International Journal of Agriculture, Environment and Bioresearch

Vol. 08, No. 01; 2023

ISSN: 2456-8643

PRINCIPAL COMPONENT ANALYSIS AND CLASSIFICATION OF SUGARCANE PRODUCTION DATA IN BURUNDI

BARANKANIRA Emmanuel[1,2*], NDIKUMANA Valentin[3] and NITUNGA Benjamin[3]

[1]Burundi Higher Institute of Education, Department of Natural Sciences, P.O Box 6983, Bujumbura, Burundi

[2]Sciences Research and Professional Development Center, P.O Box 6983, Bujumbura, Burundi

[3]Lake Tanganyika University, Department of Statistics, P.O Box 5403, Bujumbura, Burundi

https://doi.org/10.35410/IJAEB.2023.5796

ABSTRACT

This study aims at identifying groupings of individuals (years) from sugarcane production data in Burundi by applying Principal Component Analysis (PCA) and classification, and computing clusters' distances between centroids. These data were obtained from the Ministry of Environment, Agriculture and Livestock in Burundi and cover a period from 2000 to 2019, i.e. 20 years. The variables used are mean sugarcane yield, sugar production, cultivated area, sugarcane production, molasses production, herbicide inputs, mean temperature and cost of sugar production. R software version 3.6.1 was used to analyze data. This study shows that the years 2002 and 2013 are opposed by the first factorial axis while the years 2001 and 2017 are opposed by the second one. The years 2000, 2002, 2008, 2012, 2013 and 2017 are better represented on the principal factorial plane. The years 2017 and 2019 are characterized by high values of herbicide inputs and cultivated area while the years 2011, 2012 and 2013 are characterized by high values of mean temperature and mean sugarcane yield. The year 2015 is characterized by high values of sugar production, sugarcane production, cost of sugar production and molasses production.

Keywords: Principal Component Analysis, classification, sugarcane, Burundi.

1. INTRODUCTION

Sugarcane (*saccharum officinarum l.*) is an important crop insofar as it is a real economic, social and environmental issue in the world in general and in sub-Saharan African countries in particular [1]–[5]. In Burundi, agriculture is the pillar on which economy relies and the most important crop is coffee [6]. Sugar is produced by the Moso Sugar Company (SOSUMO, Société Sucrière de Moso, *in French*) and is used for consumption (meals, tea, coffee, juices) in households and during celebrations, for feeding livestock, for the manufacture of drinks by industries, and sugarcane for erosion control [7]. Historically, SOSUMO was created in 1982 to contribute to the development of the natural region of Moso. It is located on Gihofi hill in Bukemba commune, itself being in Rutana province in the southern of the country. Sugar produced by SOSUMO is well appreciated and consumed not only by Burundians but also by Great Lakes countries such Tanzania, Rwanda and Democratic Republic of Congo.

Sugarcane is also grown for its stalks, which contain a sweet juice from which sucrose or crystallizable sugar is extracted. The sugarcane cycle, its growth, maturation, production and yield are closely conditioned by climate change [4], [8], [9]. Sugarcane needs climatic conditions

Vol. 08, No. 01; 2023

ISSN: 2456-8643

such as heat, water and light, and is better adapted to high altitudes [10], [11]. Drought and cold nights, on the other hand, favour ripening.

To the best of our knowledge, there is no study that has focused on multidimensional positioning and classification of sugarcane data in Burundi. The objective of this cross-sectional study is to reduce the dimensionality of the data to a lower dimensional space while minimizing the loss of information and detecting possible groupings of years. In other words, the aim is to study the proximity between individuals (years), to detect variables on which the possible similarities or dissimilarities are based and the relationships between these variables.

2. MATERIALS AND METHODS

Source of data

Data were collected from the statistical yearbooks of the National Agricultural Surveys of Burundi and cover a period from 2000 to 2019 [12]. These data are from the Price Information System and the Agricultural Season Monitoring Information System of the Ministry of Environment, Agriculture and Livestock from 2000 to 2019.

These annual data are produced by National Institute of Statistics of Burundi (former Institute of Statistics and Economic Studies of Burundi). They include mean sugarcane yield (MEANSCY) in tons per hectare, cultivated area (CULTAREA) in hectares, sugarcane production (SCPROD) in tons, sugar production (SPROD) in tons, molasses production (MOPROD) in kilograms, herbicide inputs (HERBINP) in kilograms, mean temperature (MEANTEMP) in millimeters, and the cost of sugar production (COSTSPROD) in Burundian Francs. The Burundi National Agricultural Surveys are conducted since 2012 at two levels (enumeration area and households) in all provinces except Bujumbura-Mairie (the capital city) which is completely urban. They concern the three agricultural seasons, namely season A or *Agatasi* (in Kirundi) from 16 September to 15 February of the following year, season B or *Impeshi* (in Kirundi) from 16 February to 15 June and season C or *Ici* (in Kirundi) from 16 June to 15 September [13]. In this study, the unit of analysis is therefore the year. Data analysis was performed from May to September 2022.

Principal Component Analysis

The space in which we live and evolve is a three-dimensional space, and any higher dimensional space is beyond human reach. Thus, data visualization is most often done in a two- or three-dimensional space. It is, however, usual to add a fourth dimension (time) to have a space-time. In order to have a graphical representation of the data in a reduced space while losing as little information as possible, Principal Component Analysis (PCA) is used [14]–[16]. It is a descriptive technique of multidimensional positioning or reduction of the dimensionality of the data which makes it possible to detect groupings of observations (groups, classes, clusters) and to remove the redundant information brought by the correlated variables to obtain the true dimension of the data.

Vol. 08, No. 01; 2023

ISSN: 2456-8643

The data matrix consists of *n* individuals on which *p* variables have been measured. In other words, the objective of PCA is to find a lower dimensional space (hyper-plane) in which it is possible to represent the *n* individuals and/or the *p* variables by their coordinates on a reduced number *r* of new variables ($r<p$) or principal components. These principal components, uncorrelated, are linear combinations of the original variables and allow as much information as possible to be restored.

