1 Self-organized map: the new approach for study of genetic 2 divergence in kale

3

4

5Orlando Gonçalves Brito^{1*}, Valter Carvalho de Andrade Júnior², Alcinei Místico

6Azevedo³, Maria Thereza Netta Lopes Silva³, Ludimila Geiciane de Sá³, Clóvis

7Henrique Oliveira Rodrigues³, Ana Clara Gonçalves Fernandes³

8

9

10¹Department of Agronomy, Universidade Federal dos Vales do Jequitinhonha e Mucuri,

11Diamantina, Minas Gerais, Brazil.

12

13²Department of Agriculture, Universidade Federal de Lavras, Lavras, Minas Gerais,

14Brazil.

15

16² Institute of Agrarian Sciences, Universidade Federal de Minas Gerais, Montes Claros,

17Minas Gerais, Brazil.

18

19* Corresponding author

20E-mail: orlandocefet@yahoo.com.br (OGB)

21

22

23

24

Abstract

The objective of this study was to study the genetic divergence between genotypes of kale, to propose a methodology for the use of neural networks of the SOM type and to test its efficiency through Anderson's discriminant analysis. We evaluated 33 families of half-siblings of kale and three commercial cultivars. The design was a randomized block with four replications with six plants per plot. A total of 14 plant-level quantitative traits were evaluated. Genetic values were predicted at family level via REML / BLUP. For the study of divergence, neural networks of the SOM type (Self-organizing Map) were adopted. We evaluated different network architectures, whose consistencies of the clusters were identified by the Anderson discriminant analysis and by the number of empty clusters. After selecting the best network configuration, a dissimilarity matrix was obtained, from which a dendrogram was constructed using the UPGMA method. The best network architecture was formed with five rows and one column, totaling five neurons and consequently five clusters. The greatest dissimilarity was established between clusters I and V. The crossing between the genotypes of cluster I and those belonging to clusters III and V are the most recommended, since they aim to recombine families with characteristics of interest to the improvement and high dissimilarity. Anderson's discriminant analysis showed that the genotype classification was 100% correct, indicating the efficiency of the methodology used.

44

Keywords: *Brassica oleracea* var. *acephala*; genetic improvement; bioinformatics; biometry; neural networks.

47

48

49

32

4

50Introduction

51

52 Genetic improvement is one of the main tools to increase crop yield in a
53sustainable way [1]. However, for genotype selection in breeding programs, knowledge
54of genetic dissimilarity among the genotypes of the population to be improved is
55essential [2, 3]. Several molecular techniques have been used for genetic improvement.
56However, the use of phenotypic characters in the study of dissimilarity is more
57appropriate. This is due to its direct association with the attributes of interest in
58breeding. Thus, it is necessary to select dissimilar genotypes for recombination in order
59to provide greater genetic variability in the segregating populations [4].

60 Several studies on the evaluation of genetic divergence in brassicas have been
61carried out in different countries [5, 6, 7]. However, few studies have been done
62specifically with kale (*Brassica oleracea* var. *acephala*) [8].

63 Often genetic diversity studies have been performed using traditional
64multivariate techniques such as dendrograms, major components, and canonical
65variables. However, there is the possibility of carrying out these studies through
66computational intelligence using artificial neural networks (ANNs) [1, 9]. The main
67advantages of RNAs are their non-parametric approach, tolerance to data loss, and the
68need for detailed information about the modeling system as a design and genealogies
69[10, 11].

70 Self-organizing Map (SOM) neural networks are a type of exploratory
71multivariate analysis tool that allows, through artificial computational intelligence, to
72design high-dimensional data in a smaller dimensional space, without loss of
73information [12]. This new organization prioritizes maintaining the structure, such as
74clusters and information relationships [13]. This reinforces its constant use in several

grouping and process characterization works. However, the network topology is usually selected subjectively. In addition, at the beginning of the iterative process of SOM networks, the synaptic weights are random, which can generate several results for the same data set and network configuration. This may lead to the discrediting of this methodology, which requires the implementation of strategies to mitigate this problem.

