https://doi.org/10.1590/1678-4499.20200232

PLANT BREEDING
Article

Patterns recognition methods to study genotypic similarity in flood-irrigated rice

Antônio Carlos da Silva Júnior[1,*], Michele Jorge da Silva[1], Cosme Damião Cruz[1], Moyses Nascimento[2], Camila Ferreira Azevedo[2], Plínio César Soares[3]

1. Universidade Federal de Viçosa – Departamento de Biologia Geral – Viçosa (MG), Brazil.
2. Universidade Federal de Viçosa – Departamento de Estatística – Viçosa (MG), Brazil.
3. Empresa de Pesquisa Agropecuária de Minas Gerais – Viçosa (MG), Brazil.

ABSTRACT: Genetic diversity studies are performed based on information on a set of traits measured in a group of genotypes, considering one or more environments. The pattern recognition methods allow classifying genotypes from a set of important agronomic information. Thus, this study aimed to present and compare pattern recognition methods to inquire about the similarity of environments and genotypes in flood-irrigated rice for the recommendation of cultivars. The experiments were performed in the municipalities of Leopoldina, Lambari, and Janaúba, state of Minas Gerais, Brazil. To evaluate the pattern of similarity, 25 rice genotypes in three environments belonging to the flood-irrigated rice breeding program were used. Among these genotypes, five cultivars were used as an experimental control for the grain yield, the height of the plant, flowering, panicle length, grains filled by panicles, percentages of grains filled by panicles, in the 2012/2013 agricultural year. The methods used were mixtures of multivariate normal distributions and density-based clustering algorithm. It was observed, therefore, that the genotypes are distributed in three distinct groups, in which there are intragroup homogeneity and intergroup heterogeneity for the agronomic traits of the flooded rice culture. The methods used to assess the dissimilarity of environments using pattern recognition methods were efficient in classifying flooded rice irrigated environments.

Key words: classification, dissimilarity, environments, *Oryza sativa* L.

Received:
May 19, 2020

Accepted:
Jun. 8, 2020

Section Editor:
Carlos Scapim

*Corresponding author:
antonio.silva.c.junior@gmail.com

INTRODUCTION

Rice (*Oryza sativa* L.) is one of the most produced and consumed cereals in the world, and is characterized as the main food for more than half of the world population. With the increase in the population, the demand for grain productivity has increased over the years and it is estimated that by 2050 global rice production should increase from 60 to 110% to supply the demand of the world population (Godfray et al. 2010; Tilman et al. 2011; Ray et al. 2013; Santos et al. 2019).

Genetic diversity studies are performed using information from a set of traits measured in a group of genotypes, considering one or more environments. These studies are useful for recognizing similarity patterns and quantifying variability to explore breeding plants. The similarity pattern is generally attributed to genetic similarity by ancestry or by sharing alleles in common, which are fixed by selection.

When performed experiments in more than environment, an approach to the behavior of genotypes are common, emphasizing stability and adaptability for a given characteristic of agronomic importance, mainly grain production. Also, the

cc BY

study of dissimilarity of environments is equally important and aims to identify more discriminative and representative environments to subsequently analyze the most stable and adapted genotypes.

One way of evaluating the behavior of genotypes, given the dissimilarity of environments, but little explored among breeders, is the use of approaches based on pattern recognition methods. In this case, the evaluation of a set of traits relevant to the breeder and the influence of the environments on a possible pattern of grouping of the evaluated genotypes are considered. Thus, it is assumed that genotypes can be differentiated by genetic causes, within the environment, and by environmental causes, called macrovariations, provided by the edaphoclimatic differences to which they were subjected. The pattern recognition methods allow classifying objects, within many categories or classes expressed by the environments, from a set of important agronomic information (Bishop 2006).

Pattern recognition between genotypes provided by the dissimilarity of environments allows the breeder to make decisions to identify groups of environments in which the interaction genotype × environments (G × E) may not be significant for the set of available genotypes.

