CBAB
CROP BREEDING AND
APPLIED BIOTECHNOLOGY

ARTICLE

Bayesian network: a simplified approach for environmental similarity studies on maize

Crop Breeding and Applied Biotechnology
19: 70-76, 2019
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332019v19n1a10

Camila Baptista do Amaral[*1], Gustavo Hugo Ferreira de Oliveira[1], Kian Eghrari[1], Rodolfo Buzinaro[1] and Gustavo Vitti Môro[1]

Abstract: *The current methodologies used to evaluate environmental similarities do not allow the simultaneous analysis and categorization of the environments. The objective of this study was to verify the possibility of using the Bayesian network (BN) to detect similarities between environments for plant height, lodging, and grain yield in maize. Thirteen experimental varieties were grown in six environments to measure the traits plant height, lodging, and grain yield. The BN was constructed for each trait, using the Hill-Climbing algorithm. Results were compared with the simple part of the genotypes x environments interaction, clustering by the Lin's method and by simple correlation between environments. The Lin's method clustered environments with predominance of complex interaction for all traits. The BN is efficient to analyze environmental similarity for plant height and grain yield since it detected the highest correlations. The BN revealed no connections among the environments that presented predominance of complex interaction.*

Keywords: *Zea mays, prediction method, environmental correlation, genotype x environment interaction.*

INTRODUCTION

The primary objective of breeding programs is to release new cultivars with optimal agronomic traits. To reach this goal, the evaluation of genotypes in several years and locations is necessary and helps estimate genotype x environment interaction. This interaction prevents the generalized recommendation of genotypes and demands the study of the genotype in specific environments. This process is expensive and requires financial and human resources, which makes the research onerous. The interaction can also be used to select similar environments with predominance of simple interaction (Garbuglio et al. 2007).

Cruz and Castoldi (1991) proposed a methodology that divides the interaction into simple and complex parts, based on the decomposition of the mean square of genotypes x environments interaction (GE). Despite being adequate to evaluate this type of experiment, this methodology does not allow the simultaneous analysis and categorization of the environments, once the result is given by pairs of environments. Another frequently used methodology is the Lin's algorithm (Lin 1982), which groups the environments based on the absence of interaction.

The Bayesian network (BN) is an approach that represents cause and effect relations. Its graphical structure allows the identification of assumptions between system variables that may be obscured when using other methodologies

***Corresponding author:**
E-mail: camila.agro07@gmail.com
ORCID: 0000-0002-9148-8843

Received: 18 April 2017
Accepted: 23 July 2017

[1] Departamento de Fitotecnia, Universidade Estadual Paulista, 14.884-900, Jaboticabal, SP, Brazil

(Borsuk et al. 2004). Currently, the BN has been studied to predict variables. Felipe et al. (2015) successfully analyzed the predictive capacity of the BN considering 31 traits, which revealed that the BN can be used even with a large number of traits, and allowed their joint analysis.

The present study raises the hypothesis that the BN can be used to predict environments instead of variables, and has the potential to be used to evaluate similarity between environments with predominance of simple interaction, simultaneously. Therefore, the objective of this study was to verify the possibility of using the BN to detect similarity between environments for plant height, lodging, and grain yield in maize.

MATERIAL AND METHODS

The experiments with the Bayesian network were carried out in two locations, Jaboticabal (lat 21º 14’ 33’’ S, long 48º 17’ 10’’ W, and 565 m asl), SP, Brazil and Campo Alegre de Goiás (lat 17º 38’ 20’’ S, long 47º 46’ 55’’ W, and 884 m asl), GO, Brazil. These locations were selected due to their diverse environmental conditions. The same experiment was carried out in five different seasons in Jaboticabal (Environment 1: first season of 2009/2010; 3: second season of 2010; 4: first season of 2015/2016, under low nitrogen conditions; 5: first season of 2015/2016, under high nitrogen conditions; and 6: second season of 2016), and one season in Campos Alegre de Goiás (Environment 2: second season of 2010). Each season represented one environment to be included in the modeling approach.

Information from 13 open-pollination synthetic varieties, obtained as described by Oliveira et al. (2016), was evaluated in each environment. All experiments were arranged in a complete randomized block design, with three replications. Each plot consisted of two 5 m-long rows, and the population was corrected to 60,000 plants ha^{-1}. The management of the experiments followed the recommendations of Fornasieri Filho (2007).

