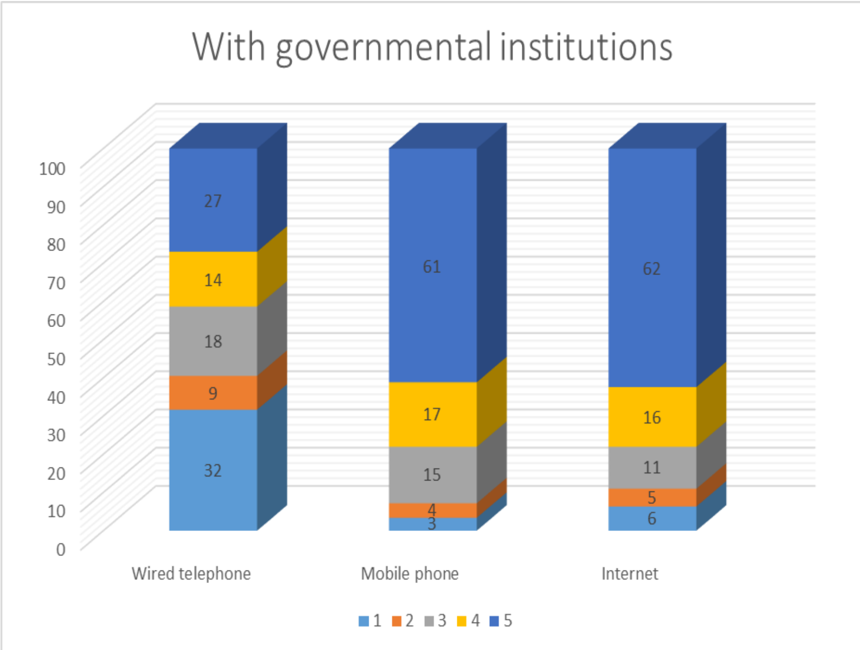


Figure 3. The importance of tools in communicating with governmental institutions (Own edition 2019)

3.4. Use of cloud services

The essence of the cloud services is that no considerable ICT knowledge is required for the enterprises and primary producers, also they do not have to employ ICT experts. Thus, the labour-related costs become available and the resulting savings can be used to purchase services, often for a smaller budget. Cloud services have several advantages, it will not only improve competitiveness but simultaneously enhance the efficiency and by correct use it may also increase data protection.

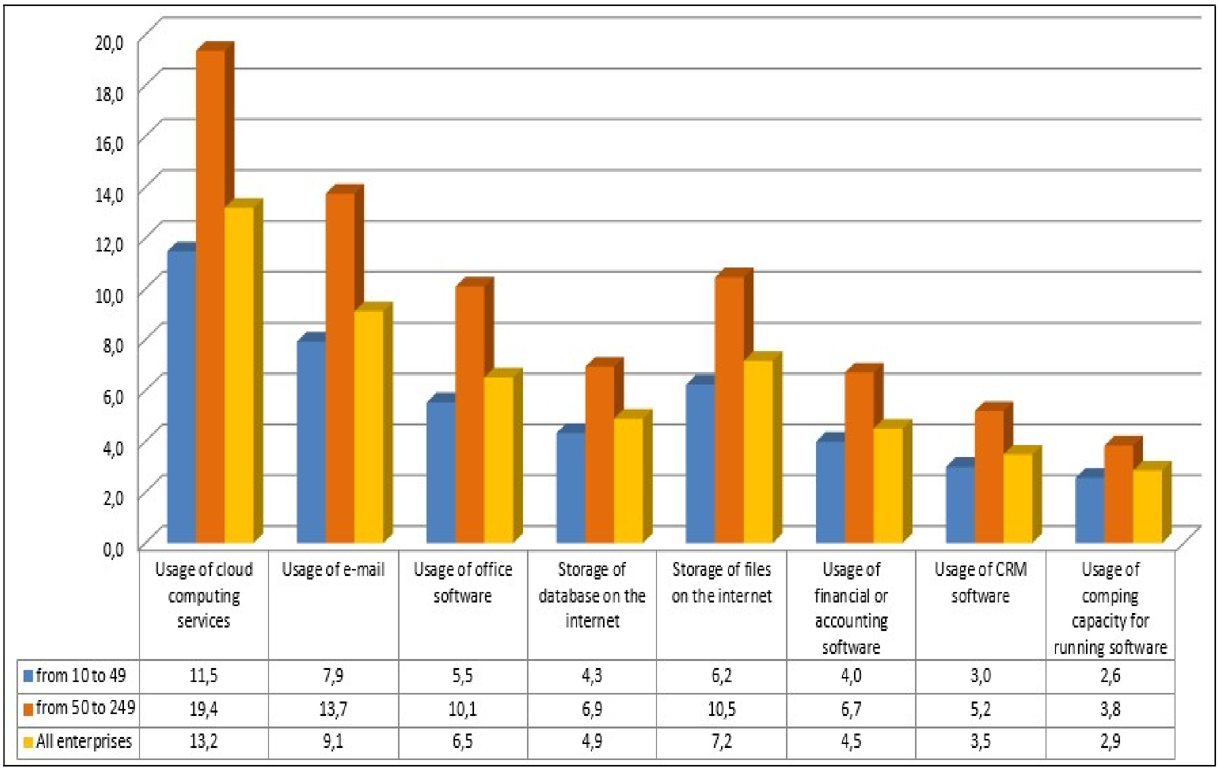
Data traffic from cloud services is steadily growing because economic competition requires the enterprises to react faster to the environmental changes. There are three kinds of cloud service models (Harding, 2011):

- IaaS, Infrastructure as a Service,
- PaaS, Platform as a Service,
- SaaS, Software as a Service.

IaaS (Cloud Infrastructure as a Service): This model is the basis of cloud services. Enterprises access memory space and installed applications in a virtual environment, however, maintenance has to be done by the users.

Paas (Cloud Platform as a Service): The service provider installs the required applications written in the selected programming language. Maintenance is also the responsibility of users.

Saas (Cloud Software as a Service): Applications of the service provider run in the cloud infrastructure and can be accessed via web browsers. They do not require large-scale investment and can be used immediately. The usage level of different cloud services by the Hungarian small and medium-sized enterprises presented is in Figure 4.



	Usage of cloud computing services	Usage of e-mail	Usage of office software	Storage of database on the internet	Storage of files on the internet	Usage of financial or accounting software	Usage of CRM software	Usage of comping capacity for running software
from 10 to 49	11,5	7,9	5,5	4,3	6,2	4,0	3,0	2,6
from 50 to 249	19,4	13,7	10,1	6,9	10,5	6,7	5,2	3,8
All enterprises	13,2	9,1	6,5	4,9	7,2	4,5	3,5	2,9

Figure 4. Usage of cloud services by size class (in the percentage of enterprises using the internet.
(Own edition based on Eurostat data 2016)

Figure 4 shows a clear difference in the usage level of cloud services between small enterprises and medium enterprises and considers everything the latter enterprise category (with 50-249 persons employed) places more emphasis on the use of cloud services.

3.5. The importance of safe storage by Internet service providers

In the survey (Figure 5), we also asked farmers for their opinion on the importance of secure storage of data with Internet Service Providers (1-not important, 5-very important). The percentage of values shows the distribution of secure storage data. Only 9% said it is not too important. Maybe they never lost important data. Nearly the half part (44%) said it is important. maybe they understand how important these data, what they stored in the cloud. In addition, 20%-20% on the Likert scale was chosen the 4 and 3, which strengthen the importance of secure data storage at the Internet Service Provider.

Based on the results, it can be concluded that, regardless of the fact that few people use cloud-based services, the reliability of cloud providers is particularly important. The average age of the farmers who rated question 5 was 47 years, while the number of respondents who ranked 1 was 56 years. They have not informatics knowledge and most of them understand, what the cloud is and store data in the cloud. Maybe some years later the younger farmers will understand the importance of the secure data on the cloud.