The objective of this study was to study the genetic divergence between families of half-siblings of leaf kale in order to select parents, to propose a new methodology for the use of SOM neural networks and to test its efficiency through Anderson's discriminant analysis.

Material and methods

Experiment Implantation and data collection

The experiment was conducted at the Institute of Agricultural Sciences (ICA) of the Federal University of Minas Gerais, located in the municipality of Montes Claros-MG, from October 2016 to August 2017. The region is located at coordinates 16° 41 'S and 43° 50 'W and altitude of 646,29 m, with climate AW type (tropical savanna climate with dry winter and rainy summer) according to classification of Köppen [14], and included in the region belonging to the Brazilian semiarid.

The experimental design was a randomized block design with four replications. The evaluated treatments consisted of 36 genotypes of kale, with 33 half-sib families and three commercial cultivars (Butter, Portuguese Butter and Georgia Butter). The plants of the half-sib families were obtained through cloning, using sprouts from an experiment conducted at the Federal University of the Jequitinhonha and Mucuri Valleys (UFVJM).

100 Sprouts were removed from the plants when they were approximately 5 cm.
 101 Subsequently, they were rooted in styrofoam trays of 72 cells filled with commercial
 102 substrate. The sprouts remained in a greenhouse for 40 days until they reached seedlings
 103 for planting. Due to the difficulty of obtaining sprouts in the commercial cultivars, the
 104 seedlings of the same ones were produced from seeds acquired in the local commerce.
 105 Both the seedlings of the half-sib families and the commercial cultivars were taken to
 106 the field when they presented the same pattern.

107 Soil preparation was carried out through plowing and two harrowing.
 108 Afterwards, beds of 1.20 m were prepared, with the aid of mechanized equipment. In
 109 the plot were planted six seedlings of the corresponding genotype. The spatial
 110 arrangement of the plants was in double row, spaced 1.00 x 0.50 m, representing a
 111 population of 20 thousand ha⁻¹ plants. Planting fertilization and other cultural treatments
 112 such as irrigation, phytosanitary management and cover fertilization were carried out
 113 according to crop needs.

114 We evaluated 14 characteristics of quantitative origin in each plant of the plot
 115 (S1 Table). The number of shoots, leaves, the average mass of fresh leaf and the
 116 production of leaves were evaluated every 15 days. At 160 days the leaf length, limbus
 117 length, petiole length, petiole base diameter, petiole diameter and leaf width were
 118 evaluated. At 170 the plant height, stem height, stem base and middle were measured.
 119 These evaluation dates occurred in these periods (160 and 170 DAT) due to the fact that
 120 the plants are already considered adults and in full production.

121

122 **Data analysis**

123 The genetic divergence between families was characterized at the family level.

124 The statistical model $\underline{y} = \underline{X}\underline{\tau} + \underline{Z}\underline{a} + \underline{W}\underline{p} + \underline{e}$, in which $\underline{y}$ is the data vector, $\underline{\tau}$ is the
125 vector of the repetitive effects (assumed as fixed) added to the general average, $\underline{a}$ is the
126 vector of the individual additive genetic effects (random), $\underline{p}$ is the vector of the plots
127 effects (random), $\underline{e}$ is the vector of errors or residuals (random). Capital letters represent
128 the incidence matrices for these effects. For this, genetic-statistical software Selegen-
129 REML / BLUP [15] was used. Subsequently, the data were standardized to obtain an
130 average of zero and a standard deviation of one. Neural networks of the SOM type were
131 adopted for the study. Different network architectures were evaluated, varying the
132 number of rows from 1 to 6 and columns from 1 to 6. Excluding the combination of a
133 row and a column, which would provide a single cluster, there were 35 tested
134 configurations.