The first step in PCA is therefore to study the correlation between the variables through the correlation matrix and the scatter plot matrix [17]. PCA first looks for the combination of variables that best visualizes the individuals, i.e. the one for which the variance is maximum. This combination is called first principal component and it maximizes the overall spread of the points in the center of gravity of the scatter plot. A second one, which is orthogonal to the first one, is sought so that it maximizes the inertia unexplained by the first one, and so on.

Links between variables and similarities/dissimilarities between individuals are explored. The individuals are projected in $\mathbb{R}^p$ space and the variables in the $\mathbb{R}^n$ space, this to mean that each individual is a *p*-component vector and each variable is an *n*-component vector.

The squared Euclidean distance between two individuals i and i' is computed as follows [18]:

$$d^2\left(i,i'\right)=\sum_{j=1}^{p}\left(x_i^j-x_{i'}^j\right)^2=\left\|x_i-x_{i'}\right\|^2 \tag{1}$$

Some variables may be much more dispersed (too much variance) or measured at very different scales. Variables with large observed values relative to others will be disproportionately important in determining the first components. To remedy this situation, all variables should be scaled before analysis to make them comparable.

Thus, instead of working on the values x_{ij} for a given individual *i* and a given variable *j*, the variables should be normalized to obtain the standardized variables z^j whose values are given by [17]:

$$z_i^j=\frac{x_i^j-\bar{x}^j}{s_j} \tag{2}$$

z_i^j being the values of the centered-reduced variable z^j, $\bar{x}^j$ the mean of the j^{th} variable x^j and s_j its standard deviation. This means that the origin of axes is moved to the center of gravity g (mean point) of the individuals scatter plot. The metric used $M = D_{\frac{1}{s^2}}$ is a diagonal matrix whose elements are the inverses of the variances of the variables. This metric, which is applied to centerd data, has the advantage of giving the same importance to the variables and of making the inter-individual distances dimensionless. The square of the distance between two individuals i and i' is then weighted by the inverse of the variance of each variable x^j in the form:

$$d^2\left(i,i'\right) = \sum_{j=1}^{p} \frac{1}{S_j^2}\left(x_i^j - x_{i'}^j\right)^2 = \left\|z_i - z_{i'}\right\|^2 \tag{3}$$

The square of the linear correlation coefficient between two variables x^j and x^k is also the square of the cosine of the angle between these variables:

$$r^2\left(x^j,x^k\right) = cos^2\left(x^j,x^k\right) = \frac{s_{jk}^2}{s_j^2 s_k^2} \tag{4}$$

s_{jk} being the covariance between the variables x^j and x^k. The squared distance between those variables is given by:

$$d^2\left(x^j,x^k\right) = 2\left[1 - r\left(x^j,x^k\right)\right] \tag{5}$$

The coordinate of each variable on each factorial axis is the linear correlation coefficient between this variable and the considered factorial axis. Each principal component c_j is a new variable obtained after projection of rows of a centered-reduced data matrix Z on the director-vector u_j of the corresponding factorial axis:

$$c_j = Zu_j \tag{6}$$

This can be written as an equations system:

$$\begin{cases} c_1 = a_{11}z_1 + a_{21}z_2 + a_{31}z_3 + \cdots + a_{p1}z_p \\ c_2 = a_{12}z_1 + a_{22}z_2 + a_{32}z_3 + \cdots + a_{p2}z_p \\ \\ c_k = a_{1j}z_1 + a_{2j}z_2 + a_{3j}z_3 + \cdots + a_{pj}z_p \\ \\ c_p = a_{1p}z_1 + a_{2p}z_2 + a_{3p}z_3 + \cdots + a_{pp}z_p \end{cases} \quad (7)$$

The vectors u_1 and u_2 are unit and orthogonal. The third principal component c_3 is obtained by projecting rows of Z on the vector orthogonal to the plan (u_1, u_2), and so on. The vector c_j (principal component) contains the coordinates of individuals on the factorial axis supporting the eigen vector u_j and $Var(c_j) = \lambda_j$ is the eigenvalue (axis inertia) corresponding to the eigen vector u_j. The eigenvalues equation is $det|R - \lambda I| = 0$ where R denotes the correlations matrix, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)^t$ the vector of eigenvalues with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

The inertia of the scatter plot to be maximized is the weighted sum of squared distances between individuals and the center of gravity g of the scatter plot:

$$I_p = \sum_{i=1}^{n} p_i d^2(i,g) \quad (8)$$

given $p_i = \frac{1}{n}$ and $\sum_{i=1}^{n} p_i = 1$. This total inertia, which is to be minimized, is the sum of inertia of projected scatter plot and the lost inertia (weighted sum of the squared distances between individuals-points and their projections):

$$I_p = \sum_{i=1}^{n} p_i d^2(i,g) = \sum_{i=1}^{n} p_i d^2(c_j,g) + \sum_{i=1}^{n} p_i d^2(i,c_j) \quad (9)$$

PCA is then applied to the correlations matrix. Hence, the total inertia equals the number of variables. Variance-covariance matrix $X^t X$ becomes then the correlations R matrix. This matrix gives eigenvalues (second stage) and each eigenvalue represents the inertia of the corresponding axis. The sum of those eigenvalues (total inertia) equals p (sum of diagonal elements of the correlations matrix). The contribution c_{yk} of each new variable x^k in the total variation in terms of percentage of inertia explained by the factorial axis n° k is:

Vol. 08, No. 01; 2023

ISSN: 2456-8643

$$c_{yk} = \frac{\lambda_k}{\sum_k \lambda_k} \times 100 \quad (10)$$

λ_k being the eigenvalue corresponding to the k^{th} factorial axis. The inertia rate or percentage of the variance explained by the r first factorial axes (mostly $r=2$) is:

$$\tau_r = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{q} \lambda_i} \quad (11)$$

The number of factorial axes used to project individuals and/or variables can be chosen visually [4]. There are three criterions to choose the number of factorial axes: elbow criteria or Cattell criteria that consists of observing a setback followed by a regular decreasing of eigenvalues (the selected axes are the ones that are before the setback), the Kaiser criterion or the Kaiser-Guttman criterion that consists of retaining axes whose inertia is higher than 1 and Karlis-Saporta-Spinalir criterion that consists of retaining axes for which:

$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}} \quad (12)$$

The eigenvalues λ are compared using Anderson confidence interval:

$$\lambda_i \, exp\left(-u_\alpha \sqrt{\frac{2}{n-1}}\right) < \lambda < \lambda_i \, exp\left(u_\alpha \sqrt{\frac{2}{n-1}}\right) \quad (13)$$

u_α being the Student quantile, n the number of observations and α the significance level. It is interesting to project individuals in a hyper-plane of the r first factorial axes ($r<p$) that explain between 80% and 85% of the total inertia. Individuals that contribute more to the formation of a factorial axis are naturally the ones whose coordinates are high (in absolute values) or distant to the gravity center.