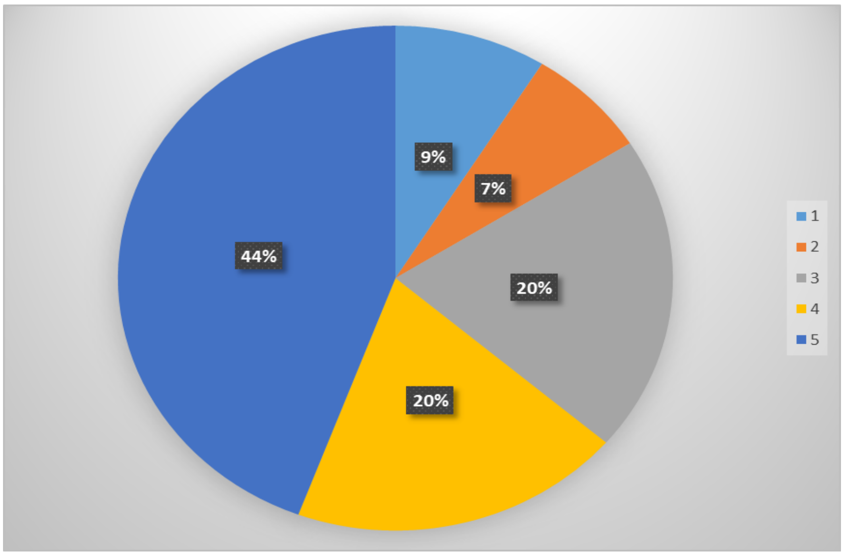


Figure 5. The importance of secure data storage by Internet Service Providers (Own edition 2019)

Conclusion

There is digital development when enterprises fully make use of the opportunities and advantages offered by digital technologies. In terms of several ICT usage indicators, the Hungarian SME sector and the primary producer sector are far lagging behind the EU average.

Appropriate usage levels of these basic IT devices and services enable this gap to be overcome and it is essential. By appropriate use of opportunities offered by ICT tools to maintain or boost the competitiveness can be easier achieved.

The ICT usage level of enterprises highly depends on the degree of IT development. We hope that the online sales proportion of enterprises will grow and they will increasingly focus on the opportunities by methods as targeted consumer access with marketing campaigns.

There are advantages in terms of procurement as well, for example, comparing products, purchase without having to queue or travel and with a short period of time for delivery at the address indicated. In light of the above, we conclude that the opportunities existing in the neglected sectors and regions which lagged behind in terms of IT are under-utilized, however, all enterprises try to adapt to the phenomenon of data boom while new market activity types and roles are developed.

References

Azam, M.S. (2014). Diffusion of ICT and SME Performance: The Mediating Effects of Integration and Utilisation, PhD Thesis, Curtin University, Australia. https://doi.org/10.1108/S1069-096420150000023005

Billing, P. (2017). Small and Medium-sized Enterprises in European Regions and Cities. ESPON EGTC.

Chikán, A., Czakó, E., and Wimmer, Á., 2014. Kilábalás göröngyös talajon. Gyorsjelentés a 2014 évi kérdőíves felmérés eredményeiről. Budapesti Corvinus Egyetem, Versenyképességi Kutató

EC Vállalkozáspolitikai és Ipari Főigazgatóság, 2016. 2015 évi SBA tájékoztató Magyarország. 1-17.

European Commission 2004. Kisvállalkozások Európai Chartája. Luxemburg. Az Európai Közösségek Hivatalos Kiadványainak Hivatala

Fleischer (2003). Az infrastruktúra-hálózatok és a gazdaság versenyképessége. A "Gazdasági versenyképesség: helyzetkép és az állami beavatkozás lehetőségei" című kutatás résztanulmánya. Pénzügyminisztérium, Budapest, 50 p.

Gilmore, A., Carson, D., and Grant, K., 2001. SME marketing in practice, Marketing Intelligence&Planning, 19 (1), 6-11. https://doi.org/10.1108/02634500110363583

Gouvea, R. – Kapelianis, D. – Kassicieh, S. (2018): Assessing the Nexus of Sustainability and Information & Communications Technology. *Technological Forecasting & Social Change.* 130 39–44. https://doi.org/10.1016/j.techfore.2017.07.023

Harding, C., 2011. Cloud Computing for Business. The Open Group Guide, The Open Group, Van Haren Publishing

Harindranath, G. Dyerson, R. and Barnes, D. (2008) ICT in small firms: Factors affecting the adoption and use of ICT in southeast England SMEs. The Electronic Journal Information Systems Evaluation Volume 11 Issue 2, pp. 52-63.

Hetyei (2001). Az informatika hozzájárulása a menedzsment feladatokhoz. 109-127p. In: Hetyei József (szerk.) Vezetői döntéstámogató és elektronikus üzleti megoldások Magyarországon, szerk. Hetyei József, Computerbooks Kiadó, 357 p.

Holicza, P., 2016 A magyar KKV szektor helyzete nemzeti és nemzetközi szinten. Vállalkozásfejlesztés a XXI. században. Budapest, 147-162.

http://ec.europa.eu. (2019) . Internal Market, Industry, Entrepreneurship, and SMEs. Retrieved from: https://ec.europa.eu/growth/smes_en (Accessed: Jan 12, 2019).

Krauszné, P. M., Krausz F., 2017. Proceedings of the Conference on Problem-based Learning in Engineering Education 30-37.

Moghavvemi, S., Mohd Salleh, N.A., Zhao, W (2011) An Empirical Study Of IT Innovation Adoption Among Small And Medium-Sized Enterprises In Klang Valley, Malaysia. Social Technologies Journal,1(2), pp. 267–282.

Motwani, J. A., Jiang, J.J., and Kumar, A., 1998: A comparative analysis of manufacturing practices of small vs. large Western Michigan organizations. Industrial Management & Data Systems, 98 (1), 8-11. https://doi.org/10.1108/02635579810199690

Nemzetgazdasági Minisztérium , 2016. KKV Évkönyv. A kis- és középvállalkozások helyzete Magyarországon 2014. Budapest, 1109.

Nyíri, A., Szakály, D., 2010. Absorb meter 89 p.

Széchenyi 2020, 2018. https://www.palyazat.gov.hu/ginop-324-82416-felhalap-iaas-paas-saas-vllalatiszolgltatsok-ikt-megoldsok-fejlesztsnek-spiaci-bevezetsnek-tmogatsa-1

Towlson, K (1993). Business information in Hungary Impressions of the contemporary scene", Aslib Proceedings, Vol. 45 Issue: 2, pp.35-37 https://doi.org/10.1108/eb051303

Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 10, No. 2.
journal.magisz.org

Digitization landscape in the Hungarian food-processing industry

Attila Sándor Debrenti[1]

INFO
Received 28-12-2019
Accepted 13-01-2020
Available on-line 28-01-2020
Responsible Editor:L. Várallyai

Keywords:
Digitalization, Business Intelligence, ERP, Food sector, Hungary, ICT, Industry 4.0

ABSTRACT

The food industry is the third-largest in Hungary, the first in Hungary in terms of the number of employees, and the first in Europe in the processing industry, as well as a significant user of resources. In order to develop the digitization of the food industry, it is necessary to get a picture of the situation. In this article, the most important functions, features, and aspects of these technologies are examined. Several international organizations are contributing to the measurement of digital transformation. Digital Enterprise Metrics is for the level of a single enterprise digitalization is measured with industry. My research focused on the technological aspects of digitalization. The research-based on a survey. According to the size of the company, the level of digitization increases and there are differences among branches in the food-processing sector.

1. Introduction

The "Food Industry Concept of Hungary 2017-2050" approved by the Government gives priority to the development of the food industry. Strengthening innovation, technological change, and digitalization in the food industry is also crucial for more efficient use of resources, reducing losses, increasing productivity, quality assurance and addressing labor shortages. Experts estimate that 80% of the food industry's problems could be solved by adapting Industry 4.0 and other digitization technologies already in use in other industries, but most food industry players are unaware of new technologies, while developers are not aware of the industry's problems. On the other hand, business success determined by a series of managerial decisions. With the advancement of information technology solutions, several tools are available to assist with large amounts of data processing, flexible querying, visualization, analysis of relationships and trends, and filtering of business process management information based on a pre-developed operating model to effectively support management decisions

In terms of production value, the food industry is the third-largest in Hungary, the first in Hungary in terms of the number of employees, and the first in Europe in the processing industry, as well as a significant user of resources. Agri-food enterprises operate in a complex and dynamic environment (Wolfert et al. 2010) and facing different challenges, such as financial issues, lack of technical skills and investment in the business, etc. (Singh et al. 2019, Zhang 2012). In this sense, several key factors can help food SMEs. Therefore, increasing productivity and efficiency in the food industry is extremely important, and the application of Industry 4.0 and digitization offers many new opportunities. In this way, new products, new services, new solutions, and approaches can be developed to use resources better, improve quality parameters, create new skills, and competencies.