The following traits were measured: plant height, determined by the distance in cm between the ground and the insertion of the flag leaf, in 10 random plants per plot; number of lodged plants per plot, determined by the breakage below the ear and maize root lodging; and grain yield per plot. After physiological maturity, the ears of both rows of the plot were hand-harvested; the grains were separated and weighed, and the grain moisture was determined. The grain yield of each plot was corrected to 13% of humidity and converted to kg ha^{-1}.

The BN is a graphical representation of a probability distribution over a set of variables (Felipe et al. 2015). The Directed Acyclic Graph (DAG) represents the BN using nodes connected by arrows, and is used as an output to the modeling approach. In this case, it is used to illustrate the association between environments. This graph characterizes a joint probability of the data, which brings scale benefits due to the factorization (Aliferis et al. 2010). In a set of variables $\{X_1, X_2,..., X_p\}$ with joint distribution $\Pr(X_1, X_2,..., X_p)$ and a DAG D that is compatible with this joint distribution (Pearl 2000), the following factorization can be performed:

$$Pr\ (X_1,X_2,...,X_\mathrm{p}) = \prod_{i=l}^{p} \Pr(X_1 | \mathbf{Pa}_i)$$

in which Pa_i are the parents of X_i in D. The BN analysis involves searching for a structure that is compatible with the joint distribution of the data. The selected structure has already been used as a prediction tool, as described by Felipe et al. (2015). In this study, the BN was only used in the context of environmental association.

For the present work, the Hill-Climbing algorithm (“search and score” approach) was used to construct the BN from the means of each plot. The model was adjusted using the package “bnlearn” of the R software (Scutari 2009). The environmental correlation was estimated, using the Pearson’s correlation, to quantify the relationship between environments associated by the BN. The magnitude of the correlations was analyzed according to the limits of interpretation of correlations proposed by Carvalho et al. (2004), where: r = 0.0 (no correlation); 0.0 < |r| < 0.30 (weak); 0.30 < |r| < 0.60 (intermediate; 0.60 < |r| < 0.90 (strong); 0.90 < |r| < 1 (very strong) and |r|= 1 (perfect). After the network was “learned” from environment data, additional conventional methodologies were applied to validate the BN. Therefore, joint analysis was performed considering all the six environments, using the following model:

$$Y_{ijk} = m + B(A_{jk}) + G_1 + A_j + GA_{ij} + E_{ijk}\ ,$$

where $\mathrm{Y}_{\mathrm{ijk}}$ is the phenotypic observation; m is the general mean; $\mathrm{B(A_{jk})}$ is the effect of k block within the j$^{\mathrm{th}}$ environment;

G_i is the effect of the ith genotype; A_j is the effect of the jth environment; GA_{ij} is the effect of the interaction between the ith genotype and the jth environment; and E_{ijk} is the random error or residue. All effects, except for error, were considered as fixed.

Environmental stratification was carried out using the conventional approach as proposed by Cruz and Castoldi (1991). This method can be used when the GE interaction is significant between the pair of environment, decomposing this interaction into two parts. The first part, denominated as simple interaction, is determined by the difference in variability between genotypes in the environments; and the second part, denominated as complex, is given by the absence of correlation between genotypes under environmental variation (Cruz et al. 2012). Moreover, this methodology allows estimating the Pearson's and Spearman's correlation. In this method, the lowest values of the percentage of simple interaction represent the most different environments.

The division of the simple part of the mean squares of the interaction (MSGxE) was performed for plant height, lodging, and grain yield, using the following formula:

$$S = (1 - r)\sqrt{Q_1 . Q_2}$$

where Q_1 and Q_2 were the mean squares of genotypes in environments 1 and 2, respectively; and r was the correlation between the genotypes means in both environments. The percentage of the simple interaction of MSGxE is expressed as follows:

$$\%S = \frac{100S}{MSG{\times}E},$$

where S = MSGxE – C, being C the complex interaction represented by $\sqrt{(1-r)^3 Q_1 Q_2}$.