The relative contribution of an individual i to the formation of the principal axis k is given by:

$$CTR(i,k) = \frac{x_{ik}^2}{\lambda_k} \quad (14)$$

It is important to check if each of the projected points on an axis (or a plan) is close to it. That is why the cosine value of the angle α between a point and its projection on the axis (or the plan) is

Vol. 08, No. 01; 2023

ISSN: 2456-8643

examined. Hence, the quality of representation of the individual i by the axis k, also called "absolute contribution of the axis k to the representation of the individual i" is given by the squared cosine:

$$AC(i,k) = \frac{x_{ik}^2}{\sum_j z_{ij}^2} \quad (15)$$

The contribution of a variable j to the constitution of a principal axis k is obtained using the formula:

$$CV(j,k) = u_{jk}^2 \quad (16)$$

The correlation between the variable j and the axis k is:

$$r_{jk} = u_{jk} \frac{\sqrt{\lambda_k}}{s_j} \quad (17)$$

Using correlations between variables and principal axes, it is possible to interpret components in terms of variables. For this, a correlation circle is drawn for variables. Projection of individuals is done on the plane given by the two first factorial axes retained. Moreover, the biplot of individuals and variables is done to facilitate simultaneous interpretation [14].

Classification

In this study, data classification was done using results of principal component analysis. Classification methods allow to group similar or dissimilar individuals in clusters using a metric. In hierarchical classification, the number of clusters in unknown and this classification leads to a classification tree called *dendrogram* obtained using an Euclidian distance [2], [19]. Besides, hierarchical descending classification assumes that all individuals are in one cluster. This cluster is then divided into two, three, and so on, clusters so that at the end n clusters are formed.

In this study, hierarchical ascending classification consists of grouping two similar individuals at the first stage to form a cluster. At the second stage, this cluster is then grouped with an individual close to the centroid of this cluster and so on using Ward's method. The total inertia I_T is then the sum of between and within clusters inertia. It is defined not only as the sum of the squares of distances between each individual and the scatter plot center of gravity g weighted by the inverse of the sample size n but also as the sum of variances or diagonal elements of the variance-covariance matrix:

Vol. 08, No. 01; 2023

ISSN: 2456-8643

$$I_T = \sum_{i=1}^{n} \frac{1}{n} d^2\left(e_i - g\right) \tag{18}$$

On the other hand, the within cluster inertia I_W to be maximized is the weighted sum of squared distances between each individual and its projection weighted by the inverse of the sample size whereas the between cluster inertia I_B to be minimized is the weighted sum of squared distances between clusters centers of gravity and the scatter plot center of gravity:

$$I_W = \sum_{i=1}^{n} \frac{1}{n} d^2\left(e_i - g_i\right) \tag{19}$$

$$I_B = \sum_{i=1}^{n} \frac{1}{n} d^2\left(g_i - g\right) \tag{20}$$

This leads to the Huygens' decomposition [20]:

$$I_T = I_W + I_B \tag{21}$$

This means that individuals of the same cluster must be as similar as possible and individuals of different two clusters as different as possible. According to Ward's algorithm, the weighted matrix of the squared distances between centers of clusters G_i and G_j is computed as follows:

$$W = \frac{p_i p_j}{p_i + p_j} d^2\left(G_i, G_j\right) \tag{22}$$

At the first stage, the distance is computed according to the relation (3). At the second stage, two nearest individuals or clusters are aggregated. The center of the new cluster is computed and the sum of the weights of individuals is attributed to this center. At the last stage, Ward's difference between individuals or clusters aggregated is computed in terms of loss of between cluster inertia or of gained within cluster inertia [4]. This procedure is repeated so as to obtain a single class at the end and the dendrogram will then be built. The best partition is obtained by the cut-off level of the classification tree which gives a low loss of within cluster inertia. In this article, classification is done using principal components. Euclidian distances between centers of clusters are also computed. R software, version 3.6.1 is used for data analysis.

3. RESULTS AND DISCUSSION

Descriptive statistics

Values of mean sugarcane yield (MEANSCY) vary between 65 tons/ha and 123 tons/ha with a mean of 84 tons/ha (Table 1). High values for range are observed for the variable representing

Vol. 08, No. 01; 2023

ISSN: 2456-8643

sugarcane production (SCPROD). An overdispersion is observed for the variable representing herbicide inputs (HERBINP).

Table 1. Descriptive statistics of quantitative variables

Variable	Frequency	Minimum	Mean	Standard deviation	Median	Maximum
MEANSCY	20	65	83.99	16.21	81	123
CULTAREA	20	2378	2852.15	201.85	2860	3250
SCPROD	20	132764	185481.34	27229.85	179107	239519
SPROD	20	14161	20126.80	2593.34	19839	25802
MOPROD	20	4469	6131.70	1382.51	5735.50	8858
HERBINP	20	10208	28650.38	37485.71	19700	182918.7
MEANTEMP	20	21.2	27.93	5.46	28.04	37.5
COSTSPROD	20	305	834.21	454.71	727	1741

Correlation analysis

Sugarcane production (SCPROD) is highly and positively correlated with sugar production (SPROD) and molasses production (MOPROD) ($r=0.94$, $p<0.001$, $r=0.87$, $p<0.001$ respectively). Cost of sugar production (COSTSPROD) is highly and positively correlated with mean sugarcane yield (MEANSCY) ($r=0.70$, $p=0.001$) and molasses production (MOPROD) ($r=0.76$, $p<0.001$). However, cultivated area (CULTAREA) and mean sugarcane yield (MEANSCY) are negatively correlated and the correlation coefficient is not significantly different from zero ($r=-0.02$, $p=0.940$) (Table 2).