However, the cloud begins to form a major stream of business environments. Businesses are increasingly questioning "when and why" is there a need for the cloud, and at the end of the acceptance process, the "how" is being sought (KPMG 2013). According to Botos's study, micro-enterprises in the agricultural sector do not look for cloud services or big data, because the enterprise management system is not used either (Botos et al. 2015).

The planned Hungarian digitalization strategy focuses on the entire food economy from the chain of raw material production to the final consumer, to the issues of industrial food processing. At present, this is the area of the Hungarian food processing industry where a solution should be found as soon as

[1] Attila Sándor Debrenti
University of Debrecen (PhD student)
debrenti@gmail.com

possible to the problems of increasing productivity, resource efficiency, labor shortages, skills shortages and which can bring tangible benefits the fastest.

Today, a wide range of management support systems and tools are available for the efficient operation of the company. The prerequisite for the application of these systems is the appropriate digitization of enterprises and their digital readiness. However, in order to develop the digitization of the food industry, it is necessary to get a picture of the situation. At what level are businesses, what tasks, what decisions, what functional areas are the current tools used, how do they judge the applicability of systems and tools and what plans do they have for developing digitization?

In this article, the most important functions, features, and aspects of these technologies for managerial decision support were examined: network infrastructure and Internet usage; use of integrated enterprise information systems (ERP); management information systems (MIS); opinions on some of the most important information technologies in business intelligence tools (BI) and Industry 4.0.

2. Literature review on measuring digitalization

2.1. The importance and impact of digitization

Across industries around the world, digital transformation is changing the way organizations of all sizes do business. Harnessing the power of the 3rd Platform (cloud, mobility, social business, and big data and analytics), organizations leverage new digital competencies to transform every step of the value chain. For organizations to remain competitive in this quickly unfolding digital era and beyond, IT must embrace digital transformation and the requisite infrastructure needed to achieve it.

While organizations are leveraging new digital competencies to revolutionize product and service delivery, it is important to remember that digital transformation is not a technology initiative but a business strategy. Advances in digital technologies are embedded in all sectors of the economy and contribute to improving productivity, reaching new markets, reducing costs, changing business processes, creating new business opportunities and new jobs.

Despite the ever-increasing saturation of manufacturing with information and communication technologies, most collected data is only used for its original purpose. This limits the potential of this data because it is almost never analysed for more complex information (Friedemann et al. 2016).

The process of digitalization has created opportunities for new products, technologies, and processes (Chinoracký & Čorejová 2019). An environment is needed that allows the integration of the systems forming a System-of-Systems (SoS). The vision of the overall contribution from the research community in manufacturing and logistics systems, over the next few years, is to bring together researchers and practitioners presenting and discussing topics in modern manufacturing modeling, management and control in the emerging field of Industry 4.0-based resilient and innovative production SoS and supply networks (Panetto et al. 2019).

Tokody (2018) examed whether there is any general or formalized technology that could be equally used in various fields, and which could help to achieve digitalization in everyday life, also in the industry or in manufacturing. Mourtzis et al. (2018) suggested an Open Platform Communications – Unified Architecture (OPC-UA) communications standard to provide a macroscopic and microscopic view of machine shops, towards the Machine Shop 4.0 concept. The proposed system is validated in a Laboratory case study. Dachs et al. (2019) investigated the relationship between the location of production activities and digital manufacturing technologies, also known as Industry 4.0 (I4.0). I argue that I4.0 supports backshoring because it provides higher productivity and flexibility which offers an incentive for firms.

Jaatun (2016) focuses on the role of accountability within information management, particularly in cloud computing contexts. The key to this notion is that an accountable Cloud Provider must demonstrate both willingness and capacity for being a responsible steward of other people's data.

Biagi & Falk (2017) presents new empirical evidence regarding the impact of ICT/e-commerce activities on labor demand. The data is based on new and unique data for 10 European countries for the period 2002–2010. A key feature of the empirical analysis is the use of several types of advanced ICT activities, such as enterprise resource planning (ERP) systems, mobile internet access, and e-commerce practices. The main result of the study is that the increase in ICT/e-commerce activities over time has not led to a decline in jobs. This holds true for both manufacturing and service industries, as well as for SMEs and large firms. For ERP systems and websites, there is some evidence of positive effects.

According to Pacheco et al (2019) the significance of small companies to the global economy and their intrinsic difficulties. The main barriers involving the transition towards Sustainable Product-Service Systems in manufacturing Small and Medium-sized Enterprises as well as the strategies to overcome them. Findings reveal that internal barriers associated with intrinsic characteristics of SMEs become still more sensitive during the transition (e.g., limited financial resources, the lack of competences, follower mentality and resistance to change).

Wong et al (2019) investigated the effects of relative advantage, complexity, upper management support, cost, market dynamics, competitive pressure and regulatory support on blockchain adoption for operations and supply chain management among Small-Medium Enterprises (SMEs) in Malaysia.

Currently, the agri-food sector takes advantage of modern machinery, tools, and emerging information and communication technologies (ICTs) that consider the Internet of Things (IoT) capabilities. These implementations have given way to a new era of agri-food production called ‘Agri-Food 4.0’, where automation, connectivity, digitalization, the use of renewable energies and the efficient use of resources are predominant in this sector (Miranda et al. 2019).

Manufacturing digitalization and the growth of big data promises to foster more responsive supply chains and to close gaps between manufacturers and consumers, leading to highly-connected manufacturing operations, mass customization and more sustainable production (Zaki et al. 2017).

An elaborate business model is the fundament of every firm as it describes the manner of functioning. In the course of the new developments of digitization technologies such as Big Data and Data Science, business models and processes have changed tremendously and new business models pop up as never before. Although theory and practice show increased interest in the potentials of new business models through digitization, prior research has not demonstrated yet the main driver of its significant use (Härting et al. 2018). Cisco worked with IDC to develop a five-stage Digital Network Readiness Model to help organizations envision a clearer path to a network that can support all their digital aspirations (Greene et al. 2017).

Several authors have examined some sectors and segments of the Hungarian agri-food sector. Botos et al (2018) investigated the use of ICT in a preliminary study. Füzesi & Herdon (2010) and Herdon & Füzesi (2011) examined product tracking implementations and quality assurance solutions for information systems in the meat industry. Herdon et al (2012) presented a prototype of a Digital Business Ecosystem for SMEs.

2.2. Measuring digitization

As the digital transformation spreads across every sector and affects every aspect of society, measuring its distinct features and dynamics will become increasingly challenging. New approaches will be needed – and the digital tools and footprints created by digital activities can form part of the solution. The digital transformation is also being felt across all dimensions of data production and use. For example, qualitative information is increasingly becoming a source of quantitative evidence.

Several international organizations are contributing to the measurement of the digital transformation through initiatives, some of which are described in the G20 Toolkit for Measuring the Digital Economy. These include, but are not limited to, work on key ICT indicators within the Partnership on Measuring ICT for Development led by the ITU, UNCTAD and the United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Statistics (UIS). The OECD works closely with several of these organizations, including the World Trade Organisation (WTO) on the issue of measuring digital trade, and the IMF on measuring the implications of the digital economy for macroeconomic statistics (OECD 2019). Kotarba (2017) covers an analysis of metrics used to measure digitalization activities.

Readiness is a developmental stage that describes inclination, willingness, and preparedness to perform an action. Meanwhile, digital is defined as the device and application of digital technology. Thus, digital readiness is defined as inclination and willingness to switch to and adopt digital technology and the readiness to create new innovative opportunities by using this technology in order to bring an individual, organization, industry, and country to achieve their goals faster and with greater results.

