DOI: 10.34101/ACTAAGRAR/1/8587

Analysis of sweet corn nutritional values using multivariate statistical methods

László Huzsvai[1] – Péter Fejér[2] – Árpád Illés[2] – Csaba Bojtor[2] – Csilla Bojté[2] – Éva Horváth[2] – Cintia Demeter[2]

[1]Faculty of Economics and Business. Institute of Sectoral Economics and Methodology. Department of Research Methodology and Statistics, University of Debrecen

[2]Faculty of Agricultural and Food Sciences and Environmental Management, Institute of Land Utilisation, Technology and Regional Development, Faculty of Agricultural and Food Sciences and Environmental Management, University of Debrecen

szintia.demeter@gmail.com

SUMMARY

Processing large amounts of data provided by automated analytical equipment requires carefulness. Most mathematical and statistical methods have strict application conditions. Most of these methods are based on eigenvalue calculations and require variables to be correlated in groups. If this condition is not met, the most popular multivariate methods cannot be used. The best procedure for such testing is the Kaiser-Meyer-Olkin test for Sampling Adequacy. Two databases were examined using the KMO test. One of them resulted from the sweet corn measured in the scone of the study, while the other from the 1979 book of János Sváb. For both databases, MSA (measures sampling adequacy) was well below the critical value, thus they are not suitable e.g. for principal component analysis. In both databases, the values of the partial correlation coefficients were much higher than Pearson's correlation coefficients. Often the signs of partial coefficients did not match the signs of linear correlation coefficients. One of the main reasons for this is that the correlation between the variables is non-linear. Another reason is that control variables have a non-linear effect on a given variable. In such cases, classical methods should be disregarded and expert models better suited to the problem should be chosen in order to analyse the correlation system.

Keywords: *multivariate methods; Kaiser – Meyer – Olkin (KMO) Test for Sampling Adequacy; nutritional values of plants*

INTRODUCTION

Sweet corn has been grown on 360–380 thousand hectares globally in recent years. Its main producers are the USA, the European Union and Thailand. Sweet corn production of Hungary is also at the forefront on international level, ahead of France in terms of production area, making it the leading producer in the EU. In recent years (2014–2019), the yield of sweet corn in Hungary varied between 460–515 thousand tons. Most of the production is processed by the canning industry while a smaller proportion by the refrigeration sector. Consumption of processed sweet corn is 1–1.5 kg person^{-1} per year. Consumption of processed sweet corn is 1–1.5 kg person^{-1} per year. Consumption as fresh goods shows a growing tendency. In terms of sweet corn exports, Hungary is the second largest exporter of processed sweet corn after the USA. 95% of the produced goods are exported. The quality of sweet corn hybrids is largely determined by genetic components. The grain quality of sweet corn is highly dependent on its nutritional values and the applied cultivation technology (Brecht et al., 1990; Garwood et al., 1976; Dániel, 1978; Hadi, 2003; Balázs, 2004; Nagy, 2006; Orosz, 2009; Lente, 2012). The nutritional role of sweet corn is significant. It has a protein content similar to that of peas, its fat content is four times higher, while its carbohydrate content is twice as that of peas. Compared to maize feed, its starch content is approximately half and is replaced by sugars, mainly polysaccharides. Super sweet corns have a starch content of less than 10%, but their sucrose content is over 30% (2.5–3% for normal sweet corn, 1.5% for feed maize). Fresh sweet corn is a great source of vitamins C and A. An ear of sweet corn covers more than 10% of the vitamin C demand (DV) of a person and more than 6% of vitamin A. An ear of cooked sweet corn has 4 grams of protein, 1.8 grams of fat, 5.4 grams of sugar, 2.8 grams of fibre and just 113 kcal. A large amount of soluble fibre lowers cholesterol levels. Sweet corn contains less crude fibre that would hinder digestibility, however it is richer in crude fats and vitamins. 100 g of raw grains contain 0.19 mg B1, 0.12 mg B2, 1.64 mg B3 and 9.2 mg of vitamin C (Bajtay, 1979). Obeng et al. (2020) found that the pro vitamin A (PVA) concentrations of sweet corn hybrids differ significantly by genotype. In the scope of research, they concluded that the amount of PVA does not depend on the phenotypic appearance of sweet corn. Ray et al. (2019) measured the oil and protein content of maize hybrids in their studies, and found that protein and oil content were higher in the fertilized group compared to the control group. In the course of their research, they also found that genotype determines the yield and quality of maize. NPK fertilizer applied at higher doses increased the amount of macronutrients in maize grains. The results show that protein concentration can be increased with nitrogen fertilizer. In their research Kang et al. (1986) showed a direct effect between grain filling and the decrease of water content. The chemical composition of the maize grain (starch, sugar content, protein and amino acid quality), plays an important role in the development of water content in maize grains. In their experiments, Calvo-Brenes and O'Hare (2020) measured the carotenoid content of harvested sweet corn at different time intervals. It was found that sweet corn could be stored at -80 °C for 3 months and at 4 °C for 15 days so that the carotenoid content of the harvested sweet corn did not change. Carotenoids are important for the human body, they have a protective