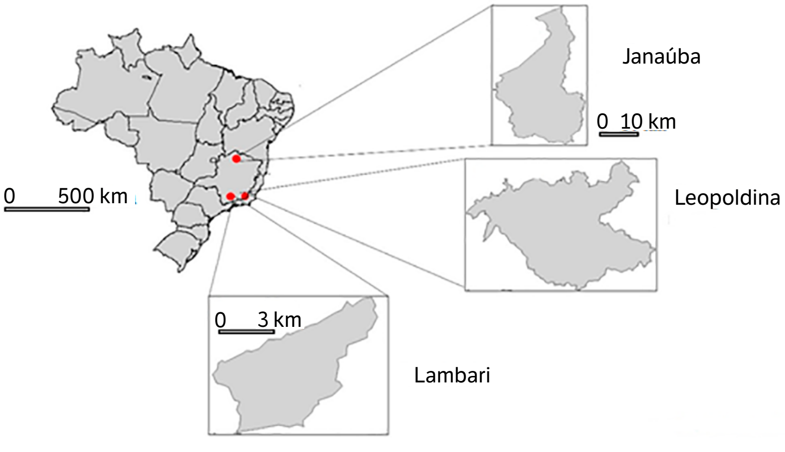
Thus, this study aimed to use pattern recognition methods (mixture of normal distributions and density-based clustering algorithm) to study the similarity of environments and genotypes in flood-irrigated rice for the recommendation of cultivars.

MATERIAL AND METHODS

Description of the experiments

The experiments were conducted in the state of Minas Gerais, Brazil, in the experimental field of the Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG), in the municipalities of Leopoldina (latitude 21° 31' 48.01" S, longitude 42° 38' 24" W), Lambari (latitude 21° 58 '11.24" S, longitude 45° 20' 59.6" W) and Janaúba (latitude 15° 48' 77" S, longitude 43° 17' 59.09" W) (Fig. 1).

To inquire about the pattern similarity, 25 rice genotypes were evaluated in three environments belonging to the flood-irrigated rice breeding program. Among these genotypes, five cultivars were used as experimental controls ('Rubelita', 'Seleta', 'Ourominas', 'Predileta', and 'Rio Grande') for the following traits: grain yield (Kg·ha^{-1}), height of plant (cm), flowering (days), panicle length (cm), grains filled by panicles, percentages of grains filled by panicles, in the 2012/2013 agricultural year. The experimental design used in all experiments was randomized blocks with three replications. The value for cultivation and use (VCU) tests were conducted on floodplain soils with continuous flood irrigation. The cultural treatments were carried out according to the recommended for the cultivation of irrigated rice in the evaluated regions (Soares et al. 2005).



Fonte: https://gadm.org/maps.html

Figure 1. Location of experiments in three regions of Minas Gerais, Brazil.

In the Leopoldina experimental field, the seedlings were previously formed in nurseries and, later, transplanted at a spacing of 0.20 m on the line. In other, sowing was carried out on the planting line with a density of 300 seeds·m^{-2}. The tests were carried out in floodplain soils with continuous flood irrigation. Irrigation started around 10 to 15 days after seedling emergence, in the case of planting with seeds, or when the seedlings were established in the soil. The irrigation depth was gradually increased according to the development of the plants.

Statistical analysis

The statistical model described in Eq. 1 was considered to each original observation:

$$Y_{ijl} = \mu + B/E_{jl} + G_i + E_j + GE_{ij} + e_{ijl} \tag{1}$$

where: Y_{ijl} is the observation in the lth block, evaluated in the ith genotype and jth environment; μ is the general average of the experiments; B/E_{jl} is the effect of block l within environment j; G_i is the effect of the ith genotype (i = 1,2 ... g); E_j is the effect of the jth environment (j = 1,2 ... k); GE_{ij} is the random effect of the interaction between genotype i and environment j; e_{ijl} is the random error associated with the Y_{ijl} observation.

The pattern recognition analysis was made from adjusted values ($Y^*_{ijk} =\hat{\mu} + \hat{G}_i + \hat{e}_{ijl}$), which adjust the phenotypic values for the effects of block, environment and the interaction of genotypes and environments. With the adjusted value, pattern recognition analyzes were performed using mixtures of multivariate normal distributions and density-based clustering algorithm.

Mixtures of multivariate normal distributions

In this analysis, it was considered that there are, in the set of environments, homogeneous subgroups whose data could be characterized by probability distributions, supposedly normal. As a whole, a multivariate normal distribution is assigned to each component of the mixture. Thus, it is expected that the clusters represent sets of environments with an ellipsoidal data arrangement, centered on the mean vector μ_k, and with other geometric characteristics, such as volume, orientation, and shape, determined by the covariance matrix Σ_k of dimension v × v.