There is a need to develop indicators that can show how far businesses in different sectors have progressed in the digital transformation journey, i.e. how digitally mature they are. The digital maturity indicator combines four components to provide a comprehensive picture of how digitally mature Swedish companies are (OECD 2017). The four components are systems for enterprise resource planning (ERP), systems for customer resource management (CRM), social media, market and integration (systems for e-invoice, e-sales, and supply chain).

According to the PMG study, five domains define your digital business aptitude. First, they have a clear vision of what digital means to their business and they have embedded their strategy with digital DNA. Second, a critical mass of digital talent, third, digitized core internal processes, fourth, flexible and agile, lastly, they have re-designed theirs.

Digital Economy Metrics is commonly viewed that the term “digital economy” was introduced by Don Tapscott in his publication: The digital economy: promise and peril in the age of networked intelligence (Tapscott, 1997). Digital Density Index (DDI) is developed by Oxford Economics and Accenture jointly a Digital Density Index (DDI) measuring how digital technologies impact economic growth (Macchi et al. 2015). Digital Economy and Society Index (DESI) is developed within the framework of ”Europe 2020 Strategy”, the European Commission introduced a performance measurement system to track the evolution of the EU member states in digital competitiveness (European Commission, 2016). Digital Society Metrics can be described as a society in which the usage of ICT is common across demographic parameters of the population. Digital citizens function in the digital economy using the available digital public and commercial infrastructure for conducting life activities.

Digital Industry Metrics for the purpose of this paper, the term “digital industry” is defined as the application of digitalization in any type of industry. It is, therefore, not limited to the ICT/new technology sector that produces digital solutions, but it covers all manufacturing or service delivery where such digital solutions are used. Digital Enterprise Metrics is for the level of a single enterprise digitalization can be measured with industry metrics presented in the previous section. However, there is a large additional measurement area that is not covered explicitly by the IDI. Additional KPIs describe the status and performance of eCommerce and digital customer dialog in an enterprise.

Ruiz-Rodríguez et al. (2018) constructed a synthetic index of digital development (Enterprise Digital Development Index –EDDI-) and analyzed the countries in the EU and Spanish regions. The variables of the index come from the “Community survey on ICT usage and e-commerce in enterprises” of Eurostat. Many companies fail to understand that the use of digital technologies requires different preparations. One of them is the requirement of an adequate level of digital readiness provides a viewpoint of the adequacy of existing models. Nasution et al (2018) describe future directions to evaluate the readiness of companies.

3. Material and Methods

The research goal defined the research method. The use of information and decision support tools to support management was at the heart of the assessment of the situation of the digitalization of food business enterprises. The research (in 2019) examined the role and relationship of integrated enterprise information systems (ERP) or decision support, business analysis tools (BI), which are used independently of one another in the Food Industry 4.0 technology change. The preparation and implementation phases of the survey are shown in Figure 1. The anonymous online questionnaire was filled by Hungarian food industry companies, which contain 6 question groups with 44 questions. The question groups focused on the main IT topics in the digitalization transformation. The survey was only possible with a questionnaire survey. The preparation and implementation phases of the questionnaire were as follows.

Questions groups were:

- General enterprise information
- IT infrastructure and Internet usage
- Enterprise information systems (ERP)
- Using Management Information Systems (CIS)
- Using Business Analytics (BI) Tools
- Industry 4.0

Within the above question groups the following types were used: Yes/No, List (radio), Multiple choices, Array (5 point choice), Date/Time, Array (Yes/No/Uncertain), Array dual scale, Short free text.

The research tools were LimeSurvey, Mailing List Server Publishing on Web sites, Excel, Power BI (Desktop, Publishing Server to distribute the results).

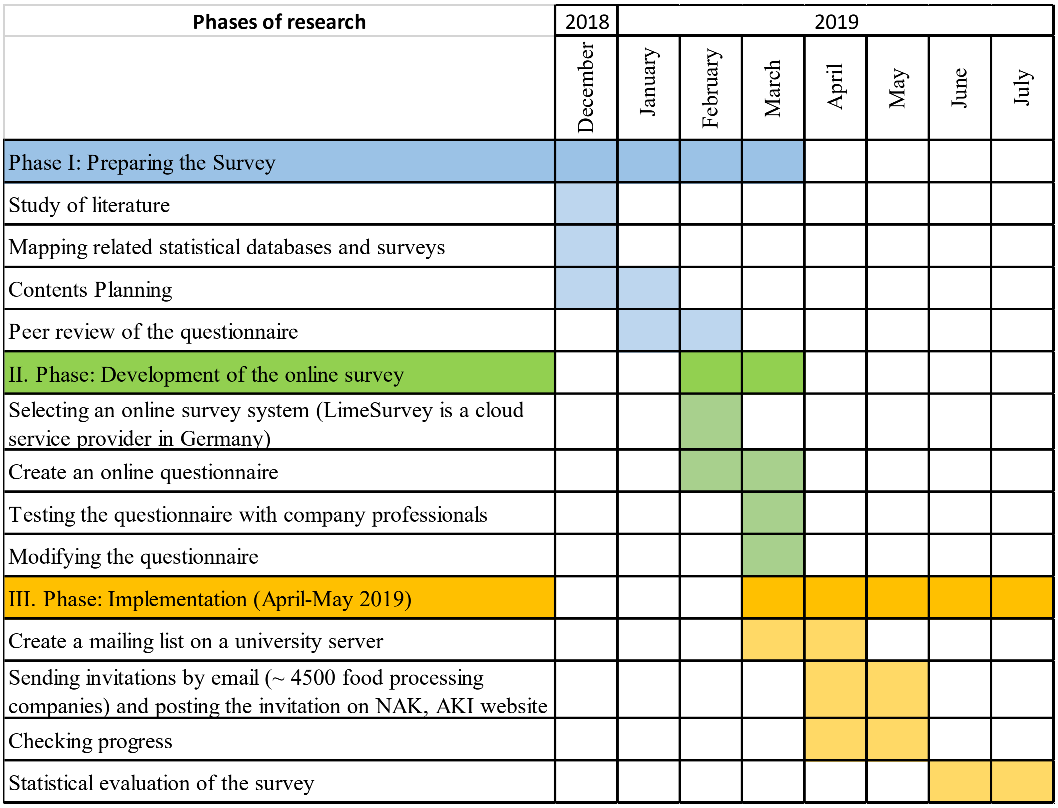


Figure 1. The research phases

Limesurvey is a professional questionnaire engine with a pre-installed environment and database. Once one entered the system, he can create the questionnaire he wants to do by first defining the frame of the questionnaire, then the question groups, and the questions themselves (Figure 2). The used survey tools and evaluation flow are shown in Figure 3.