Table 2. Correlations matrix

	MEANSCY	CULTAREA	SCPROD	SPROD	MOPROD	HERBINP	MEANTEMP	COSTSPROD
MEANSCY	1.00							
CULTAREA	-0.02	1.00						
SCPROD	0.64*	0.35	1.00					
SPROD	0.52*	0.27	0.94*	1.00				
MOPROD	0.55*	0.53*	0.87*	0.79*	1.00			
HERBINP	0.02	0.57*	0.22	0.10	0.31	1.00		
MEANTEMP	0.29	0.25	0.21	0.19	0.38	-0.16	1.00	
COSTSPROD	0.70*	0.47*	0.72*	0.61*	0.76*	0.13	0.49*	1.00

*: significant correlation

Eigenvalues, variables and individuals coordinates

More than seventy percent (72.1%) of total variance are captured by the two first factorial axes (two first principal components) and 85.1% of the variability of the data are explained by the three first factorial axes (Table 3). According Kaiser criterion, the three first factorial axes for

Vol. 08, No. 01; 2023

ISSN: 2456-8643

which eigenvalues are higher than one are selected. The two first eigenvalues are not significantly different with respect to Anderson confidence intervals (CI).

Table 3. Eigenvalues, confidence intervals and percentages of explained variance

Dimension	Eigenvalues	CI 95%	% of variance	% of cumulative variance
1	4.26	[2.26, 8.05]	53.3	53.3
2	1.50	[0.80, 2.84]	18.8	72.1
3	1.04	[0.55, 1.97]	13.0	85.1
4	0.59	[0.31, 1.11]	7.4	92.4
5	0.32	[0.17, 0.61]	4.1	96.5
6	0.13	[0.07, 0.25]	1.7	98.2
7	0.11	[0.06, 0.21]	1.4	99.6
8	0.03	[0.02, 0.06]	0.4	100.0

Figure 1 shows screeplot of eigenvalues (black) with the percentage of cumulative explained variance (red). A significant drop in inertia between the second and the third eigenvalue is observed (Figure 1).

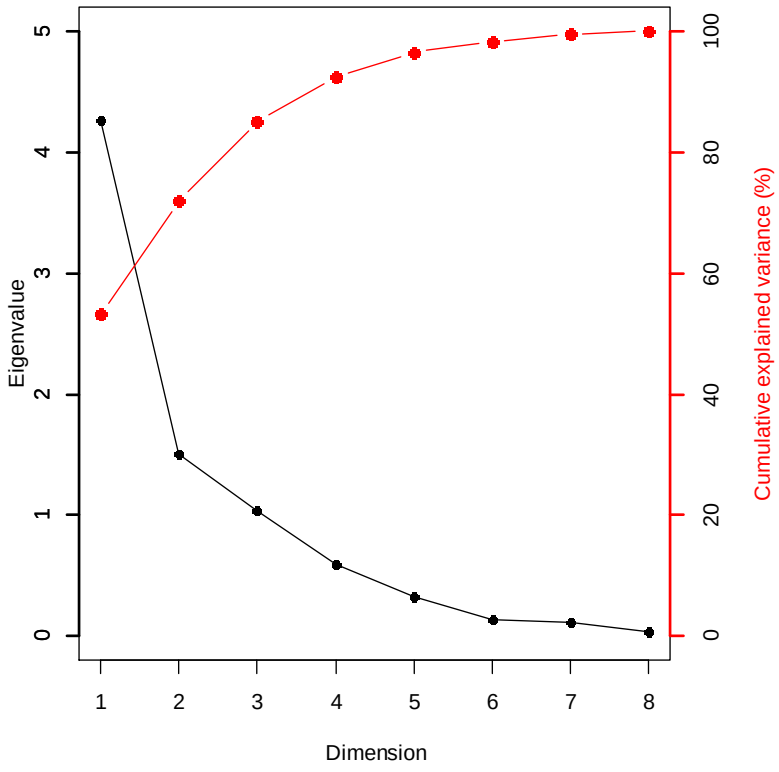


Figure 1. Screeplot of eigenvalues and percentage of explained variance

Vol. 08, No. 01; 2023

ISSN: 2456-8643

This drop in inertia suggests limiting ourselves to the first two factorial axes since 72.1% of the information contained in the data are captured by the first two principal components according to the elbow criterion (Cattell criterion). In other words, this figure shows a step (bend) followed by a regular decrease, which makes it possible to select the first two first factorial axes. Individuals (years) and variables will therefore be projected in the main plane.

All variables are positively and significantly correlated with the first factorial axis. Molasse production (MOPROD), sugarcane production (SCPROD), cost of sugar production (COSTSPROD) and sugar production (SPROD) are better represented with respect to the first axis considering the squares of cosine values (Figure 2, Table 4). These variables contribute at 75.6% to the formation of the first factorial axis. Herbicide inputs (HERBINP) is better represented with respected to the second axis and contributes at 46.8% to the building of this axis. Molasse production (MOPROD) and sugarcane production (SCPROD) are better represented in the main plane. The first factorial axis does not oppose anything. The second axis opposes mean sugarcane yield (MEANSCY) and herbicide inputs (HERBINP).