135 At the beginning of the iterative process of SOM networks, the synaptic weights
136 are random, this can lead to several results for the same data set and network
137 configuration. Therefore, the best network architecture was established from 1000
138 trainings performed for each of the 35 configurations. For each of them the average hit
139 rate was estimated through Anderson's discriminant analysis [16] and the smaller
140 number of empty clusters. The best network architecture was selected from the highest
141 average hit rate and the lowest number of empty clusters, simultaneously, following the
142 parsimony principle.

143 Having determined the best configuration (topology) of the network, it was
144 submitted to 1000 new trainings, and later a matrix of dissimilarity was built. The
145 dissimilarity between each pair of genotypes was determined from the frequency at
146 which they were allocated in different clusters.

147 Then the dendrogram of dissimilarity was made from the UPGMA method
148(Unweighted Pair Group Method using Arithmetic averages). The number of clusters of
149the dendrogram was established according to the number of neurons formed in the best
150network configuration. The consistency of the clustering was verified through
151Anderson's discriminant analysis.

152 For the complementation of the clustering study, the graphical dispersion of the
153dissimilarity matrix in two-dimensional projection [17] was performed. In this
154procedure, the dissimilarity measures are converted into scores relative to two variables
155that, when represented in scatter plots, will reflect the distances originally obtained.

156 All analyzes were done in R [18] software (S2 Table). For the use of SOM
157networks, the RSNNS package was used. In order to obtain the dendrogram, the hclust
158function was used.

159

160Results

161

162 It was found that the best topology was established with five rows and one
163column (Fig 1A). This configuration was the one that allowed the Anderson Anderson
164discriminant analysis (98.72%) (Fig 1A) and showed a lower number of empty clusters
165(zero) (Fig 1B). Therefore, it was possible to form five neurons (five lines x one
166column), implying the formation of five clusters.

167

168Figure 1 - Average percentage of correct classifications by Anderson discriminant
169analysis (A) and percentage of empty clusters considering 35 SOM (B) network
170configurations.

171

172 From the best network architecture, the dissimilarity matrix was obtained from
173 Kohonen's self-organizing maps (Fig 2). Lighter colors indicate smaller genetic
174 distances between genotypes, while darker colors represent greater distances, that is,
175 greater genetic divergence. It was observed that there was high genetic divergence
176 between most of the genotypes (Fig 2). Commercial cultivars (COM1, COM2 and
177 COM3) were similar to each other but divergent in relation to most half-sib families.
178 The families most genetically similar to the commercial cultivars were F8, F22 and F23.

179

180 Figure 2 - Graphical representation of the dissimilarity matrix obtained by Kohonen
181 self-organizing maps. "F" refers to the half-sib family and "Com" to the commercial
182 cultivar.

183

184 Analyzing the clusters formed by the UPGMA method (Fig 3), it is verified that
185 the first cluster (I) was composed of five half-sib families (F3, F4, F7, F16 and F22) and
186 the three commercial cultivars. These materials were characterized by low number of
187 sprouts, leaves and yield, as well as intermediate values of the base and middle stem
188 diameters (Fig 4).

189

190 Figure 3 - Dendrogram obtained by the UPGMA method from the dissimilarity matrix
191 obtained by SOM type networks for the composition of the clusters (I, II, III, IV and V)
192 considering 14 quantitative characteristics in leaf cabbage genotypes.

193

194 Figure 4 - Graphical representation of the standardized genetic values for the study of
195 the contribution of each characteristic to the composition of the clusters. Characteristics:
196 NS: number of shoots; NL: number of leaves; AMFL: average mass of fresh leaf; LY:

197 leaf yield; LL: leaf length; LBL: limbus length; PL: petiole length; DPB: diameter of
198 the base of the petiole; DPM: diameter of the petiole medium; LW: leaf width; PH:
199 plant height; SH: stem height; SBD: stem base diameter; RD: diameter of the base of
200 the stem. * Cluster.