DOI: 10.34101/ACTAAGRAR/1/8587

effect on eye health and play a preventive role in cardiovascular diseases as well as in the case of certain forms of cancer (Burt el al., 2011; Ma and Lin, 2010; Rosen and Hu, 2018).

Currently, more and more modern automatic analytical equipment is used to determine the nutritional values of plants; components are most often expressed given as a percentage of dry matter. If all possible nutritional values are measured, the sum results in 100%. In the case of classic multivariate statistical methods, this often causes problems, as the obtained results lead to an axiom or the conditions for the applying the methods are not met. The most commonly applied real multivariate statistical methods are based on eigenvalue calculations. Examples include principal component analysis (PCA), factor analysis (FA), and discriminant analysis (DA). For these methods, it is required that variables are group-correlated with each other and must be higher than the values of partial correlation. The best procedure for this is the Kaiser-Mayer-Olkin test (KMO). The test compares the sum of squares of Pearson's correlation coefficients to the sum of squares of partial correlation coefficients. The resulting indicator is MSA (measure of sampling adequacy). The critical value of the test is 0.5. At higher values, the application conditions for PCA and FA are met. However, the value is lower, these useful procedures cannot be applied because the results obtained will not be true. This means that even the opposite of the obtained results might be true. In this case, pairwise correlations are much stronger than multiple correlations. What could be the reason for this? In the present study, this is examined in a sweet corn experiment and based on the example of Sváb (1979).

MATERIALS AND METHODS

For the statistical analysis of the nutritional values of sweet corn hybrids, plants were randomly selected for sampling. In the course of sampling, 78 average samples were formed. The grains on the sweet corn ear were sampled along its entire length, from the base to its apex with a sharp knife. For laboratory tests, samples were transported in liquid nitrogen and stored frozen at -84 °C until processing.

Laboratory methods

In this method, maize samples were ground together with dry ice and then stored at -18 °C in an open container in a freezer until the dry ice sublimed. 0.6 g of the ground sample was weighed into a 50 ml centrifuge tube. 6 ml of 100% ethanol was added, vortexed for 30 seconds and then sonicated in a cooled ultrasonic bath for 5 minutes. 3 mL of 10% NaCl solution and 10 mL of hexane were added and vortexed for 30 seconds, followed by 3 minutes of centrifuging at 5000 rpm until the phases separated. The upper hexane layer was pipetted into an evaporator tube. The hexane extraction was repeated twice more until the lower aqueous-alcoholic phase became colourless. The collected hexane fractions were evaporated to dryness under a stream of nitrogen at room temperature in the dark. 2 mL of MeOH containing 0.1% BHT was added to the evaporated residue. It was dissolved by vortexing and ultrasound, and the solution was filtered through a 0.22 μm syringe filter into an HPLC vial. It was stored in a freezer at -18 °C until HPLC analysis (Kimura et al., 2007).