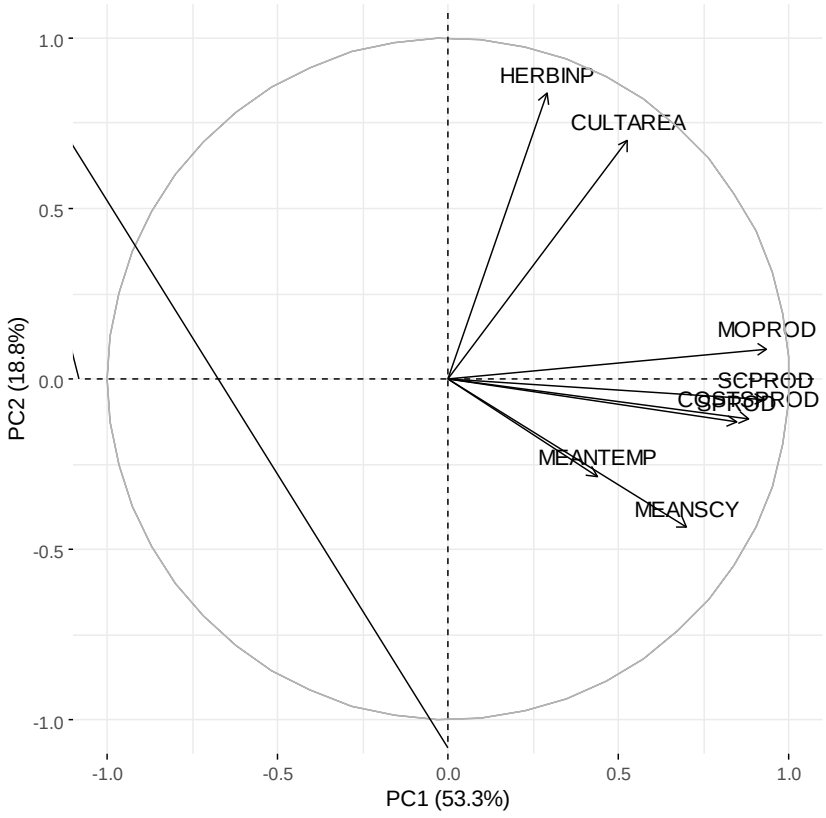


Figure 2. Correlations circle

Vol. 08, No. 01; 2023

ISSN: 2456-8643

Table 4. Coordinates, quality of representation and contribution of variables

	Coordinates		Squares of cosines		Contribution	
Variable	Dim.1	Dim.2	Dim.1	Dim.2	Dim.1	Dim.2
MEANSCY	0.70	-0.43*	0.49	0.19	11.46	12.58
CULTAREA	0.53	0.70	0.28	0.49	6.51	32.47
SCPRODC	0.93	-0.06*	0.86	0.00	20.15	0.24
SPROD	0.84	-0.13	0.71	0.02	16.74	1.07
MOPROD	0.93	-0.09*	0.87	0.01	20.51	0.53
HERBINP	0.29*	0.84	0.08	0.70	1.95	46.78
MEANTEMP	0.44*	-0.29	0.19	0.08	4.48	5.45
COSTSPROD	0.88	-0.12*	0.78	0.01	18.20	0.89

* : not significant

The first factorial axis opposes 2002 and 2013 whereas the second axis opposes 2001 and 2017 (Figure 3, Table 5). Besides, 2000 and 2013 are better represented in relation to the first factorial axis with regard to the values of the squared cosines. Figure 3 (biplot) helps to find notable individuals in the main plane, and interpretation of positioning individuals and variables.

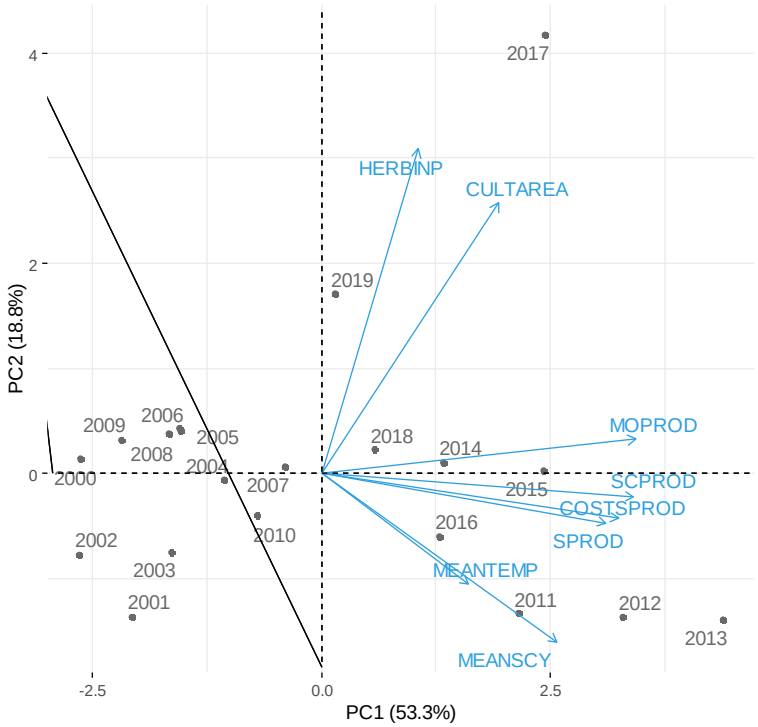


Figure 1. Biplot of variables and individuals

Vol. 08, No. 01; 2023

ISSN: 2456-8643

The year 2017 is the best represented individual in relation to the second factorial axis. The years 2000, 2002, 2008, 2012, 2013 and 2017 are better represented in relation to the principal factorial plane (Figure 3, Table 5). The years 2012 and 2013 contribute the most to the construction of the first factorial axis (35.3%) while the year 2017 contributes the most to the construction of the second one (57.9%). The years 2017 and 2019 are characterized by high values of herbicide inputs (HEBRINP) and cultivated area (CULAREA) while the years 2001, 2002 and 2003 are characterized by low values of those variables. The years 2011, 2012 and 2013 are characterized by high values of mean temperature (MEANTEMP) and mean sugarcane yield (MEANSCY) while the years 2005, 2006, 2008 and 2009 are characterized by low values of those variables. The year 2015 is characterized by high values of sugar production (SPROD), sugarcane production (SCPROD), cost of sugar production cost (COSTSPROD) and molasses production (MOPROD). The years 2004, 2007, 2010 and 2018 are poorly represented in relation to the whole plane.