Another estimation method was proposed by Lin (1982), which considers the sum of squares for the interaction between genotypes and pairs of environments, and subsequently clusters of environments with non-significant interaction. Afterward, the method estimates the sum of squares between genotypes and groups of three environments each time, and uses the F test to evaluate the possibility of creating a new group. A sum of square of the pairs of environments was estimated, using the means, according to Cruz et al. (2012), by:

$$MSGxE = \frac{1}{2}\left[d^2_{jj'} - \frac{1}{t}(Y_{.j} - Y_{.j'})^2\right]$$

where $d^2_{jj'} = \Sigma_i(Y_{ij} - Y_{ij'})^2$. The highest values represent the most similar environments. The Genes software (Cruz 2013) was used to analyze the algorithms of Cruz and Castoldi (1991) and Lin (1982).

RESULTS AND DISCUSSION

The experimental coefficient of variation (Table 1) was classified as intermediate for plant height and grain yield, and as very high for lodging, according to Fritsche-Neto et al. (2012). The coefficient of variation is an adequate method to evaluate the experimental precision and the estimated mean accuracy (Cargnelutti Filho and Storck 2007). Lodging usually presents high values of phenotypic coefficient of variation, as reported by Nzuve et al. (2014), due to the difference of influence of the environment in the plots for the trait.

The joint analyses of variance revealed significant effects at 1% probability for environments, genotypes, and the GE for all traits (Table 1), indicating the presence of differences among environments, variability among genotypes, and different response of genotypes to environmental condition. Quantitative traits usually present genotypes x environments interaction (Fan et al. 2007), requiring the unfolding of the interaction into simple and complex interactions.

The unfolding of the GE in the percentage of simple effect for pair of environments by the method proposed by Cruz and Castoldi (1991) showed that most of it were composed by simple interaction between the pairs of environments for all traits (Table 2).

Synthetic varieties are composed of a high number of genotypes, leading to great variability within the population, as demonstrated by Semagn et al. (2014), who found significant genetic variability within open-pollination varieties. This variability implies high stability, defined as the ability to maintain performance throughout multiple environments (Mansfiel and Mumm 2014). In this case, the stability results in predominance of simple interaction, that is, the great

Table 1. Summary of joint analyses of variance of plant height, lodging, and grain yield of 13 maize varieties in six environments

Trait	**Mean squares**				**CV** (%)
	Environment (E)	**Genotype** (G)	**G x E**	**Error**	
Plant height	8582.065**	1639.735**	200.686**	114.810	5.42
Lodging	1794.664**	84.834**	39.204**	14.410	63.40
Grain yield	105758246.000**	9772840.000**	777180.000**	408180.000	12.16

** Significant at 1% by the F test.

number of genotypes in the population confers the ability to predict the mean performance, regardless of the environmental effects.

The environments clustering based on the Lin's method (Lin 1982) formed one group for plant height and lodging, and four groups for grain yield (Table 3), as expected, due to the higher percentage of complex interaction between pairs of environments for plant height and lodging. For plant height, the group was formed with the environments 1, 3, 4, 5 and 6. However, the decomposition of the genotypes x environments interactions indicated predominance of the complex interaction between pairs 1 x 4, 1 x 6, 3 x 4, 3 x 6, 4 x 6 and 5 x 6 (Table 2), suggesting inconsistency between the results of the unfolding of the effect of the interaction and the results obtained with the Lin (1982)'s method. The same disparity was observed for lodging, where the pairs 2 x 6, 2 x 5, 3 x 6, 4 x 5, 4 x 6 and 5 x 6 presented predominance of complex interaction and were allocated in the same group.

Table 2. Percentage of the simple interaction determined by the method of Cruz and Castoldi (1991) for plant height, lodging, and grain yield, in the competition experiments of maize varieties in six environments

Environment	**Plant height**	**Lodging**	**Grain yield**
1 x 2	42.68	57.97	73.72
1 x 3	53.23	57.84	70.20
1 x 4	48.15	86.83	43.14
1 x 5	68.53	78.74	56.75
1 x 6	31.17	63.75	30.58
2 x 3	80.31	59.49	56.81
2 x 4	91.03	69.29	61.16
2 x 5	90.31	48.16	82.13
2 x 6	57.02	19.95	29.38
3 x 4	46.78	86.66	62.93
3 x 5	55.97	61.81	61.79
3 x 6	33.76	33.09	26.25
4 x 5	59.74	35.27	57.68
4 x 6	29.39	13.38	56.63
5 x 6	24.90	6.17	54.10
Mean	54.20	51.89	54.88

The groups formed for grain yield were 1-2-3-5, 2-6, 4-5 and 3-4 (Table 3), and the pair 2-6 presented 70.62% of complex interaction. Cruz et al. (2012) also reported differences between the environment clustering using the Lin's algorithm and the environment clustering using the method of Cruz and Castoldi (1991), where the interaction was predominantly simple The authors also observed that this inconsistency was not a barrier since the interaction detected within the group was simple. However, the use of the simple interaction became a problem since the result is given for each pair of environment, making it difficult to stratify the environments.