Determination of ash content was performed according to the MSZ 20501/1-87 standard by means of annealing. The organic matter content of the sample decomposes during annealing and the following incineration, the remainder being the ash content.

For the determination, 1 g of sample was weighed with analytical accuracy (W1) into a pre-annealed incineration crucible with a known weight (W2). It was heated on an electric hot plate for 30 minutes until a dry carbon residue formed and then incinerated in a 550 °C oven for 6 hours. The crucible was then cooled to room temperature in a desiccator and the mass was weighed back (W3). The ash content was calculated using the following formula: crude ash %= (W3-W2)/W1.

The starch content was extracted by the method of Chow and Landhäusser (2004). A 25 mg sample was dissolved three times in 80% ethanol, vortexed for 1 minute and then boiled at 95 °C for 10 minutes. After extraction, the samples were centrifuged at 3500 rpm for 10 minutes. The extraction was repeated until the samples became colourless. This was followed by washing with distilled water and acetone to remove ethanol. After washing, the remaining samples were dried, and then the light absorbance of the samples was measured at 570 nm, and the starch content of the sample was calculated using a formula.

Crude protein content was determined by the Dumas method on a LECO FP528 type device. In the device, the organic matter content of properly shredded and homogenized samples decomposes into CO_2, H_2O and nitrogen oxides in various oxidation states in a stream of oxygen above 800 °C. These are reduced to molecular nitrogen by an inert carrier gas passed through a glowing copper coil, the amount of which can be determined volumetrically with a thermal conductivity detector. Procedure of determination: A given volume of sample was weighed into special tin foils that did not contain nitrogen. These were placed into the device that automatically calculated the crude protein content based on the predefined Kjeldahl conversion factor and the measured N%:

Protein% = N% (Kjeldahl factor)

The applied Kjeldahl factor was 6.25 for all samples

The determination of crude fibre content was performed according to the Wende method using a FIBERTEC SYSTEM M type semi-automatic device. The operating principle is that the sample goes through acidic and then alkaline hydrolysis, washed with distilled water, degreased, then the residue is dried in an oven and reweighed. It is then incinerated and the value corrected for the ash content is the crude fibre content.

For the determination, 0.5 g of the sample was weighed into the filter crucible with analytical accuracy

DOI: 10.34101/ACTAAGRAR/1/8587

(W1). After hydrolysis and degreasing, the crucibles were dried for 2 hours in an oven at 130 °C, cooled in a desiccator and weighed (W2). It was then heated in a 500 °C oven for 5 hours and, after cooling, the mass was weighed again (W3). The crude fibre content was calculated according to the following formula:

Crude fibre % = (W2-W3)/W1

Crude fat content was determined by means of a TECATOR SYSTEM HT-6 semi-automatic device. The principle of the procedure is that the properly processed and homogenized sample is extracted in the extractor with water and a peroxide-free solvent. As the solvent tank is heated, its vapours reach the condenser through a side tube and then condense there and flow through the sleeve containing the sample. During the 20-minute, so-called hot extraction, the sample sleeves are in the solvent tank and then the cold extraction is performed for 40 minutes after their removal from the tank. Then the solvent is distilled and the remaining dissolved material is dried in an oven.

With analytical accuracy, 3g of sample (W1) was weighed into the pre-dried sample holder sleeves. The weight of the empty solvent tanks was also measured (W2). The applied solvent was 30 ml of hexane for each sample. After extraction and removal of hexane, the residue in the solvent tanks was dried at 105 °C for 30 minutes, cooled in a desiccator and weighed again (W3). The crude fat content was calculated according to the following formula:

Crude fat % = (W3-W2)/W1.