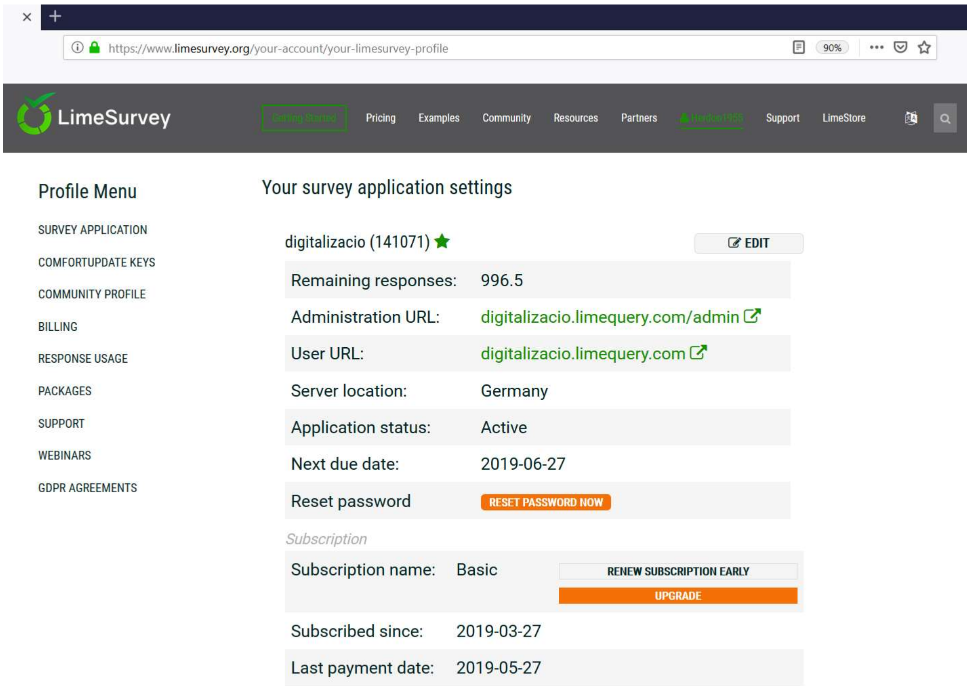


Figure 2. Lime Survey settings

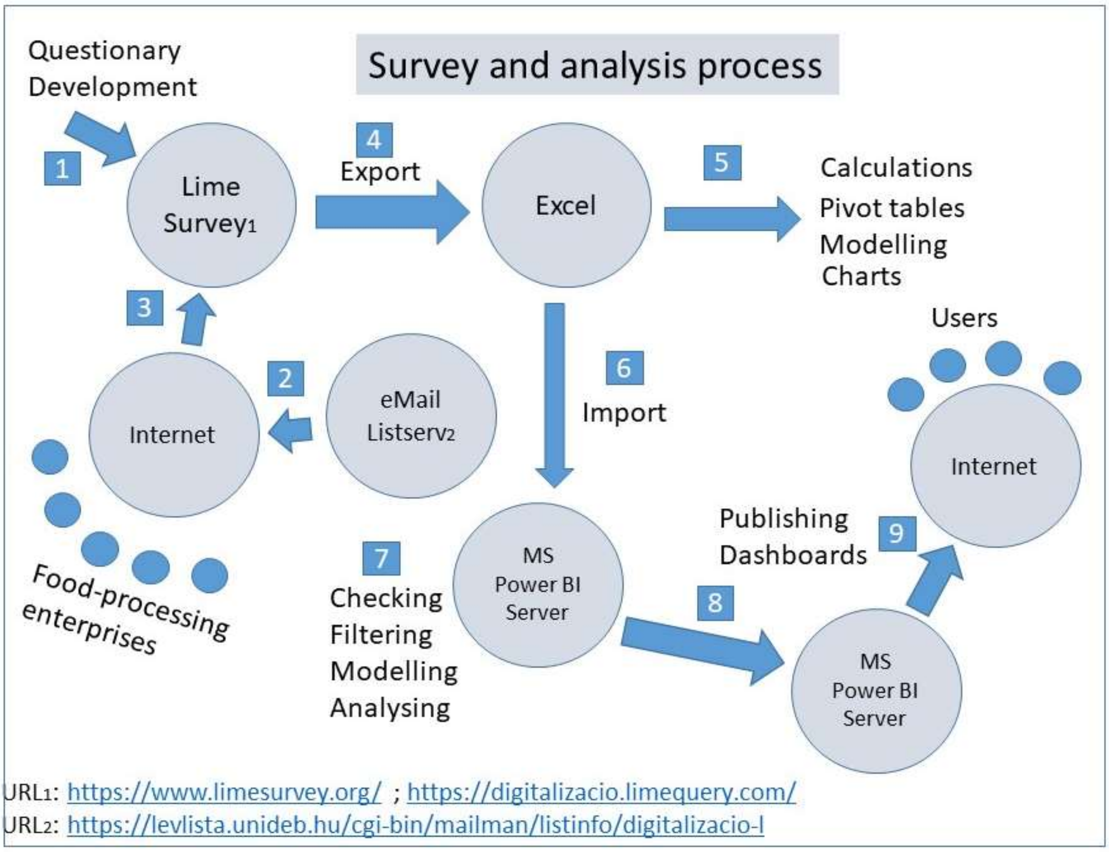


Figure 3. The used tools and processes

The surveyed population consisted of Hungarian food processing companies. During the examination, I tried to obtain a representative sample. Because the sample was sent to all enterprises from the entire population, respondents were randomly selected, so each sampling unit had the same chance of responding (to the sample).

The general requirement for sample selection, as described above, is that the sample is representative of the population from which it was taken, that is, to reflect well the composition and characteristics of the population. So I aimed for a representative sample with the same characteristics as the primary population. The aim was to ensure that the sample population reflects the characteristics of the sample population. Table 1 shows the sample size and degree of representativeness.

Some feature of the respondents:

- Sent requests via e-Mail Listserv: 4000; Access the on-line Questionnaire: 390; The number of the filled questionnaire: 205
- Number of the fillers by position (this question was optional): Top manager 71, Middle manager 39, Administrative worker 30, IT professional, 15 Other 8, Operational leader 7.

Table 1. The survey sample by branches.

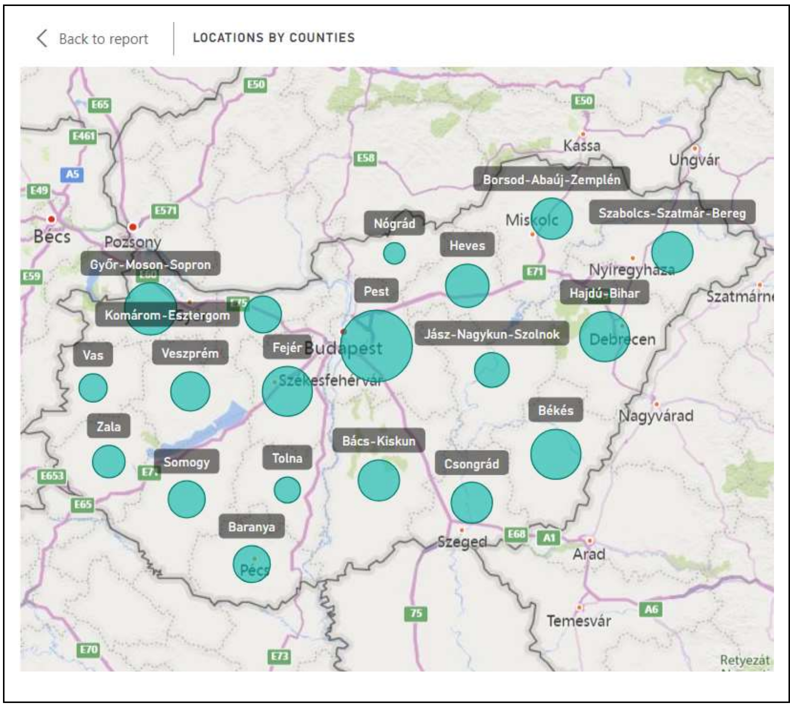
Branch code	Branch	Number of enterprises B	Number in the sample A	Percentage B/A
104	Manufacture of vegetable and ani	60	7	12%
109	Production of feed	100	12	12%
105	Milk Processing	100	11	11%
106	Manufacture of grain mill product	100	10	10%
101	Meat Industry	500	38	8%
108	Manufacture of other food produ	600	44	7%
103	Processing of fruits and vegetable	500	27	5%
107	Manufacture of bakery and farina	1200	36	3%
110	Beverage industry	1400	33	2%

Considering the size distribution of respondents by size category (Table 2), three-quarters of micro enterprises have sales below HUF 50 million, while the small part of small enterprises are between HUF 100 million and HUF 1 billion, while in the case of medium enterprises Part One has net annual sales of more than HUF 1 billion. 93% of large companies have sales of over 1 billion HUF.

Table 2. The sample by Annual Revenue / Company size.

Annual net sales / Company size	Micro-	Small-	Middle-	Large-	All enterprises
<2 million HUF	9	1	0	0	10
2-10 million HUF	17	0	1	0	18
10 and 50 million HUF	22	3	1	0	26
50-100 million HUF	8	6	1	0	15
100-200 million HUF	3	21	2	0	26
200-500 million HUF	6	22	1	0	29
500 million HUF and 1 billion HUF	0	21	4	1	26
>1 billion HUF	0	13	28	14	55
Altogether	**65**	**87**	**38**	**15**	**205**

The number of food-processing companies can be seen in Figure 4. The map of Hungary shows the geographical distribution of the company. The circle and the numbers on the table show the numbers of the enterprises.