Table 5. Coordinates, quality and contribution of individuals

	Coordinates		Squared cosines		Contribution	
Year	PC1	PC2	PC1	PC2	PC1	PC2
2000	-2.63	0.14	0.93	0.00	8.12	0.06
2001	-2.07	-1.36	0.43	0.18	5.02	6.14
2002	-2.64	-0.77	0.81	0.07	8.17	1.97
2003	-1.64	-0.74	0.56	0.12	3.15	1.83
2004	-1.06	-0.06	0.52	0.00	1.32	0.01
2005	-1.53	0.40	0.67	0.05	2.75	0.55
2006	-1.55	0.43	0.65	0.05	2.82	0.62
2007	-0.40	0.07	0.13	0.00	0.19	0.02
2008	-1.66	0.38	0.79	0.04	3.23	0.47
2009	-2.18	0.32	0.39	0.01	5.59	0.35
2010	-0.69	-0.40	0.14	0.05	0.56	0.54
2011	2.15	-1.32	0.50	0.19	5.43	5.84
2012	3.29	-1.36	0.76	0.13	12.72	6.17
2013	4.39	-1.39	0.89	0.09	22.57	6.40
2014	1.33	0.11	0.55	0.00	2.07	0.04
2015	2.43	0.03	0.67	0.00	6.96	0.00
2016	1.28	-0.59	0.40	0.09	1.93	1.17
2017	2.44	4.17	0.23	0.66	6.97	57.91
2018	0.58	0.23	0.06	0.01	0.40	0.17
2019	0.15	1.71	0.00	0.51	0.03	9.75

Hierarchical ascending classification

Vol. 08, No. 01; 2023

ISSN: 2456-8643

Figure 4 shows the positioning of each individual with respect to the axes with a distribution of individuals in clusters using concentration ellipses of level 0.80. The third cluster does not show a concentration ellipse due to the small number of individuals in this cluster (only one individual). Figure 5 shows the dendrogram formed by the grouping of individuals at each stage and their levels of similarity.

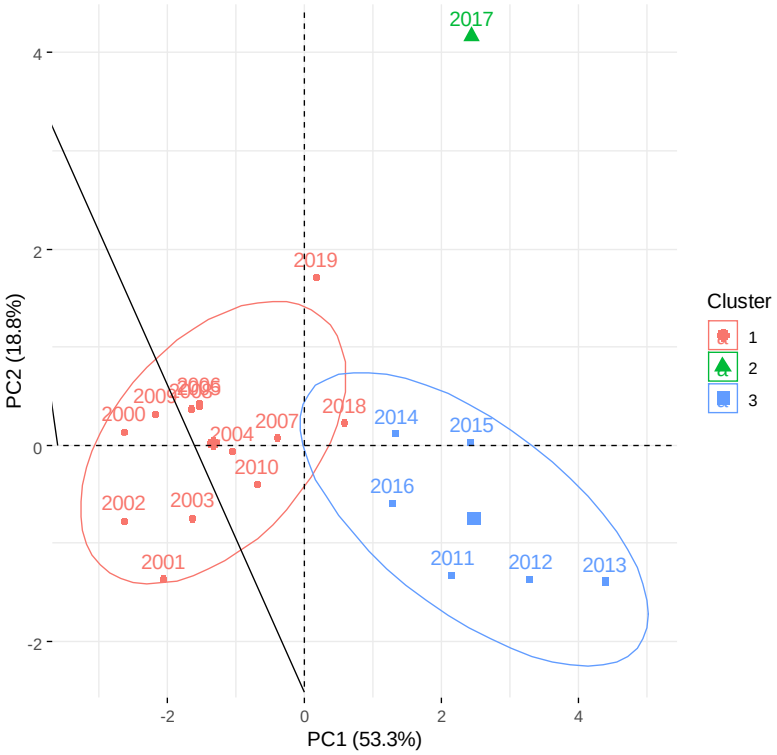


Figure 2. Classification of individuals

There are three clusters (Figure 4, Figure 5). The first one is composed by the years 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2018 and 2019. The second one contains the years 2011, 2012, 2013, 2014, 2015, 2016 and the last one the year 2017. All those three clusters are mutually exclusive and well discriminated insofar as it is possible to draw a straight line separating them as best as possible. The clusters that are further apart are the first and the second one since the distance between their centroids is 5.98. Distance between the centroids of the second and the third clusters is 5.20.

Vol. 08, No. 01; 2023

ISSN: 2456-8643

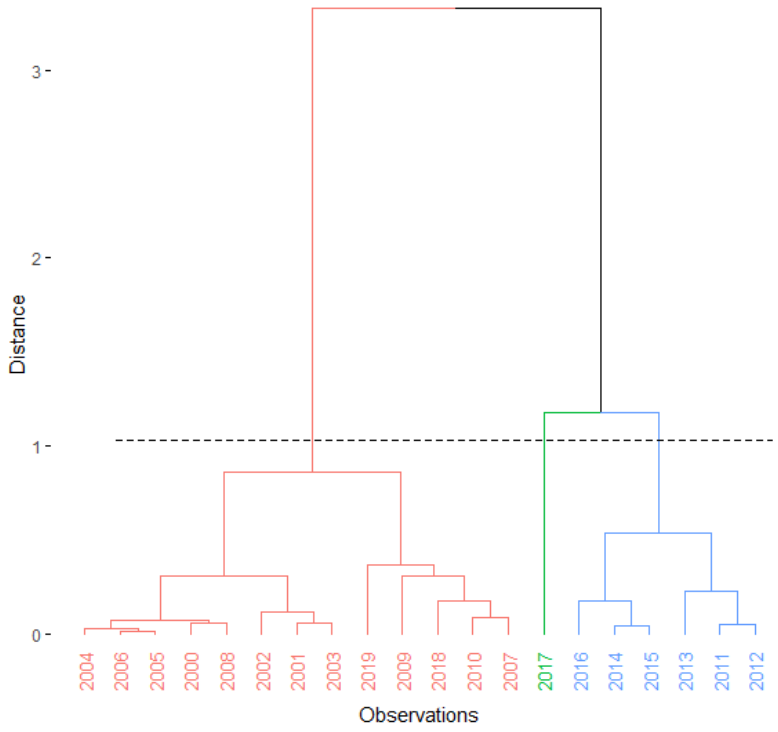


Figure 5. Dendrogram

As shown in Figure 4, final subdivision is made of 3 clusters. The first cluster on the far left of Figure 5 consists of thirteen individuals: 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2018 and 2019. The second cluster, to the right of the first one, consists of one individual (2017). The third cluster on the far right consists of six individuals: 2011, 2012, 2013, 2014, 2015 and 2016. The different groupings show the similarity of the years.