Determination of moisture content was performed according to the MSZ 20501/1-87 standard. This is an indirect method of determining moisture content by drying, in which a sample of a given weight is dried to constant weight.

3g of the sample was measured with analytical accuracy (W2) into lids with a previously known weight (W1), which were dried in an oven and cooled in a desiccator. The sample was dried at 70 °C, cooled in a desiccator and its weigh was measured again (W3). The moisture content was calculated according to the following formula:

Moisture % = (W3-W2)/W1.

Statistical methods

Statistical analyses were performed using version 4.0.2 of the R Core Team (2020) statistical software package. The distribution of the variables were plotted on a histogram and the bivariate relationships on a scatterplot. Bivariate linear correlation coefficients (Pearson's r-value) were determined. The degree of closeness was indicated by the size of numbers, while significant values were indicated by asterisks.

The correlation system was characterized by the Kaiser - Mayer - Olkin test. This shows how dependent a variable is on other variables. The index is the quotient of Pearson's coefficients of determination and partial coefficients of determination.

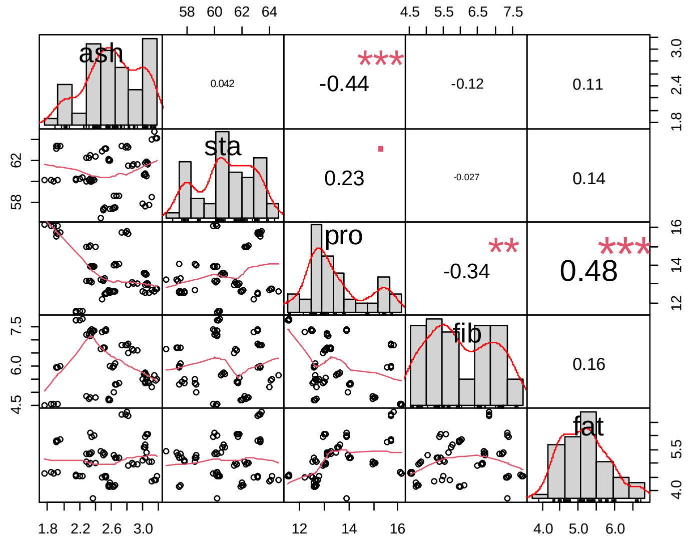
RESULTS AND DISCUSSION

Distribution of the nutritional values of sweet corn and the pairwise correlations are shown in *Figure 1.*

MSA values for the KMO test were determined.

Kaiser-Meyer-Olkin factor adequacy
Overall MSA = 0.19
MSA for each item =
ash sta pro fib fat
0.14 0.29 0.26 0.11 0.18

Figure 1. **Distribution and correlations of the nutritional values of sweet corn. The font size of the numbers expresses the strength of the correlation**



The significance levels are <0.05 (*), <0.01 (**), and <0.001 (***)

Table 1. **Pearson's correlation coefficients**

	ash	sta	pro	fib	fat
ash	1.000	0.042	-0.44	-0.117	0.11
sta	0.042	1.000	0.23	-0.027	0.14
pro	**-0.436**	0.233	1.00	-0.336	0.48
fib	**-0.117**	-0.027	**-0.34**	1.000	0.16
fat	**0.111**	0.142	**0.48**	**0.162**	1.00

KMO values for each variable are well below the critical value of 0.5, which means that the values of partial coefficients are much higher than the values of the Pearson's coefficients (*Table 1*).

Partial coefficients (*Table 2*) are much higher than simple correlation coefficients. Pairwise correlations are much stronger. Correlations are distorted by something or there is a problem with the measurements. Another reason might be that the correlation between the variables is not linear, such as the correlation between ash and fibre content. Based on the partial correlation coefficients (analysing the lower triangle), ash is negatively related to protein and fibre, but positively related to oil content. Protein content and fibre show a negative, protein and oil a positive correlation. Fibre content and oil are also positively related.