Parsimonious parameterizations of the covariance matrices, for each environment group, can be obtained through Eq. 2:

$$\Sigma_{k=} \lambda_k D_k A_k D_k^T \tag{2}$$

where: λ_k is a scalar that controls the volume of the ellipsoid, A_k is a diagonal matrix that specifies the shape of the density contours with $det\ (A_k) = 1$ and D_k is an orthogonal matrix that determines the orientation of the corresponding ellipsoid (Banfield and Raftery 1993; Celeux and Govaert 1995). In one dimension, there are only two models denoted by *E* for equal variance and *V* for a variable variance. In the multivariate configuration, the volume, shape, and orientation of covariance can be limited to being equal or variable between groups. Thus, 14 possible models with different geometric characteristics can be specified. Table 1 presents all of these models with the corresponding distribution structure type, volume, shape, orientation, and associated model names.

The model is chosen using the Bayesian information criterion (BIC) (Scrucca et al. 2016), according to Eq. 3:

$$BIC_{M,k} = 2l(\hat{\Psi};y)\text{-}\nu_{M,k}\log(n), \tag{3}$$

where, $l(\hat{\Psi};y)$ is the logarithmic of the maximized likelihood function (Supplemental Material available); $\nu_{M,k}$ is the number of independent parameters to be estimated in the model *M*; and *k* is the number of components in the mixture, supposedly equal to the number of environments analyzed. According to the BIC expression presented by Scrucca et al. (2016), the higher the BIC value, the better the model.

Table 1. Parameterizations of the intragroup covariance matrix Σ_k for multidimensional data and the corresponding geometric characteristics.

Model	Σ_k	Distribution	Volume	Shape	Orientation
EII	λI	Spherical	Equal	Equal	-
VII	$\lambda_k I$	Spherical	Variable	Equal	-
EEI	λA	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k A$	Diagonal	Variable	Equal	Coordinate axes
EVI	λA_k	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k A_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	λDAD^T	Ellipsoidal	Equal	Equal	Equal
EVE	λDA_kD^T	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k DAD^T$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k DA_kD^T$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda_k DAD_k^T$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k D_kAD_k^T$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda D_kA_kD_k^T$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k D_kA_kD_k^T$	Ellipsoidal	Variable	Variable	Variable

Density-based clustering algorithm

The density-based clustering algorithm to discover the number of clusters (environments) was created to identify different forms of groupings and the presence of noise in the databases (Ester et al. 1996)[4]. It uses the concept of center-based density since the density of a point in the data set is the number of points within a neighborhood radius. This algorithm contains two input parameters, the radius and the minimum number of points in a given radius. However, center-based density makes it possible to classify a point in dense regions (center point), at the limit of a dense region (limit point) or in a sporadically occupied region (noise point).

To evaluate the classification performance objectively, the confusion matrix was used, in which the frequency observed on the diagonal represents the elements correctly classified. The marginal column represents the total of elements classified for a category *i*. On the other hand, the marginal line represents the total of reference elements sampled for a category *i*.

The GENES software (Cruz 2016) was used to perform the analyses, integrated with Matlab (Matlab 2011) and R (R Core Team 2019).

RESULTS AND DISCUSSION

Table 2 shows the result of the joint analysis of variance of 25 rice genotypes evaluated in three environments. The estimate of the coefficient of variation (CV%) was low for all characteristics, indicating adequate experimental precision, as demonstrated in other studies related to the culture of irrigated rice (Hosan et al. 2010; Silva et al. 2011; 2019; 2020; Costa et al. 2002; Streck et al. 2017; Santos et al. 2019). For the effect of genotypes, there was statistical significance for the traits of grains filled by panicles and the percentage of grains filled by panicles, and no significance was observed for the effect of genotypes in the joint analysis for the other traits. The difficulty in detecting differences between the general means of such genotypes can be justified by the advanced stage of genetic improvement in which these genotypes are found for these traits. There was significance ($p < 0.01$) for the effects of the environment, except for grains filled by panicles, and for the genotype interaction by environments (G × E), except for the height of plant and percentage of grains filled by panicles. Consequently, the behavior of the genotypes was influenced by environmental conditions, justifying the use of methodologies that are capable of classifying environments according to clustering methods.