201

202 The commercial cultivars showed the lowest values for all the characteristics
203 when compared with the half-sib families evaluated. This was only observed for the
204 petiole diameters (DBP and DPM) and stem (SBD and RD). It was also observed that
205 the diameter of the base of the petiole presented similar values for all families, except
206 for the family F8 (higher value) and F20 (smaller value).

207 The second cluster (II) consisted of seven half-sib families (F8, F18, F21, F23,
208 F25, F28 and F32). These families, in general, presented intermediate values for the
209 number of sprouts and diameter of the middle of the petiole, in addition to plants with
210 smaller stem base diameter and smaller width of leaves. As observed for cluster I, these
211 families also presented lower agronomic performance than the other genotypes, with
212 low leaf production (Fig 4). The third cluster (III) was formed by ten families of half-
213 brothers (F5, F9, F10, F11, F12, F13, F14, F29, F30 and F31). This cluster was
214 characterized by presenting families with higher genetic values for practically all the
215 characteristics evaluated, especially for yield, number of shoots, stem base diameter and
216 leaf width.

217 The fourth and fifth cluster (IV and V) were composed of five and six half-sib
218 families, respectively. These clusters were composed of families with higher plant
219 height and stem height, and cluster V also presented high number of shoots and high
220 leaf yield. It should be emphasized that the families of clusters IV and V, also, have
221 characteristics that are not of interest for the improvement of the crop, plants of low

diameter of the bases of the stem (cluster IV), with high values for height of plants and petiole length (mainly in cluster V), as well as a high number of buds and leaves too large (high values of LL and LW) (Fig 4).

In a possible selection, clusters I, III and V should be prioritized, as they present half-sibling families with agronomic characteristics of interest for improvement, such as lower plant heights and lower number of shoots (cluster I), larger stem diameter (cluster III) and high leaf yield (cluster III and V). The recombination of these characteristics may contribute to a greater heterotrophic effect in the segregating population, which is of interest.

It was verified that the cluster IV was the most homogeneous, presenting smaller average genetic distance within the cluster (0.232) (Table 1). This can be verified through the smaller variation of color existing between the families within each assessed characteristic (Fig 4). The fact that this cluster presents few genotypes may also have contributed to this greater homogeneity.

236

Table 1 - Average distances within (main diagonal in bold) and between (outside diagonal) clusters based on the dissimilarity matrix obtained by Kohonen's Self-Organizing Maps.

Cluster	I	II	III	IV	V
I	0,261	0,684	0,990	0,986	1,000
II	-	0,271	0,991	0,873	0,993
III	-	-	0,352	0,937	0,734
IV	-	-	-	0,232	0,707
V	-	-	-	-	0,258

240

When comparing clusters, it is verified that the greatest divergence was established between cluster I and V, with genetic distance between them equal to 1.00. The graphical dispersion of the dissimilarity matrix shown in Fig 5 shows that the

dispersion between and within the clusters was high, and that cluster I was very distant from clusters III, IV and V. It is also possible to verify that families F8 and F23 were positioned very close to cluster I (Fig 5), but not enough to be grouped in cluster I by the UPGMA method (Fig 3). This was probably due to the presence of some characteristic strongly associated to the other genotypes that compose the clusters II. In relation to the efficiency of the clusters using the Anderson discriminant analysis (Table 2), it was verified that the classification adopted allowed a rate of 100% accuracy.

251

Figure 5 - Graphic dispersion of the dissimilarity matrix in 1000 SOM type networks with five rows and one column (5 neurons). The colors represent the four distinct clusters formed: I (●), II (●), III (●), IV (●) e V (●).

255

Table 2 - Percentage of correct classifications by the Anderson discriminant function for each cluster formed by the UPGMA method.