Back to report | THE ENTERPRISE NUMBER

Location (County)	Count of ID
Pest	25
Győr-Moson-Sopron	16
Békés	15
Fejér	15
Hajdú-Bihar	15
Heves	12
Bács-Kiskun	11
Borsod-Abaúj-Zemplén	11
Csongrád	11
Szabolcs-Szatmár-Bereg	11
Veszprém	10
Baranya	9
Komárom-Esztergom	9
Somogy	9
Jász-Nagykun-Szolnok	8
Zala	7
Vas	5
Tolna	4
Nógrád	2
Total	**205**

Figure 4. Geographical locations of the companies in the Sample (N=205) (Created by Power BI)

4. Results

My research focused on the technological aspects of digitalization. More methods developed to measure the digital readiness on different levels from global, macroeconomic, industry end enterprise levels. The following levels of digitalization and methods with samples of relevant metrics are covered.

4.1. The structures of the developed model to measure the readiness

The Digital Economy and Society Index (DESI) measures the progress of EU countries towards a digital economy and society. As such, it brings together a set of relevant indicators on Europe's current digital policy mix(European Comission 2019). The DESI has a three-layer structure. It is composed of five principal dimensions, each divided in a set of sub-dimensions, which are in turn composed by individual indicators. The dimensions are 1 Connectivity, 2 Human capital, 3 Use of internet services, 4 Integration of digital technology, 5 Digital public services.

Some dimensions, sub-dimensions and individual indicators are more relevant than others, and for such a reason they were given higher weight in the computation of the final index score for each country.

Table 3 presents the overall weights attributed to the main DESI dimensions, which reflect the EU's digital policy priorities.

Table 3. Structure of the DESI

DESI Dimension	Weight
1 Connectivity	25%
2 Human Capital	25%
3 Use of Internet Services	15%
4 Integration of Digital Technology	20%
5 Digital Public Services	15%

The questions (and their sub-questions) what contains the data and used to create a DE-FS structure were the following in the survey (Table 4):

Table 4. The questions (* with their subquestions in the questionnaire).

ID of Question	Question
8	Does your company have a local (internal) computer network? *
9	What type of Internet access does the company have? *
10	What is the Internet used for business purposes? *
12	Does your business have a website? *
14	Does your company have a subscription to any of the following cloud-based services? Free services are not included! *
15	Does your company have an integrated or independent corporate governance information system? *
21	Please indicate which activities are supported (or in which area of the company operation) the corporate governance information system is regularly used. *
24	Does your company have a management information system? *
29	What are management information systems used for? *
31	Do you use business intelligence tools/applications? *
32	Evaluate which business analytics technologies are relevant to your company! (1 = Not relevant, 5 = Very relevant) *

The information technological aspect for companies links to the connectivity, use of internet services and the integration of digital technology. I used these aspects to evaluate the companies and branches for the Hungarian food-processing companies by the created Digital Enterprise Index for Food-processing Sector (DEI-FS). The DEI-FS has a three-layer structure as depicted in Table 5. It is composed of 3 principal technological dimensions, each divided in a set of sub-dimensions, which are in turn composed by individual indicators.

Table 5. The structure of Digital Enterprise Index for Food-processing Sector DEI-FS.

<table>
<tr><th>Dimension</th><th>Weight</th><th>Sub-dimension</th><th>Weight</th><th>Indicator</th><th>Weight</th></tr>
<tr><td rowspan="3">1 Connectivity</td><td rowspan="3">0.42</td><td>1a Wired Broadband Internet</td><td>0.15</td><td>Fixed Broadband Internet Access</td><td>1.00</td></tr>
<tr><td>1b Mobile Broadband Internet</td><td>0.35</td><td>Wireless Internet Access</td><td>1.00</td></tr>
<tr><td>1c Internal local computer network</td><td>0.50</td><td>Operating an Internal LAN</td><td>1.00</td></tr>
<tr><td rowspan="2">2 Use of Internet Services</td><td rowspan="2">0.25</td><td rowspan="2">2a Transactions</td><td rowspan="2">1.00</td><td>Purchasing Products, Services (Electronic Purchasing)</td><td>0.50</td></tr>
<tr><td>Sale of products, services (e-commerce)</td><td>0.50</td></tr>
<tr><td rowspan="3">3 Integration of Digital Technology</td><td rowspan="3">0.33</td><td rowspan="3">3a Business Digitization</td><td rowspan="3">0.15</td><td>Electronic Information Sharing (Advertising / Marketing)</td><td>0.40</td></tr>
<tr><td>Website Operation</td><td>0.30</td></tr>
<tr><td>Using a Subscription Cloud Service</td><td>0.30</td></tr>
</table>

		3b E-commerce	0.10	Electronic Procurement / E-Commerce	1.00
		3c ERP	0.25	Using Enterprise Management Information System (ERP)	0.55
				It is used regularly in at least 10 areas of company operations	0.15
				Used in at least 5 but less than 10 areas of enterprise operations	0.15
				Used in less than 5 main functional units of enterprise operations	0.15
		3d Executive Information System	0.25	Using a Management Information System	0.55
				Analyzing aggregate data	0.15
				Making plan-fact comparisons	0.15
				Time series analysis	0.15
		3e BI Tools / Applications	0.25	Using Business Intelligence Tools / Applications	0.50
				Using BI/Analytical functions*	0.50

*Using BI/Analytical functions: This indicator consists of the following sub-indicators (using the following tools): Analysing; Monitoring, controlling; Forecasts; Graphic Visual Dashboards; data warehouse; Data Mining Tools; Complex Analysis Technologies; Cloud-Based BI; Using Mobile Based BI with 0,05 weight).

12 questions and their sub-questions related to the indicators which are in Table 3. The weights were determined using DESI and other methods as well as literature references. Of course, it is possible to change the weights in the hierarchical weight structure in the created calculator.

In DEI-FS, the aggregation of indicators into sub-dimensions, of sub-dimensions into dimensions, and of dimensions into the overall index was performed from the bottom up using simple weighted arithmetic averages following the structure of the index (Table 5).

As an example, the top-level DEI-FS score for enterprises, branches, enterprise groups E was calculated using the formula: *DEI-FS* (E)= $Connectivity$(E)$*0.25$ + $Use\ of\ Internet\ Services$ (E)$*0.15$ + $Integration\ of\ Digital\ Technology(E)*0.2$

Where Con(E) is the score obtained by enterprise E in the Connectivity dimension, and so on for the remaining dimensions in the formula.

4.2. The rank by enterprise size

Figure 5. below shows the level of development of enterprises within the enterprise size categories with the modified "DEI-FS" index. We can observe that as the size of the company increases, the level of development also increases, so it can be said that large enterprises (68%) and medium-sized enterprises (52%) are the most digitally advanced. The average level of development of all enterprises is 41%. Below this is the indicator for micro-enterprises (29%), while the average for small enterprises is the average for all enterprises.

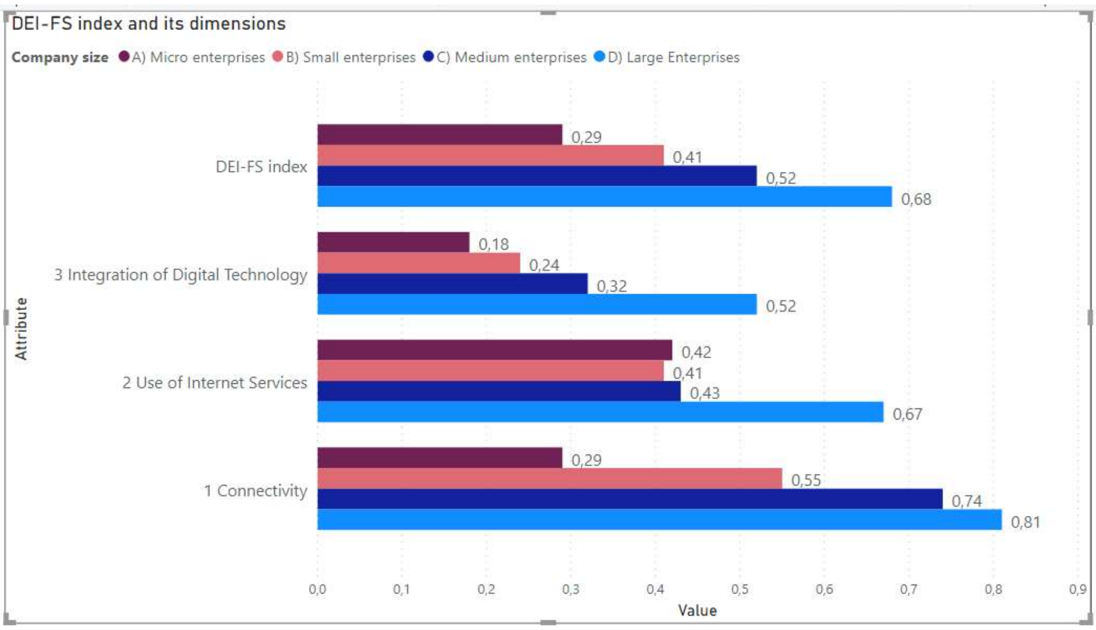


Figure 5. The digital development level by " DEI-FS " index within the size category (N=205)