[4] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland: AAAI.

Table 2. Result of the joint variance analysis for grain yield (GY), height of plant (HP), flowering (FLO), panicle length (PL), grains filled by panicles (GP), percentages of grains filled by panicles (GPO) of the 25 flood-irrigated rice genotypes evaluated in three environments in the state of Minas Gerais.

SV	DF	Mean Square					
		GY	HP	FLO	PL	GP	PGP
Genotype	24	682447^{ns}	237^{ns}	65.91^{ns}	8.11^{ns}	477*	0.00614**
Environment	2	49408974**	15951**	13556**	6.48**	3447^{ns}	0.01265**
G×E	48	1384692**	315^{ns}	51.03 **	8.84**	272**	0.00114^{ns}
Residue	144	419169	311	3.41	2.34	160	0.00161
CV(%)		14.26	18.70	2.01	6.49	13.03	4.34
Average		4539	94.35	91.94	23.59	97.10	0.88

**, *, ns: significant at 1%, 5% and not significant by the F test; SV: source of variation; DF: degrees of freedom; CV: coefficient of variation.

Among all the adjusted models presented in Table 1, the two that presented the highest Bayesian information criterion (BIC) values were VEI (diagonal distribution, variable volume, equal shape, and coordinate axis orientation; BIC = -2901.55) and VVI (diagonal distribution, variable volume, variable shape and the orientation of coordinate axes; BIC = -2918.59), associated with five and three components of mixtures, respectively (Fig. 2).

Table 3 shows the number of genotypes allocated to each of the five and three components of mixtures considering, respectively, the VEI and VVI models.

The model considering a mixture with three components allocated approximately 25 genotypes in each component. This result is interesting since, due to the edaphoclimatic differences in each location, such as temperature and humidity, it is expected to obtain a mixture composed of three components. On the other hand, the mixture model composed of five components divided the genotypes into two other groups.

Therefore, Table 4 shows a confusing matrix of classification of genotypes in the different environments, which obtained a prediction accuracy of 97.33% (representing the number of correct classifications on the total genotypes). In this table, environment 1 obtained 100% classification of the 25 genotypes, while environments 2 and 3 presented an error when classifying the genotypes in their respective environments.

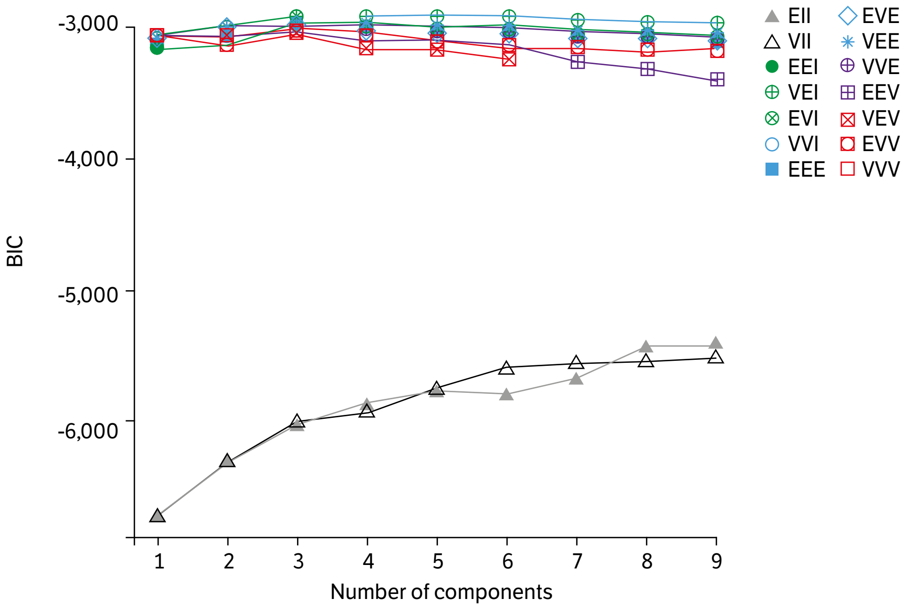


Figure 2. Bayesian information criterion (BIC) of the adjusted models considering different numbers of components of mixtures.