Cluster	I	II	III	IV	V
I	100	0	0	0	0
II	0	100	0	0	0
III	0	0	100	0	0
IV	0	0	0	100	0
V	0	0	0	0	100

258

259 Discussion

260

The efficiency of the use of neural networks in breeding programs has been presented in several studies, such as the selection of more productive genotypes of sugarcane (*Saccharum* spp.) [19] and the evaluation of genetic divergence (*Vitis vinifera* L.) [20], papaya (*Carica papaya* L.) [1] and guava (*Psidium guajava* L.) [9].

265 In kale the neural networks were used by Azevedo et al. [11] in leaf area
266prediction. The technique proved to be feasible to estimate the leaf area and to assist in
267the selection of superior genotypes. However, there are no other studies carried out with
268the crop and RNAs, which reinforces the importance of the study for application in
269breeding, especially for the study of genetic divergence.

270 The self-organizing maps began to be explored by T. Kohonen as of 1982,
271however their use is becoming more frequent [13]. The SOM-type networks present
272unmasked clustering process and perform the training based on the patterns of their
273spatial organization only in the homology (similarity) of the data, that is, without prior
274knowledge of the class to which they belong [21].

275 This allows the technique to be an appropriate procedure for grouping the
276genotypes based on morphoagronomic characters, since there is no requirement for
277calibration [22] without the need for information on experimental designs and
278genealogies, which in some situations is unknown . The high divergence observed in the
279matrix of dissimilarity (Fig 2) and the dispersion of the families in the graphical
280representation (Fig 5) is indicative of genetic variability in the population, which points
281out its potential for genetic improvement [8]. Significant genetic gains with selection
282are conditioned by the existence of such genetic variability [23].

283 The representation of the matrix by the dendrogram (Fig 3) also showed a high
284co-optic correlation, which shows high agreement in the grouping performed and the
285dissimilarity matrix [24]. This reinforces the reliability of subsequent crossing to be
286established based on the composition of the clusters using the RNAs. González-Cuellar
287and Obregón-Neira [25] also point out that the main advantage associated with
288Kohonen's self-organizing maps is the easy visualization of the established

classification, since in other classification methods involving a greater number of variables the process becomes more complex.

Although it does not present families of high yield, cluster I is composed of materials with characteristics of interest for the improvement of the kale, since it resembles the commercial cultivars. Among these characteristics, the smaller petiole length, the lower plant height and the stem, and especially the low number of shoots, stand out. This approach is closer to what is required as a commercial and agronomic standard, because in addition to resistance to the main pests, the genetic improvement of kale must prioritize genotypes with lower height, fewer shoots and more commercial leaves. This favors higher yield, facilitates crop management, harvesting, and reduces costs with plant staking. Thus, it is indicated to include these families to compose the recombinant population, along with other families of higher yield and number of leaves, as observed for the families of clusters III and V.

Another relevant observation is that for most of the families evaluated in this study, the production and number of leaves presented a positive degree of association with the increase in the number of shoots, that is, more productive families tended to have a higher number of shoots. This differs from Azevedo et al. [26], who found a low degree of association between the number of shoots and the mass of fresh leaves when evaluating genetic correlations in leaf cabbage genotypes.

This association identified in the study population is not of interest for the process of improvement of the same. This can favor the vegetative propagation, which generates higher cost with crop by the producer. The presence of this characteristic also reduces the producer's dependence on seeds, which may not be favorable when commercial cultivars are to be obtained. Therefore, it is necessary that future recombinations prioritize the crossing of low sprouting plants with high yield plants.

314 The reduction of the number of shoots can be done from crossing between the
315 clusters I families (low plants and with few buds) with those belonging to clusters III
316 and V, which presented high yield, greater number of leaves and larger stem diameters.
317 These clusters were also genetically distant (Table 1 and Fig 5). This is important in the
318 recombination process, since it increases the probability of obtaining superior genotypes
319 in yield and with agronomic characteristics of interest in the segregating population.
320 Important crosses can also be established between the families of clusters II with
321 families of clusters III and V, since they are the most genetically distant.