The anti-image covariance matrix were determined with the tolerance indices in its main diagonal; this is 1-R^2. R^2 is the value of the multiple linear coefficient of determination. It shows how independent a given variable is from other variables (*Table 3*).

Table 2. **Values of partial correlation coefficients**

	ash	sta	pro	fib	fat
ash	**1.00**	0.23	-0.74	-0.58	0.62
sta	0.23	**1.00**	0.29	0.17	-0.13
pro	**-0.74**	0.29	**1.00**	-0.69	0.76
fib	**-0.58**	0.17	**-0.69**	**1.00**	0.61
fat	**0.62**	-0.13	**0.76**	**0.61**	1.00

Table 3. **Tolerance index**

	ash	sta	pro	fib	fat
1	**0.44**	-0.14	0.25	0.27	-0.26
2	-0.14	**0.89**	-0.14	-0.11	0.08
3	0.25	-0.14	**0.26**	0.25	-0.24
4	0.27	-0.11	0.25	**0.50**	-0.27
5	-0.26	0.08	-0.24	-0.27	**0.40**

The most independent variable is starch content, the value of the tolerance index is very high (89%). Practically, this shows no correlation with any of the variables included in the study. The other four variables are also 26–50% independent of the others. This often occurs when samples are not from homogeneous conditions, e.g. other cultivation technology, different production site, other sampling technique, etc.

Sváb (1979) made a similar mistake in his book "Multivariate Methods in Biometrics" where he described PCA and FA. The two methods were presented on a flour quality database from a winter wheat experiment, but applicability conditions were not examined. The variables included in the study were farinograph value, gluten spread, gluten volume, and protein content. Data were available for 14 winter wheat cultivars. The KMO test result shows deplorable low values.

MSAi = 0.23 0.32 0.24 0.17. According to the values, none of the variables should be included in the principal component analysis. This is because three of the four variables are very strongly correlated in pairs. It is practically the same thing characterized by different metrics. These variables are the farinograph value, gluten spread, and gluten volume. Farinograph and gluten spread have a very close negative correlation (r=-0.77, partial r=-0.96). The farinograph equipment measures the amount of water required to form the dough, the time required to achieve optimal consistency, and the elasticity of the dough, i.e., the resistance required during kneading. The final value is the definition of a given area of the graph. The more gluten spreads, the sooner the dough liquefies. The amount of Farinograph and Gluten are strongly positively correlated according to the partial r (r=-0.13, partial r=0.9). Pearson's product-moment correlation value, on the other hand, shows a low negative value,

DOI: 10.34101/actaagrar/1/8587

thus the condition of linearity is probably not met. According to the r-value, there is no correlation between the two variables, the calculated p-value is 0.7. According to the partial r, the more gluten there is in the flour, the better the flour quality. This seems logical, but apart from quantity, quality matters as well. For the sake of clarity, the variables should be displayed here as well.

Linearity might be compromised even if the effect of the controlling variables is not linear, and this interferes with the value of the partial coefficients. Interpretation of partial r: a linear correlation between two variables if the value of one or more controlling variables is kept fixed. Namely, it is a vale measuring conditional correlation, it is assumed that the value of variable z outside the variables x and y was recorded. If the effect of z is not linear, it depends on where its value is recorded. Suppose that z is in a concave quadratic relationship with x and y. According to that, x and y initially show a directly proportional, later an inversely proportional relationship. This distorts the value of partial r and might even change its sign relative to the actual conditional r-value. This is due to the fact that the condition of linearity is severely violated.