Table 3. The number of genotypes allocated to each of the five and three components of mixtures considering the models VEI and VVI, respectively.

VEI model with five components		VVI model with three components	
Components	**NC**	**Components**	**NC**
1	11	1	25
2	9	2	26
3	25	3	24
4	13		
5	17		

NC: number of components; VEI: diagonal distribution, variable volume, equal shape, and coordinate axis orientation; VVI: diagonal distribution, variable volume, and variable shape and the orientation of coordinate axes.

Table 4. Confusion matrix of classification of 25 genotypes in the different environments.

Environment	1	2	3
1	25	0	0
2	0	24	1
3	1	0	24

Figure 3 shows the density-based clustering algorithm to identify the number of clusters. It was possible to observe three different groups of classification of environments. However, based on center density, it was possible to classify a point in dense regions (center point), at the limit of the region or in a sporadically occupied region (noise point).

The environment can be classified as favorable or unfavorable depending on the conditions in which it is found, thus being able to influence the classification of the genotypes. Therefore, the favorable environment corresponds to all the conditions that a given gene has to express the desirable characteristics since the genotypes perform better in these environmental conditions. Another issue that must be taken into account is the minimization of the response to uncontrollable factors since the breeder aims to produce cultivars with greater capacity for genetic resilience and more responsiveness. To the unfavorable environment, environmental conditions do not provide an expected performance, that is, a given gene is not expressed depending on the conditions in which it is exposed. For example, the genotype interaction by the environment,

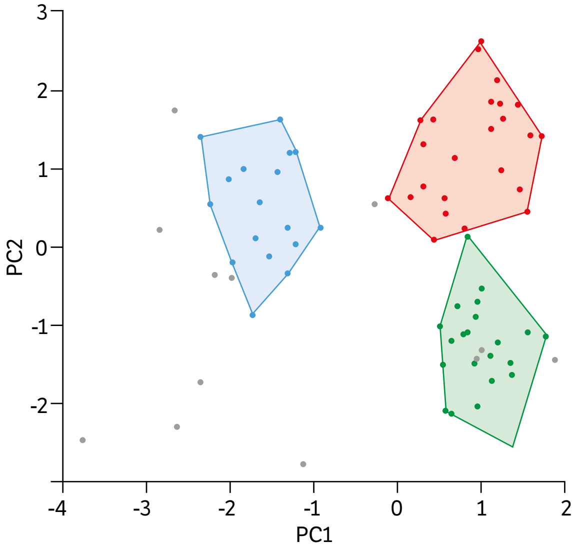


Figure 3. Density-based clustering algorithm.

in which the different behavior of the genotypes in the face of environmental variations is observed. In this case, it makes it difficult to decide to recommend a specific cultivar.

In this context, based on the results obtained, one can consider only one of the environments in the next evaluations, and the breeder should choose the best environment for the needs of his flood-irrigated rice breeding program. Criteria such as proximity to the research center and ease of access can be adopted. Also, the decrease in the number of environments will reduce the cost of evaluating the cultivars, in addition to allowing more judicious evaluations to be carried out in the remaining trials. Thus, the methods used to assess the dissimilarity of environments through pattern recognition methods provided a better classification between environments.

CONCLUSION

The methods used to assess the dissimilarity of environments using pattern recognition methods were efficient in classifying flooded rice environments.

ACKNOWLEDGMENTS

The authors thank the researchers at the Embrapa Rice and Beans Dr. Orlando Peixoto de Morais (in memory) and Paula Pereira Torga.

FUNDERS

Conselho Nacional de Desenvolvimento Científico e Tecnológico
[http://doi.org/10.13039/501100003593]

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
[http://doi.org/10.13039/501100002322]

Fundação de Amparo à Pesquisa do Estado de Minas Gerais
[http://doi.org/10.13039/501100004901]

AUTHOR'S CONTRIBUTION

Conceptualization, Silva Júnior A. C., Cruz C. D., Nascimento M. and Azevedo C. F.; Methodology, Silva Júnior A. C., Cruz C. D., Nascimento M. and Azevedo C. F.; Investigation, Silva Júnior A. C.; Cruz C. D.; Nascimento M.; Azevedo C. F. and Silva M. J.; Writing – Original Draft, Silva Júnior A. C.; Cruz C. D.; Nascimento M. and Silva M. J.; Writing – Review and Editing, Silva Júnior A. C.; Cruz C. D.; Nascimento M. and Silva M. J.; Funding Acquisition, Soares P. C.; Resources, Soares P. C.; Supervision, Silva Júnior A. C.; Cruz C. D.; Nascimento M. and Silva M. J.