This study used Principal Component Analysis as a factor analysis method and hierarchical ascending classification as a classification method. It applied these techniques to sugarcane production data. Other studies conducted elsewhere in the world have used these techniques to investigate the morphological characteristics and genetic diversity of sugarcane varieties [5],

Vol. 08, No. 01; 2023

ISSN: 2456-8643

[18]. In Burundi, sugarcane is grown mainly in swamps and for various seasons as irrigation becomes easy and sporadically on the hills. In this study, sugarcane and sugar production are strongly and positively correlated. In addition, cultivated area and mean sugarcane yield are negatively and significantly correlated. The latter result corroborates the one found in a study conducted in the Ndwedwe locality in South Africa [11]. The calculation of Pearson's linear correlation coefficient identified highly correlated variables that should not be used in Principal Component Analysis [5]. The study of Tawadare conducted in India and which focussed on phenotypic characterization and genetic diversity of sugarcane varieties showed that the first four factorial axes captures 76.72% of the total inertia, whereas it is 92.40% in our study. This noticed difference would probably be due to the fact that data used in their study are genetic and for these data, the principal components that allow a good visualization of the data are not necessarily the first two factorial axes [21]. Years 2017 and 2019 were characterized by high values of herbicides and cultivated area. Since Burundi has opened doors to foreign investors, sugarcane yield has increased. Moreover, Burundi considered it was necessary to expand the area cultivated for sugarcane. After the 2015 sociopolitical situation, socioeconomic conditions were getting better and better. Hence, cultivated area increased. As a consequence, herbicide inputs also increased to prevent sugarcane to be attacked by enemies. As other pesticides (insecticides, fungicides, acaricides, rat poisons), pesticides are increasingly used by agro-pastoralists to reduce rapidly the populations of sugarcane (and other crops) enemies and contribute to increase agricultural production and productivity. Researchers interested in sugarcane production should intensively assay this culture in other provinces such Bujumbura and Bubanza where conditions seem to be auspicious.

4. CONCLUSION

This study shows, among other results, that sugarcane and sugar production are strongly and positively correlated and that cultivated area and mean sugarcane yield are negatively correlated. Principal Component Analysis shows that molasses production and sugar production cost are better represented in relation to the first factorial axis while herbicide inputs is better represented with respect to the second one. The first factorial axis does not oppose anything while the second one opposes mean sugarcane yield and herbicide inputs. Molasses production and sugarcane production are better represented in the main plane. The biplot of individuals and variables shows that the years 2017 and 2019 are characterized by high values of herbicide inputs and cultivated area while the years 2011, 2012 and 2013 are characterized by high values of mean temperature and mean sugarcane yield. The classification analysis based on principal components leads to the formation of three clusters of years. The first cluster contains the years 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2018 and 2019. The second one has one individual: the year 2017, and the third and last one the years 2011, 2012, 2013, 2014, 2015 and 2016. The first and the third clusters are close while the first and the second ones are far apart using distances between centroids.

Vol. 08, No. 01; 2023

ISSN: 2456-8643

REFERENCES

[1] Hoarau J.-Y., Dumont T., Wei X., *et al.*(2021). Applications of Quantitative Genetics and Statistical Analyses in Sugarcane Breeding. Sugar Tech. [Online] Available: https://doi.org/10.1007/s12355-021-01012-3 [Accessed: Oct. 26, 2022].

[2] Pocovi M., Collavino N., Gutiérrez Á., *et al.* (2020). Molecular versus morphological markers to describe variability in sugar cane (Saccharum officinarum) for germplasm management and conservation. Rev. FCA UNCUYO, 52(1) pp. 40–60.

[3] Raza I., Farooq M.A., Masood M.A., *et al.* (2017). Exploring Relationship among Quantitative Traits of Sugarcane Varieties Using Principal Component Analysis. Science, Technology and Development, 36(3) pp. 142–146.

[4] Sumbele S.A., Fonkeng E.E., Akongte P., et al. (2021). Characterization of sugarcane germplasm collection and its potential utilization for evaluation of quantitative traits. Afr. J. Agric. Res., 17(2), pp. 273–282.

[5] Tawadare R., Thangadurai D., Khandagave R.B., et al. (2019). Phenotypic Characterization and Genetic Diversity of Sugarcane Varieties Cultivated in Northern Karnataka of India based on Principal Component and Cluster Analyses. Brazilian Archives of Biology and Technology, 62 [Online] Available: https://doi.org/10.1590/1678-4324-2019180376 [Accessed: Oct. 12, 2022].

[6] Bamber P., Guinn A. and Gereffi G. (2014). Burundi in the Agribusiness Global Value Chain: Skills for Private Sector Development. [Online] Available: https://doi.org/10.13140/RG.2.1.3194.6329 [Accessed: Oct. 14, 2022].

[7] Youlton C., Wendland E., Anache J., et al. (2016). Changes in Erosion and Runoff due to Replacement of Pasture Land with Sugarcane Crops. Sustainability, 8(7), 685. [Online] Available: https://doi.org/10.3390/su8070685 [Accessed: Aug. 3, 2022].

[8] Linnenluecke M.K.. Nucifora N. and Thompson N. (2018). Implications of climate change for the sugarcane industry. WIREs Climate Change, 9(1) [Online] Available: https://doi.org/10.1002/wcc.498 [Accessed: Oct. 16, 2022].

[9] Pipitpukdee S., Attavanich W. and Bejranonda S. (2020). Climate Change Impacts on Sugarcane Production in Thailand", Atmosphere, 11(4), 408 pp. 1-15.