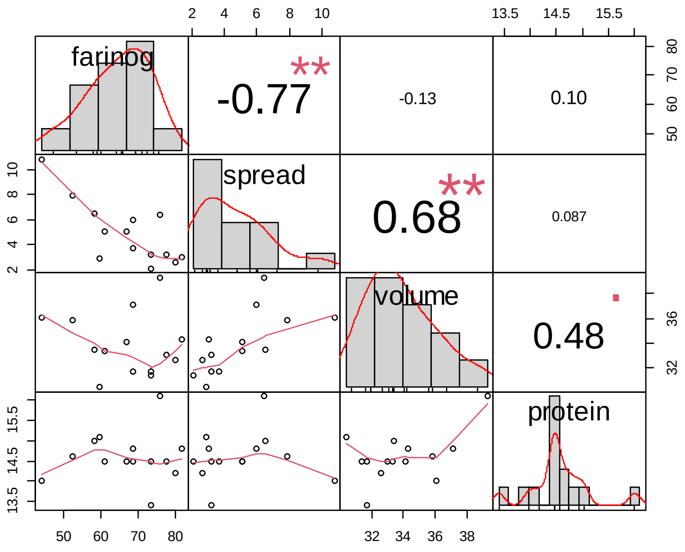
Figure 2 shows the pairwise scatterplots of the variables symmetric to the main diagonal. It is sufficient to analyse the upper triangle matrix. The second figure in the first row shows the correlation between gluten spread and farinographic value. The inversely proportional linear correlation is clearly visible. In the same row, the third figure is a graph of gluten volume and farinographic value. The linear correlation is not shown here. As a function of gluten content, the farinograph value initially increases, then decreases and then increases again. It is like a cubic function. The fourth subsection shows the effect of protein on the farinograph.

There is no correlation, farinographic value might be any number within a narrow range of protein.

The third sub-diagram of the second row shows a directly proportional linear correlation between gluten volume and spread. The fourth sub-figure shows the linear independence of protein content and spread. If there is a correlation between the two variables, it is certainly not linear.

In the last sub-figure of the third row, the mean linear correlation between protein content and gluten volume can be recognized. It should be emphasized that these correlations are apparent and only valid between sample varieties and they are generally valid correlations. It is presumable that by repeating the analysis with modern wheat varieties, completely different findings could be recorded.

Figure 2. **Distribution and correlations of flour quality, Source: Sváb János, 1979. The font size of the numbers expresses the strength of the correlation**



The significance levels are <0.05 (*), <0.01 (**), and <0.001 (***)

There is a strong positive correlation between gluten spread and protein content according to the partial coefficient (0.95). The higher gluten volume, the better it spreads. The original r-value was 0.68, both indicators show the same correlation. Pearson's product-moment correlation value for gluten volume and protein content was 0.48 and that of the partial index was 0.71. There is no contradiction here either, the higher the protein content of the flour, the more gluten it contains. The partial coefficient (-0.62) indicates a stronger than average negative correlation between gluten spread and protein content. High

DOI: 10.34101/ACTAAGRAR/1/8587

protein content reduces gluten spread. Interestingly, the partial correlation indices of all six possible pairwise correlations are significantly higher than the values of the simple coefficients. In multiple cases, apparent independence is identified as a strong correlation by the partial index. Pairwise correlations are thus disturbed by other variables, sometimes causing effects that are incomprehensible due to nonlinearity. Consequently, this database is not suitable for principal component analysis, resulting in low KMO values.

CONCLUSIONS

Most multivariate statistical methods based on eigenvalue calculations impose strict application conditions on the data. One of the most important of these conditions is that the multiple linear correlation should be larger than the pairwise correlation. This is measured by the KMO test. However, when measuring the nutritional values of plants, this condition is often not met. There might be several reasons for this; it can be measurement methods, cultivation technology and perhaps most importantly, very often the condition of linearity is not met. If the correlation between the components is not linear, the value of Pearson's and partial correlation coefficient might change its sign. Another interesting effect is that the absolute values of the partial coefficients will be larger than the product-moment correlation values, which causes logical uncertainty. It means that the correlation between two components is closer than the correlation among multiple components. Analyses performed in the absence of applicability conditions might provide false results, i.e. the opposite of the obtained results might be true. If this property of nutritional value studies is immanent, completely different methods should be sought to analyse the data. In these cases, expert models, deterministic and semi-deterministic models, which combine the advantages of deterministic and stochastic models, can be used.