The environments clustering by the Lin's algorithm aims to allocate in the same group the environments that presented lack of interaction (Peluzio et al. 2012). According to Mendonça et al. (2007), this is a less selective environment clustering method than the simple interaction, which leads to differences between the methods.

Table 3. Environments clustering for plant height, lodging, and grain yield, according to the Lin (1982)'s algorithm. Environments not clustered in the groups were not allocated by the method

Trait	**Environments**	**MSe/r**	**F-calculated**	**Critical value of the F-distribution**
Plant height	4, 5, 3, 6 and 1	40.62	1.06	1.44
Lodging	4 , 5, 6, 2 and 3	3.69	0.77	1.44
Grain yield	2, 3, 1 and 5	182804	1.34	1.49
	2 and 6	184712	1.36	1.81
	4 and 5	203789	1.50	1.81
	3 and 4	242703	1.78	1.81

MSe/r = Mean Square of error/replications.

An explanation for the inconsistencies is that the Lin's algorithm estimates the sum of squares between genotypes and pair of environments to form the groups, while the lower value is used to form the initial group. The significance of the interaction between pairs of environments is tested considering the sum of square and the mean square of the GE interaction (Cruz et al. 2012). The division of the GE interaction proposed by Cruz and Castoldi (1991) decomposes the interaction component, without considering the significance, while Lin uses a variance ratio to cluster the environments.

In the BN, it is possible to observe the joint distribution and conditional dependence of a data set for prediction purposes. Thus, the information provided by the BN could be used to determine the predictive capacity of the environments, allowing environments clustering when this information is associated with those provided by simple or complex interaction between environments.

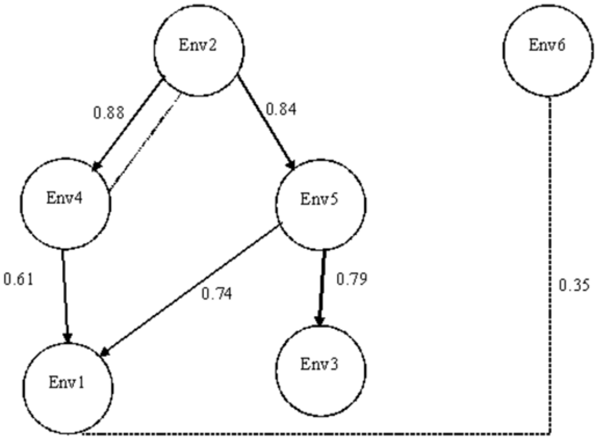

Figure 1. Directed Acyclic Graph representation considering six environments for plant height.

The BN for plant height detected the most similar pair of environments, 2 x 4, with almost 91% of simple interaction, according to the method proposed by Cruz and Castoldi (1991), and classified this pair as the most important (Figure 1). The representation of the DAG, besides the visualization of the relations between traits, allows the categorization of the parameters, since the most relevant environments are allocated in the upper part of the figure (Felipe et al. 2015).

The correlations between the connected environments were classified as intermediate or strong (Figure 1). The BN also predicted the highest correlation between environments 2 x 4 (r= 0.88), and showed no connection between the less correlated environments, 1 x 6 (r= 0.35).

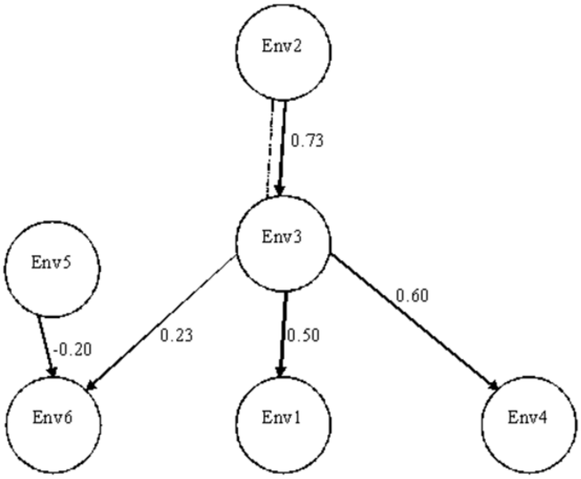

Figure 2. Directed Acyclic Graph representation considering six environments for lodging.