322 The genetic distance between clusters and within clusters (Table 2) and
323 graphical dispersion (Fig 5) show that although the UPGMA method forms clusters
324 with families of similar characteristics, they present high variability among themselves,
325 and also within clusters formed. This corroborates with that observed in the
326 dissimilarity matrix formed from Kohonen's self-organizing maps (Fig 2), which
327 indicated high divergence among genotypes. This can occur because the population is in
328 the early stages of breeding.

329 As for the efficiency of the adopted methodology, Anderson's discriminant
330 analysis showed that it is adequate and efficient. This is in agreement with Barbosa et
331 al. [1], which highlights that the discriminant analysis is useful in checking the
332 consistency of clusters, and is recommended for this type of study. Thus, it is clear the
333 consistency of the methodology used and its indication for future work.

334

335 **Conclusions**

336

337 Crossings between families belonging to clusters I and families of clusters III
338 and V are the most recommended to allow the selection of plants with lower number of

339 shoots, plant height, leaf length, leaf width, petiole length and greater number of leaves,
340 leaf yield and stem diameters.

341 The proposed methodology for the use of SOM networks is efficient, since it
342 allows the selection of the network configuration without subjectivity and high rate of
343 correctness by the Anderson descending analysis, forming clusters in agreement with
344 the morphological data.

345

346 **Supporting information**

347

348 **S1 Table.** Database for variables evaluated in half-sib families progenies of leaf kale.

349 (xlsx)

350 **S2 Table.** Script programming. (DOCX)

351

352 **Acknowledgment**

353

354 To CNPq, FAPEMIG and CAPES for the granting of scholarships and financial
355 resources for the development of the project. This study was financed in part by the
356 Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) -
357 Finance Code 001. To UFVJM for the opportunity to develop the study and germplasm
358 supply.

359

360 **Author Contributions**

361

362Conceptualization: Orlando Gonçalves Brito, Valter Carvalho de Andrade Júnior,
363Alcinei Místico Azevedo.

364Data Curation: Orlando Gonçalves Brito, Alcinei Místico Azevedo.

365Formal Analysis: Orlando Gonçalves Brito, Alcinei Místico Azevedo.

366Investigation: Orlando Gonçalves Brito, Maria Thereza Netta Lopes Silva, Ludimila
367Geiciane de Sá, Clóvis Henrique Oliveira Rodrigues, Ana Clara Gonçalves Fernandes.

368Resources: Valter Carvalho de Andrade Júnior.

369Writing – Original Draft Preparation: Orlando Gonçalves Brito, Valter Carvalho de
370Andrade Júnior, Alcinei Místico Azevedo.

371Writing – Review & Editing: Orlando Gonçalves Brito, Alcinei Místico Azevedo,
372Valter Carvalho de Andrade Júnior.

373

374References

375

3761. Barbosa CD, Viana AP, Quintal SSR, Pereira MG. Artificial neural network
377 analysis of genetic diversity in *Carica papaya* L. Crop Breed. Appl. Biotechnol.
378 2011; 11: 224-231. doi: 10.1590/S1984-70332011000300004.

379

3802. Shah MA, Khan AI, Awan FS, Sadaqat HA, Bahadur S, Rasheed, A, et al. Genetic
381 diversity of some tomato cultivars and breeding lines commonly used in Pakistani
382 breeding program. Turkish JAF Sci. Tech. 2015; 3: 126-132. doi:
383 10.24925/turjaf.v3i3.126-132.249.