ACKNOWLEDGEMENTS

"Project no. TKP2020-IKA-04 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2020-4.1.1-TKP2020 funding scheme and with the support of the EFOP-3.6.3-VEKOP-16-2017-00008 project."

REFERENCES

Balázs, S. (2004): Zöldségtermesztők kézikönyve. Mezőgazda Kiadó. Budapest.

Bajtay, I. (1979): Csemegekukorica törzsek teljes és részleges diallél keresztezésekkel. Kandidátus értekezés. Budapest.

Brecht, J.K.–Sargent, S.A.–Hochmuth, R.C.–Tervola, R.S. (1990): Postharvest quality of supersweet (SH2) sweet corn cultivars. Proc. Annu. Meet.103. 283–288.

Burt, A.J.–Grainger, C.M.–Smid, M.P.–Shelp, B.J.–Lee, E.A. (2011): Allele minig of exotic maize germplasm to enhance macular carotenoids. Crop Science. 991–1004.

Calvo-Brenes, P.–O'hare, T. (2020): Effect of freezing and cool storage on carotenoid content and quality of zeaxanthin-biofortified and standard yellow sweet-corn (*Zea mays* L.) Journal of Food Composition and Analysis. 86. 103353.

Dániel, L. (1978): A csemege és pattogatni való kukorica termesztése. Mezőgazdasági Kiadó. Budapest.

Garwood, D.L.–Mcardle, F.J.–Vanderslice, S.F.–Shannon, J.C. (1976): Postharvest carbohydrate transformations and processed quality of high sugar maize genotipes. J.Am.Soc.Hortic.Sci.101. 400–404.

Hadi, G. (2003): Csemegekukorica. [In: Radics, L. (szerk.) Növénytermesztés határok nélkül.] Szaktudás Kiadó Ház. Budapest.

Kang, M.S.–Zuber, M.S.–Colbert, T.R.–Horrocks, R.D. (1986): Effects of certain agronomic traits on and relationship between rates of grain-moisture reduction and grain fill during the filling period in maize Field Crops Research.14. 339–347.

Kimura, M.–Kobori, C.N.–Rodriguez-Amaya, D.B.–Nestel, P. (2007): Screening and HPLC methods for carotenoids in sweetpotato, cassava and maize for plant breeding trials. Food Chemistry, 100(4), 1734–1746.

Lente, Á. (2012): A legfontosabb agrotechnikai tényezők hatása a csemegekukorica hibridek agronómiai tulajdonságaira és termésére. Doktori PhD értekezés, Debrecen

Ma, L.–Lin, X. (2010): Effects of lutein and zeaxantin on aspects of eye health. J. Sci.Food.Agric.90. 2–12.

Nagy, J. (2006): Maize production Akadémiai Kiadó. Budapest.

Obeng-Bio, E.–Badu-Apraku, B.–Ifi, B.E.–Danquah, A.–Blay, E.T.–Dadzie, M.A. (2020): Phenotypic characterization and validation of provitamin A functional genes in early maturing provitamin A-quality protein maize (*Zea mays*) inbred lines. Plant Breeding.139: 575–588.

Orosz, F. (2009): Termesztéstechnológiai elemek hatása a csemegekukorica koraiságára. Doktori (PhD) értekezés. Budapest.

Ray, K.–Banerjee, H.–Dutta, S.–Hazra, A.K.–Majumdar, K. (2019): Macronutrients influence yield and oil quality of hybrid maize (*Zea mays* L.). PLoS One. 14. 5. 1–23.

R Core Team: (2020): A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rosen, R.B.–Hu, D.N. (2018): Zeaxantin for tumor threatment. USA Patent. 15/974. 333.

Sváb, J. (1979): Többváltozós módszerek a biometriában. Mezőgazdasági Kiadó, Budapest, 1979. ISBN: 9632300114