The complex interaction was predominant between pairs 1 x 2, 1 x 4, 1 x 6, 3 x 4, 3 x 6, 4 x 6 and 5 x 6 for plant height (Table 2), indicating that the environment 6 is the less similar, which was demonstrated by the DAG. The only discrepancy was the pair 1 x 4, which presented complex interaction of 51% and was connected by the BN. However, this could be explained by the lower magnitude of the complex interaction, which is almost close to 50%.

Although the BN for lodging was not efficient in discriminating the pair with the highest percentage of simple interaction (pair 1 x 4), it was able to demonstrate the most correlated environments (pair 2 x 3 (r = 0.73)). The correlations were classified as weak, intermediate or strong, and a discrepancy was observed between environments 5 x 6, for this pair presented the lowest correlation (r = -0.20) and complex interaction of 93.83%. Complex interaction was also detected between pairs 2 x 5, 2 x 6, 3 x 6, 4 x 5, and 4 x 6 (Table 2).

For grain yield, the pair with higher percentage of simple interaction in the DAG (Figure 3) was not identified (pair 2 x 5). The environmental correlations were classified as intermediate and strong, and the BN did not identify the highest correlation (pair 2 x 5), probably due to the proximity of the correlation values of this pair with the correlation values of the pair 1 x 3, which was indicated by the BN. Discrepancies were not observed, and therefore the pairs 1 x 4, 1 x 6, 2 x 6, and 3 x 6, which presented predominance of complex interaction, were not associated by the BN.

For plant height and lodging, environment 2 was considered as the most important in the DAG, while for grain yield, the same environment was considered as one of the least important. The similar results for plant height and lodging can be explained by the close association between these two traits (Shi et al. 2016). In the case of grain yield, despite the association with plant height and lodging, correlation is not always observed (Rafiq et al. 2010, Nzuve et al. 2014).

Considering all traits, the BN was efficient in indicating, directly or indirectly, the pair of environments with higher percentage of simple interaction for plant height and grain yield, and not efficient in the case of lodging, which can be explained by the high experimental error of this trait (Table 1). High coefficient of variation indicates high environmental variability (Keshavarzi et al. 2015), which decreases the predictive power of the BN. Therefore, the BN was effective in identifying similarities between environments for plant height and grain yield, and consequently facilitated the joint analysis of the environments *per se*. The BN could be advantageous to plant breeders since it allows using a great number of environments and does not require parity analysis.

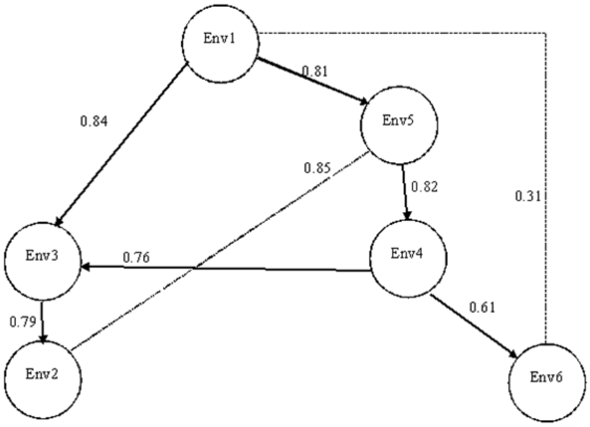

Figure 3. Directed Acyclic Graph representation considering six environments for grain yield.

CONCLUSIONS

The Bayesian network is efficient in connecting environments with predominance of simple interaction for traits with high experimental precision, while the Lin's method allocated in the same group environments with complex interaction. Therefore, the Bayesian network is a practical method to analyze an environmental net and detect similarity, without the need for the pairwise analysis of the environments. In addition, this method has the potential to cluster environments; however, the results must be associated with information on the type of interaction predominating between environments.

REFERENCES

Aliferis CF, Statnikov A, Tsamardinos I, Mani S and Koutsoukos XD (2010) Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. **Journal of Machine Learning Research 11**: 171-234.