384

3853. Andrade EKV, Andrade Júnior VC, Laia ML, Fernandes JSC, Oliveira AJM,
386 Azevedo AM. Genetic dissimilarity among sweet potato genotypes using

- 387 morphological and molecular descriptors. Acta Sci. Agron. 2017; 39: 447-455. doi:
388 10.4025/actasciagron.v39i4.32847.
389
3904. Figueiredo AST, Resende JTV, Schwarz K, Marondin JC, Galvão AG, Resende
391 NCV. Genetic divergence among processing tomato hybrids and formation of new
392 segregating populations. Ciênc. Agrotec. 2017; 41: 279-287. doi: 10.1590/1413-
393 70542017413033016.
394
3955. Koutita O, Tertivanidis K, Koutsos T, Koutsika-Sotiriou M, Skaracis G. Genetic
396 changes between C0, C3, C5 cycles of selection and S4 lines of two cabbage
397 (*Brassica oleracea* L. var. *capitata* L.) open pollinated populations based on RAPD
398 markers. Biotechnol. 2008; 7: 475-480. doi: 10.3923/biotech.2008.475.480.
399
4006. Christensen S, Bothmer RV, Poulsen G, Maggioni L, Phillip M, Andersen BA, et
401 al. AFLP analysis of genetic diversity in leafy kale (*Brassica oleracea* L. convar.
402 *acephala* (DC.) Alef.) landraces, cultivars and wild populations in Europe. Genet.
403 Resour. Crop Evol. 2011; 58: 657-666. doi: 10.1007/s10722-010-9607-z.
404
4057. Belete YS, Kebede SA, Gemelal AW. Multivariate analysis of genetic divergence
406 among Ethiopian Mustard (*Brassica carinata* A. Brun) genotypes in relation to seed
407 oil quality traits. Int. J. Agric. Res. 2011; 6: 494-503. doi:
408 10.3923/ijar.2011.494.503.
409

4108. Azevedo AM, Andrade Júnior VC, Pedrosa CE, Valadares NR, Fernandes JSC,
411 Ferreira MRA, et al. Divergência genética e importância de caracteres em genótipos
412 de couve. Hortic. Bras. 2014; 32: 51-57. doi: 10.1590/S0102-05362014000100008.
413
4149. Campos BM, Viana AP, Quintal SSR, Barbosa CB, Daher RF. Heterotic group
415 formation in *Psidium guajava* L. by artificial neural network and discriminant
416 analysis. Rev. Bras. Frutic. 2016; 38: 151-157. doi: 10.1590/0100-2945-258/14.
417
41810. Silva GN, Tomaz RS, Sant'Anna IC, Nascimento M, Bhering LL, Cruz, CD.
419 Neural networks for predicting breeding values and genetic gains. Sci. agric. 2014;
420 71: 494-498. doi: 10.1590/0103-9016-2014-0057.
421
42211. Azevedo AM, Andrade Júnior VC, Sousa Júnior AS, Santos AA, Cruz CD, Pereira
423 SL, et al. Eficiência da estimação da área foliar de couve por meio de redes neurais
424 artificiais. Hortic. Bras. 2017; 35: 14-19. doi: 10.1590/s0102-053620170103.
425
42612. Gontijo ESJ, Oliveira FSD, Fernandes ML, Silva GA, Roeser HMP, Friese K,
427 2014. Application of a multivariate exploratory analysis technique in the study of
428 dissolved organic matter and metal ions in waters from the eastern quadrilátero
429 ferrífero, Brazil. J. Braz. Chem. Soc. 2014; 25: 208-218. doi: 10.5935/0103-
430 5053.20130285.
431
43213. Stefanovic P, Kurasova O. Visual analysis of self-organizing maps. Nonlinear
433 Anal. Model. Control. 2011; 16: 488–504.
434

43514. Köppen W. Climatologia: con un estudio de los climas de la tierra. México: Fondo
436 de Cultura Econômica; 1948.

437

43815. Resende MDV. Software Selegen-REML/BLUP: a useful tool for plant breeding.
439 Crop Breed. Appl. Biotechnol. 2016; 16: 330-339. doi: 10.1590/1984-
440 70332016v16n4a49.

441

44216. Anderson TW. An introduction to multivariate statistical analysis. New York: Jhon
443 Wiley & Sons; 1958.

444

44517. Cruz CD. GENES - a software package for analysis in experimental statistics and
446 quantitative genetics. Acta Sci. Agron. 2013; 35: 271-276. doi:
447 10.4025/actasciagron.v35i3.21251.