4.3. The rank of the braches and regions

Examining the level of development by sector, Figure 6 shows that the most advanced sectors are the production of oil, the production of other foodstuffs, the production of mill products and the processing of vegetables and fruits. The most backward in this respect (not counting the fish processing and tobacco production sectors, where 1-2 fillings came) are dairy, bakery, pasta, and feed production.

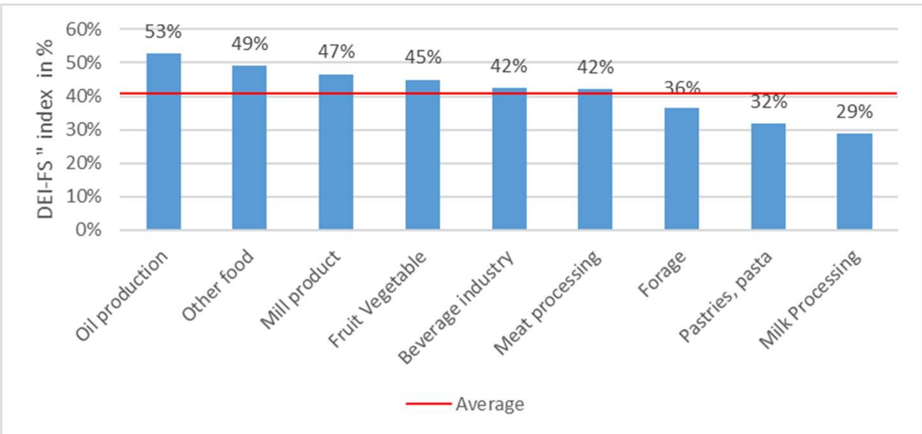


Figure 6. The digital development level of branches by " DEI-FS " index (N=205)

The three tools of the analysis were the Lime Survey (this was used for preliminary statistical analysis of the survey data), Microsoft Excel (partly for data cleansing and preliminary analysis, transformations), the third tool was the Power BI (data modeling, analysis, reports dashboard- ok). One of the tiles in the industry comparison report (Figure 7) shows the values of the DEI-FS index dimensions by industry. The complex index and your indicators have been studied in detail, and a number of qualitative characteristics have been analyzed with the Business Intelligence Tool.

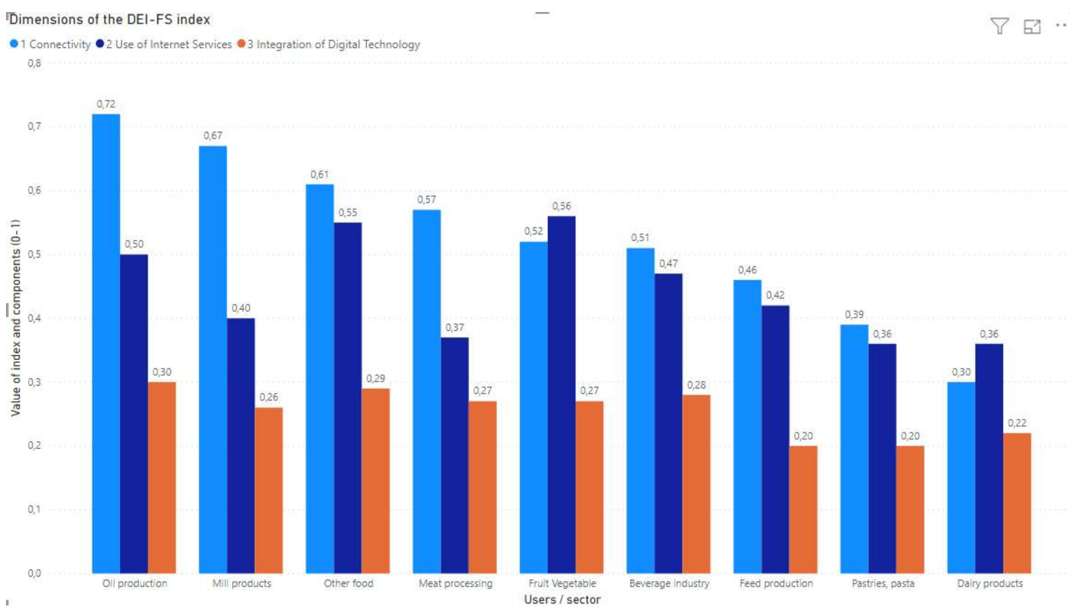
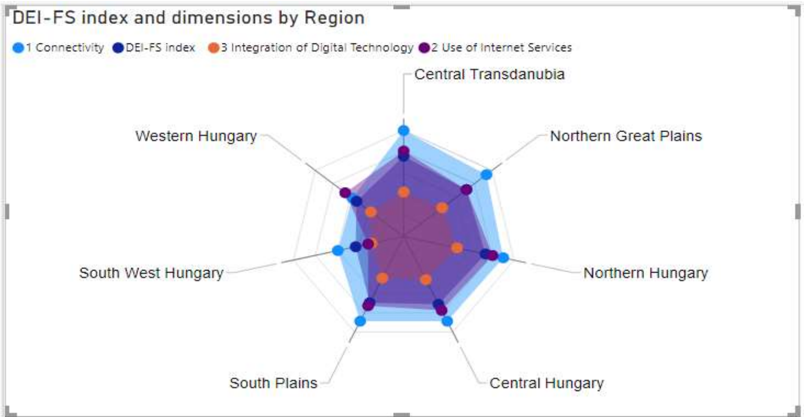


Figure 7. The dimensions of the DEI-FS index by sector branches (N=205) (Made with Power BI)

The development by regions is shown in Figure 8 which shows that the most backward regions are Western and Southern Transdanubia. However, the other regions are not much more developed than average.



Region	1 Connectivity	2 Use of Internet Services	3 Integration of Digital Technology	DEI-FS index
Central Transdanubia	0,62	0,50	0,26	0,47
Northern Hungary	0,56	0,50	0,30	0,46
Central Hungary	0,55	0,48	0,28	0,44
Northern Great Plains	0,58	0,44	0,27	0,44
South Plains	0,55	0,45	0,27	0,43
Western Hungary	0,36	0,41	0,23	0,33
South West Hungary	0,37	0,20	0,18	0,27

Figure 8. The digital development level of regions by " DEI-FS " index (N=205)) (Made with Power BI)

5. Conclusion

The complex rating index shows differences in development between company size, industry and country regions. The survey is considered representative, suggesting that it may be suitable for self-evaluation of individual companies or for the use of a calculator simulator in which individual weights can be set at intervals. On the other hand, the developed system may be suitable for analyzing progress over time, repeating the survey every year, and publishing the Dashboard interactively with the Power BI or other BI systems.

Acknowledgments

In particular, I thank the Leaders and Employees of the Hungarian Research Institute of Agriculture Economics, the Hungarian Chamber of Agriculture and the Hungarian Federation of Food Manufacturers for supporting the research work.

References

Biagi F, Falk M. 2017. The impact of ICT and e-commerce on employment in Europe, Journal of Policy Modeling, Volume 39, Issue 1, Pages 1-18. https://doi.org/10.1016/j.jpolmod.2016.12.004.