[10] Simões W., de Oliveira A., Salviano A., *et al.* (2021). Efficient irrigation management in sugarcane cultivation in saline soil. Brazilian Journal of Agricultural and Environmental Engineering, 25(9) pp. 626–632 [Online] Available: http://dx.doi.org/10.1590/1807-1929/agriambi.v25n9p626-632 [Accessed: Oct. 26, 2022].

[11] Zulu N.S., Sibanda M. and Tlali B.S. (2019). Factors Affecting Sugarcane Production by Small-Scale Growers in Ndwedwe Local Unicipality, South Africa. Agriculture, 9(8), 170, [Online] Available: https://doi.org/10.3390/agriculture9080170 [Accessed: Aug. 11, 2022].

Vol. 08, No. 01; 2023

ISSN: 2456-8643

[12] NISBu (2018). National Agricultural Survey of Burundi: Campaign 2016-2017. [Online] Available: https://bi.chm-cbd.net/fr/implementation/documents-envir-biodiv/enabi-campagne-2016-2017 [Accessed: Nov. 5, 2022].

[13] NISBu (2012). National Agricultural Survey of Burundi 2011-2012: Results of the session A », pp. 103 [Online] Available: https://bi.chm-cbd.net/sites/bi/files/2019-10/enq-nat-agri-bi-2011-2012-sais-a.pdf [Accessed: Dec. 16, 2022].

[14] Jolliffe I.T. and Cadima J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065) pp. 1-16 [Online] Available: https://doi.org/10.1098/rsta.2015.0202 [Accessed: Oct. 13, 2022].

[15] Rea-Suárez R., Figuered L., De Sousa-Vieira O., *et al.* (2018). Genotype by environment interactions for damage caused by Diatraea spp. Borers in sugarcane. Acta Agron. 67(2) pp. 355–361. [Online] Available: https://doi.org/10.15446/acag.v67n2.65881 [Accessed: Dec. 26, 2022].

[16] Shadmehr A., Ramshini H., Zeinalabedini M., et al. (2017). Phenotypic Variability Assessment of Sugarcane Germplasm (Saccharum officinarum L.) and Extraction of an Applied Mini-Core Collection. Agriculture, 7(55) pp. 1–15 [Online] Available: https://doi.org/doi:10.3390/agriculture7070055 [Accessed: Oct. 13, 2022].

[17] Gómez D., Hernandez L., Yabor L., *et al*. (2018). Euclidean distance can identify the mannitol level that produces the most remarkable integral effect on sugarcane micropropagation in temporary immersion bioreactors. Journal of Plant Research, 131(9), pp. 1–6 [Online] Available: https://doi.org/10.1007/s10265-018-1028-7 [Accessed: Oct. 13, 2022].

[18] Aamer M., Anwar M.R., Mustafa G., *et al*. (2018). Principal Component Analysis (PCA) of Some Morphological and Quality Traits in Sugarcane (Saccharum officinarum L.). Journal of Natural Sciences Research, 8(14) pp. 22-26 [Online] Available: https://www.iiste.org/Journals/index.php/JNSR/article/view/43528/44844. [Accessed: Dec. 16, 2022].

[19] Manechini J.R.V, Santos P.H.dS, Romanel E., *et al.* (2021). Transcriptomic Analysis of Changes in Gene Expression During Flowering Induction in Sugarcane Under Controlled Photoperiodic Conditions. Front. Plant Sci., pp. 1-18 [Online] Available: https://doi.org/doi: 10.3389/fpls.2021.635784 [Accessed: Oct. 13, 2022].

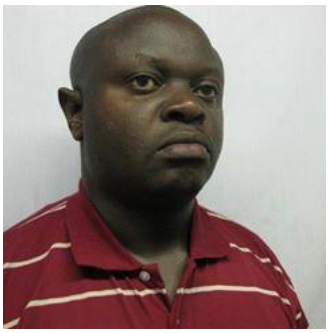
[20] Anderson F.L. (2021). Huygens' Principle geometric derivation and elimination of the wake and backward wave. Scientific Reports, 11(1) [Online] Available: https://doi.org/10.1038/s41598-021-99049-7 [Accessed: Oct. 10, 2022].

[21] Yeung K.Y. and Ruzzo W.L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics, 17(9), pp. 763–774. [Online] Available: https://doi.org/10.1093/bioinformatics/17.9.763 [Accessed: Oct. 8, 2022].

Vol. 08, No. 01; 2023

ISSN: 2456-8643

Author Profile



BARANKANIRA Emmanuel received the B.S. in Mathematics (Applied Pedagogy) from University of Burundi in 2004, a M.S. degree in Statistics, Epidemiology and Biostatistcs from Catholic University of Louvain (Belgium) in 2008, a M.S. degree in Mathematics-Biostatistics from National Higher School of Agronomy (France) in 2012 and a PhD in Biostatistics from University of Montpellier (France) in 2016. During 2004-2006, 2009-2011, 2017 up to now, he is a Teacher-Researcher in Biostatistics at Burundi Higher Institute of Education (École Normale Supérieure), Ministry of National Education and Scientific Research. He is part of Sciences Research and Professional Development Center and chairman of Statistical Consulting and Geoinformatics Business Research Center (STACOGERC).

NDIKUMANA Valentin received a B.S. in Statistics from Lake Tanganyika University (Burundi) in 2020. He is part of Statistical Consulting and Geoinformatics Business Research Center (STACOGERC).

Vol. 08, No. 01; 2023

ISSN: 2456-8643

NITUNGA Benjamin received a B.S. in Statistics from Lake Tanganyika University (Burundi) in 2021. He is part of Statistical Consulting and Geoinformatics Business Research Center (STACOGERC).

Acknowledgements: The authors are very grateful to the Ministry of Environment, Agriculture and Livestock for providing the data.

Financial support: None.

Conflict of interest: None.

Authorship: BARANKANIRA Emmanuel, NDIKUMANA Valentin and NITUNGA Benjamin developed the statistical analysis design. NDIKUMANA Valentin obtained data from the Ministry of Environment, Agriculture and Livestock of Burundi. BARANKANIRA Emmanuel and NDIKUMANA Valentin carried out statistical analysis and, with NITUNGA Benjamin, wrote the draft of the article and contributed equally to results interpretation.