448

44918. R CORE TEAM, 2016. R: A language and environment for statistical computing. R
450 Foundation for Statistical Computing, Viena.

451

45219. Brasileiro BP, Marinho CD, Costa PMA, Cruz CD, Peternelli LA, Barbosa MHP.
453 Selection in sugarcane families with artificial neural networks. Crop Breed. Appl.
454 Biotechnol. 2015; 15:72-78. doi: 10.1590/1984-70332015v15n2a14.

455

45620. Mancuso S. Clustering of grapevine (*Vitis vinifera* L.) genotypes with Kohonen
457 neural networks. Vitis 2001; 40: 59-63.

458

45921. Astel A, Tsakovski S, Barbieri P, Simeonov V. Comparison of self-organizing
460 maps classification approach with cluster and principal components analysis for
461 large environmental data sets. Water Res. 2007; 41: 4566-4578. doi:
462 10.1016/j.watres.2007.06.030.
463
46422. Delgado S, Morán F, Mora A, Merelo JJ, Briones C. A novel representation of
465 genomic sequences for taxonomic clustering and visualization by means of self-
466 organizing maps. Bioinformatics 2015; 31: 736-744, doi:
467 10.1093/bioinformatics/btu708.
468
46923. Tilahun G, Mekbib F, Fikre A. Genetic divergence and character association of
470 kabulitype chickpea (*Cicer arietium*) genotype under rainfed conditions in Ethiopia.
471 Curr. Res. Agric. Sci. 2015; 2: 123-131. doi:
472 10.18488/journal.68/2015.2.4/68.4.123.131.
473
47424. Silva AR, Dias CTS. A cophenetic correlation coefficient for Tocher's method.
475 Pesq. Agropec. Bras. 2013; 48: 589-596. doi: 10.1590/S0100-
476 204X2013000600003.
477
47825. González-Cuéllar F, Obregón-Neira N. Self-organizing maps of Kohonen as a river
479 clustering tool within the methodology for determining regional ecological flows
480 ELOHA. Ing. Univ. 2013; 17: 311-323.
481
48226. Azevedo AM, Andrade Júnior VC, Santos AA, Souza Júnior AS, Oliveira AJM,
483 Ferreira MAM. Population parameters and selection of kale genotypes using

- 484 Bayesian inference in a multi-trait linear model. *Acta Sci. Agron.* 2017; 39:25-31.
- 485 doi: 10.4025/actasciagron.v39i1.30856.

1

2Figure 1 - Average percentage of correct classifications by Anderson discriminant
3analysis (A) and percentage of empty clusters considering 35 SOM (B) network
4configurations.

5

6

7Figure B - Graphical representation of the dissimilarity matrix obtained by Kohonen
8self-organizing maps. "F" refers to the half-sib family and "Com" to the commercial
9cultivar.

10

11Figure C - Dendrogram obtained by the UPGMA method from the dissimilarity matrix
 12obtained by SOM type networks for the composition of the clusters (I, II, III, IV and V)
 13considering 14 quantitative characteristics in leaf cabbage genotypes.

14

15Figure D - Graphical representation of the standardized genetic values for the study of
 16the contribution of each characteristic to the composition of the clusters. Characteristics:
 17NS: number of shoots; NL: number of leaves; AMFL: average mass of fresh leaf; LY:
 18leaf yield; LL: leaf length; LBL: limbus length; PL: petiole length; DPB: diameter of
 19the base of the petiole; DPM: diameter of the petiole medium; LW: leaf width; PH:
 20plant height; SH: stem height; SBD: stem base diameter; RD: rod diameter . * Cluster.

21

22

23

24

25

26 Figure E - Graphic dispersion of the dissimilarity matrix in 1000 SOM type networks
 27 with five rows and one column (5 neurons). The colors represent the four distinct
 28 clusters formed: I (●), II (●), III (●), IV (●) e V (●).