Borsuk ME, Stow CA and Reckhow KH (2004) A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. **Ecological Modelling 2**: 219-239.

Cargnelutti Filho A and Storck L (2007) Estatísticas de avaliação da precisão experimental em ensaios de cultivares de milho. **Pesquisa Agropecuária Brasileira 1**: 17-24.

Carvalho FD, Lorencetti C and Benin G (2004) **Estimativas e implicações da correlação no melhoramento vegetal**. UFPel, Pelotas, 142p.

Cruz CD (2013) Genes: a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum Agronomy 3**: 271-276.

Cruz CD and Castoldi FL (1991) Decomposição da interação genótipos x ambientes em partes simples e complexa. **Revista Ceres 219**: 422-430.

Cruz CD, Regazzi AJ and Carneiro PCS (2012) **Modelos biométricos aplicados ao melhoramento genético**. UFV, Viçosa, 514p.

Fan XM, Kang MS, Chen H, Zhang Y, Tan J and Xu C (2007) Yield stability of maize hybrids evaluated in multi-environment trials in Yunnan, China. **Agronomy Journal 1**: 220-228.

Felipe VP, Silva MA, Valente BD and Rosa GJ (2015) Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. **Poultry Science 4**: 772-780.

Fritsche-Neto R, Vieira R.A, Scapim CA, Miranda GV and Rezende LM (2012) Updating the ranking of the coefficients of variation from maize experiments. **Acta Scientiarum. Agronomy 1**: 99-101.

Fornasieri Filho D (2007) **Manual da cultura do milho**. Funep, Jaboticabal, 576p.

Garbuglio DD, Gerage AC, Araújo PD, Fonseca Junior NDS and Shioga PS (2007) Análise de fatores e regressão bissegmentada em estudos de estratificação ambiental e adaptabilidade em milho. **Pesquisa Agropecuária Brasileira 2**: 183-191.

Keshavarzi A, Sarmadian F, Omran ESE and Iqbal M (2015) A neural network model for estimating soil phosphorus using terrain analysis. **The Egyptian Journal of Remote Sensing and Space Science 2**: 127-135.

Lin CS (1982) Grouping genotypes by a cluster method directly related to genotype-environment interaction mean square. **Theoretical and Applied Genetics 3:** 277-280.

Mansfield BD and Mumm RH (2014) Survey of plant density tolerance in US maize germplasm. **Crop Science 1**: 157-173.

Mendonça O, Pípolo VC, Garbuglio, DD and Fonseca Junior NDS (2007) Análise de fatores e estratificação ambiental na avaliação da adaptabilidade e estabilidade em soja. **Pesquisa Agropecuária Brasileira 11**: 1567-1575.

Nzuve F, Githiri S, Mukunya DM and Gethi J (2014) Genetic variability and correlation studies of grain yield and related agronomic traits in maize. **Journal of Agricultural Science 9**: 166.

Oliveira GH, Buzinaro R, Revolti L, Giorgenon CH, Charnai K, Resende D and Moro GV (2016) An accurate prediction of maize crosses using diallel analysis and best linear unbiased predictor (BLUP). **Chilean Journal of Agricultural Research 3**: 294-299.

Pearl J (2000) **Causality: models, reasoning and inference**. Cambridge University Press, New York, 685p.

Peluzio JM, de Deus Gerominni G, da Silva JPA, Afférri FS and Vendruscolo JBG (2012) Stratification and environmental dissimilarity for evaluation of soybean cultivars in the state of Tocantins. **Bioscience Journal 3**: 332-337.

Rafiq C, Rafique M, Hussain A and Altaf M (2010) Studies on heritability, correlation and path analysis in maize (*Zea mays* L.). **Journal of Agricultural Research 1**: 35-38.

Scutari M (2009) **Learning Bayesian networks with the bnlearn R package**. CRC Press, Flórida, 241p.

Semagn K, Magorokosho C, Ogugo V, Makumbi D and Warburton ML (2014) Genetic relationships and structure among open-pollinated maize varieties adapted to eastern and southern Africa using microsatellite markers. **Molecular Breeding 3**: 1423-1435.

Shi DY, Li YH, Zhang JW, Liu P, Zhao B and Dong ST (2016) Effects of plant density and nitrogen rate on lodging-related stalk traits of summer maize. **Plant Soil and Environment 7**: 299-306.