Botos Sz, Felföldi J, Várallyai L, Péntek Á, Szilágyi R. 2018. Analysis the advanced ICT usage of the Hungarian SME sector for preparing a domestic agri-food research APSTRACT - Applied Studies in Agribusiness and Commerce 11 : 3-4 pp. 147-154. https://doi.org/10.19041/APSTRACT/2017/3-4/20

Botos Sz, Herdon M, Várallyai L. 2015. Readiness for future internet services in rural areas. In. Procedia Economics and Finance 19. 2015. p. 383-390. https://doi.org/10.1016/S2212-5671(15)00039-8

Chinoracký R, Čorejová T. 2019. Impact of Digital Technologies on Labor Market and the Transport Sector, Transportation Research Procedia, Volume 40, 2019 Pages 994-1001, https://doi.org/10.1016/j.trpro.2019.07.139.

Dachs B, Kinkel S, Jäger A. 2019. Bringing it all back home? Backshoring of manufacturing activities and the adoption of Industry 4.0 technologies, Journal of World Business, Volume 54, Issue 6, 2019 101017 https://doi.org/10.1016/j.jwb.2019.101017.

European Commission. 2016. Digital Single Market. Digital Economy & Society, February 2016. [online] Available at: https://ec.europa. eu/digital-single-market/en/europe-2020-str [Accessed at: 10.12.2019]

Friedemann M, Trapp T.U, Stoldt J, Langer T, Putz M. 2016. A Framework for Information-driven Manufacturing, Procedia CIRP, Volume 57, 2016 Pages 38-43, https://doi.org/10.1016/j.procir.2016.11.008.

Füzesi I, Herdon M. 2010. Food Tracing and Interoperability of Information Systems in the Hungarian Meat Industry AGRIS ON-LINE Papers in Economics and Informatics 2 : 2 pp. 39-48.

Greene N, Parker P, Perry R. 2017. Is Your Network Ready for Digital Transformation? January 2017, IDC White Paper Sponsored by Cisco

Härting R-H, Reichstein C, Schad, M. 2018. Potentials of Digital Business Models – Empirical investigation of data driven impacts in industry, Procedia Computer Science, Volume 126, 2018 Pages 1495-1506. https://doi.org/10.1016/j.procs.2018.08.121.

Herdon M, Füzesi I. 2011. Information Technologies in Quality Management Systems of Meat Product Chains In: Zacharoula, Andreopoulou; Basil, Manos; Nico, Polman; Davide, Viaggi (szerk.) Agricultural and environmental informatics, governance and management: emerging research applications New York (NY), Amerikai Egyesült Államok : IGI Global, (2011) pp. 207-226. https://doi.org/10.4018/978-1-60960-621-3.ch012

Herdon M, Raffai M, Péntek Á, Rózsa T. (2010) Digital Business Ecosystem Tools as Interoperability Drivers. In: Bernus P., Doumeingts G., Fox M. (eds) Enterprise Architecture, Integration and Interoperability. EAI2N 2010. IFIP Advances in Information and Communication Technology, vol 326. Springer, Berlin, Heidelberg https://doi.org/10.1007/978-3-642-15509-3_11

Herdon M, Várallyai L. Péntek Á. (2012) Digital business ecosystem prototyping for SMEs. Journal of Systems and Information Technology 14 : 4 pp. 286-301. https://doi.org/10.1108/13287261211279026

Jaatun M.G, Pearson S, Gittler F, Leenes R, Niezen M. 2016. Enhancing accountability in the cloud, International Journal of Information Management, 0 2016. https://doi.org/10.1016/j.ijinfomgt.2016.03.004.

Kotarba M. 2017. Measuring digitalization - key metrics, Foundations of Management, Vol. 9, ISSN 2080-7279. https://doi.org/10.1515/fman-2017-0010

KPMG LLP. 2013. "2013 Food and Beverage Industry Outlook Survey", 2013, URL: http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/food-beverage-outlook-survey-2013.pdf

Macchi, M., Berthon, B., Robinson, M., 2015. Accenture Digital Density Index – Guiding Digital Transformations. [online] Available at: https://www.scribd.com/document/377691886/Accenture-Digital-Density-Index-Guiding-Digital-Transformation [Accessed at 10.12.2019]

Miranda J, Ponce P, Molina A, Wright P. 2019. Sensing, smart and sustainable technologies for Agri-Food 4.0, Computers in Industry, Volume 108, Pages 21-36. https://doi.org/10.1016/j.compind.2019.02.002

Mourtzis D, Milas N, Athinaios N. 2018. Towards Machine Shop 4.0: A General Machine Model for CNC machine-tools through OPC-UA, Procedia CIRP, Volume 78, 2018 Pages 301-306. https://doi.org/10.1016/j.procir.2018.09.045

Nasution R.A, Rusnandi L.S.L, Qodariah E, Arnita D, and Windasari N.A. The Evaluation of Digital Readiness Concept: Existing Models and Future Directions. Asian Journal of Technology Management, vol. 11, no. 2, 2018, pp. 94-117, doi:10.12695/ajtm.2018.11.2.3

OECD. 2017. Measuring Digital Maturity in Firms. Directorate for Science, Technology and Innovation Committee on Digital Economy Policy. DSTI/CDEP/MADE(2017)3

OECD. 2019. Measuring the Digital Transformation www.oecd.org/about/publishing/corrigenda.htm

Pacheco D.A.J Caten C.S, Jung C.F, Sassanelli C, Terzi S. 2019. Overcoming barriers towards Sustainable Product-Service Systems in Small and Medium-sized enterprises: State of the art and a novel Decision Matrix, Journal of Cleaner Production, Volume 222, 2019 Pages 903-921. https://doi.org/10.1016/j.jclepro.2019.01.152

Panetto H, Iung B, Ivanov D, Weichhart G, Wang X. 2019. Challenges for the cyber-physical manufacturing enterprises of the future, Annual Reviews in Control, Volume 47, 2019 Pages 200-213. https://doi.org/10.1016/j.arcontrol.2019.02.002

Ruiz-Rodríguez F, Lucendo-Monedero A.L, González-Relaño R. 2018. Measurement and characterisation of the Digital Divide of Spanish regions at enterprise level. A comparative analysis with the European context, Telecommunications Policy, Volume 42, Issue 3, 2018 Pages 187-211. https://doi.org/10.1016/j.telpol.2017.11.007

Singh R Kr, Luthra S, Mangla S K, Uniyal S. 2019. Applications of information and communication technology for sustainable growth of SMEs in India food industry. Resources, Conservation and Recycling, Volume 147, 2019, Pages 10-18. https://doi.org/10.1016/j.resconrec.2019.04.014

Tapscott D. 1997. The digital economy: promise and peril in the age of networked intelligence, New York: McGraw-Hill, ISBN 0-07-063342-8.

Tokody T. 2018. Digitising the European industry - holonic systems approach, Procedia Manufacturing, Volume 22, 2018 Pages 1015-1022. https://doi.org/10.1016/j.promfg.2018.03.144

Wolfert S, Verdouw CN, Verloop CM, Beulens AJM. 2010. Organizing information integration in agri-food – A method based on a service-oriented architecture and living lab approach. Computers and Electronics in Agriculture, Vol. 70. pp. 389-405. https://doi.org/10.1016/j.compag.2009.07.015

Wong L-W, Leong L-Y, Hew J-J, Tan G.W-H, Ooi K-B. 2019. Time to seize the digital evolution: Adoption of blockchain in operations and supply chain management among Malaysian SMEs, International Journal of Information Management, 0 2019. https://doi.org/10.1016/j.ijinfomgt.2019.08.005.

Zaki M, Theodoulidis B, Shapira P, Neely A, Surekli E. 2017. The Role of Big Data to Facilitate Redistributed Manufacturing Using a Co-creation Lens: Patterns from Consumer Goods, Procedia CIRP, Volume 63, 2017 Pages 680-685. https://doi.org/10.1016/j.procir.2017.03.350.

Zhang W. 2012. Study on Base Management Pattern of Food Producing Enterprise, Physics Procedia, Volume 24, Part B, 2012, Pages 1108-1113. https://doi.org/10.1016/j.phpro.2012.02.165