REFERENCES

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803-821. https://doi.org/10.2307/2532201

Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer-Verlag.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern Recognition 28, 781-793. https://doi.org/10.1016/0031-3203(94)00125-6

Costa, N. H. A. D., Seraphin, J. C. and Zimmermann, F. J. P. (2002). Novo método de classificação de coeficientes de variação para a cultura do arroz de terras altas. Pesquisa Agropecuária Brasileira, 37, 243-249. https://doi.org/10.1590/S0100 204X2002000300003

Cruz, C. D. (2016). Genes Software – extended and integrated with the R, Matlab and Selegen. Acta Scientiarum. Agronomy, 38, 547-552. https://doi.org/10.4025/actasciagron.v38i3.32629

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M. and Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. Science, 327, 812-818. https://doi.org/10.1126/science.1185383

Hosan, S. M., Sultana, N., Iftekharudduala, K. M., Ahamed, N. U. and Mia, S. (2010). Genetic divergence in landraces of Bangladesh rice (*Oryza sativa* L.). The Agriculturists, 8, 28-34. https://doi.org/10.3329/agric.v8i2.7574

Matlab (2010). Matlab Version 7.10.0. Natick, Massachusetts: The Math Works Inc.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Accessed Mar. 7 2020]. Available at: https://www.R-project.org/.

Ray, D. K., Mueller, N. D., West, P. C. and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. PLoS ONE, 8, e66428. https://doi.org/10.1371/journal.pone.0066428

Santos, I. G., Carneiro, V. Q., Silva Junior, A. C., Cruz, C. D. and Soares, P. C. (2019). Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. Acta Scientiarum. Agronomy, 41, e39803. https://doi.org/10.4025/actasciagron.v41i1.39803

Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2016). Mclust 5: clustering, classification and density estimation using gaussian finite mixture models. The R Journal, 8, 289-317. https://doi.org/10.32614/RJ-2016-021

Silva, E. F., Silva, V. A. C, Guimarães, J. F. R. and Moura, R. R. (2011). Divergência fenotípica entre genótipos de arroz de terras altas. Revista Brasileira de Ciências Agrárias, 6, 280-286. https://doi.org/10.5039/agraria.v6i2a1183

Silva, G. N., Silva Júnior, A. C., Sant'Anna, I. C., Cruz, C. D., Nascimento, M. and Soares, P. C. (2019). Projeção de distâncias como método auxiliar na classificação de arroz irrigado quanto a adaptabilidade e estabilidade. Revista Brasileira Biometria, 37, 229-243. https://doi.org/10.28951/rbb.v37i2.383

Silva, G. N., Silva Júnior, A. C., Sant'Anna, I. C., Cruz, C. D., Nascimento, M. and Soares, P. C. (2020). Similarity networks for the classification of rice genotypes as to adaptability and stability. Pesquisa Agropecuária Brasileira, 55, e01017. https://doi.org/10.1590/s1678-3921.pab2020.v55.01017

Soares, P. C., Melo, P. G. S, Melo, L. C. and Soares, A. A. (2005). Genetic gain in an improvement program of irrigated rice in Minas Gerais. Crop Breeding and Applied Biotechnology, 5, 142-148. [Accessed Mar. 7 2020]. Available at: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/24014/1/bd6b8337-4583-a49e.pdf

Streck, E. A., Aguiar, G. A., Magalhaes Júnior, A. M., Facchinello, H. K. and Oliveira, A. C. (2017). Variabilidade fenotípica de genótipos de arroz irrigado via análise multivariada. Revista Ciência Agronômica, 48, 101-109. https://doi.org/10.5935/1806-6690.20170011

Tilman, D., Balzer, C., Hill, J. and Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. Proceedings of the National Academy of Sciences of the United States of America, 108, 20260-20264. https://doi.org/10.1073/pnas.1116437108