Journal of Biodiversity and Environmental Sciences (JBES)
ISSN: 2220-6663 (Print) 2222-3045 (Online)
Vol. 13, No. 4, p. 54-62, 2018
http://www.innspub.net

INNSPUB
RESEARCH PAPER
OPEN ACCESS

Comparative study on Saurel *Trachurus trachurus* fish obtained from the Algerian littoral using multivariate statistical methods (principal component analysis and hierarchical analysis)

Souheïla Azzouz[1*], Lyamine Mezedjri[1], Ali Tahar[2]

[1] *NVS Department, Faculty of sciences, University of 20 Août 55, Skikda 21000, Algeria*

[2]*Laboratory of Vegetable Biology and Environment, departmentof biology, Faculty of sciences, University of Badji Mokhtar, Annaba 23000, Algeria*

Article published October 06, 2018

Key words: *Trachurus trachurus*, Algerian littoral, Statistical analysis, PCA, Hierarchical analysis.

Abstract

A comparative biometric study is carried out by applying multivariate statistical methods (principal components analysis and hierarchical analysis) on samples of the teleost fish Saurel *Trachurus trachurus* fished from the Algerian coasts during the year 2012-2013 in order to distinguish significant differences between different sites and both sexes in the sites. The sampling is done from seven sites located on the Algerian coastline from East to West: El-kala, Annaba, Skikda, Collo, Jijel, Algiers, Oran, carrying out on each fish 36 morphometric and meristic measurements. Then, multivariate statistical methods; PCA and a hierarchical analysis are applied. The principal component analysis PCA allows us to group the seven sites (sexes) into five homogeneous groups, for each of the morphometric and meristic characteristics taken into consideration. Whereas, in the case of the 36 variables, the PCA allows us to find four groups. The hierarchical analysis of data made it possible to group the 36 variables studied into ten homogeneous groups, and to show the links and the similarities between these variables. The hierarchical analysis also allowed, with a level of similarity of 98.16%, the grouping of the seven sites (sexes) of sampling according to the 36 morphometric and meristic variables, into four distinct homogeneous groups. This method illustrates well and highlights the relationships between the studied variables thanks to the correlation coefficient.

***Corresponding Author:** Souheïla Azzouz ✉ azzouz.souhi@yahoo.fr.

Introduction

The Saurel *Trachurus trachurus* (Linnaeus, 1758), teleost fish, on which we chose, belongs to the group of small pelagic which represent the most important volumes in the catches in Algeria. The works on Saurel are numerous: Barraca (1964); Geldenhuys (1973); Wengrzyn (1976); Kerstan (1985) and (1986); Borges (1991); Tsangridis and Fillippousis (1991); Karlou-Riga (1995); Ben Salem, *et al.*, (1981); Ben Salem (1983); Borges (1978); Korichi (1988); Ben Salem (1988); Nowogrodzka (1943); Russel (1976); Giovanardi and Romanelli (1990).

This experimental scientific work is made for an object of a comparative biometric (morphometric) study based on multivariate statistical analyzes of Saurel fish from all over the Algerian littoral in order to enrich the biological knowledge of the studied species in the Algerian halieutic resources.

Materials and methods

Data collection

This study is carried out on fish caught by sardine boats, trawlers and small crafts, from seven sites on the Algerian coastline from East to West: El-kala, Annaba, Skikda, Collo, Jijel, Algiers, and Oran.

A sample of at least 30 individuals is taken into consideration at each site. Each person is wrapped in a plastic film immediately after collection to avoid damage, and is put in the freezer at -20 °C. In the laboratory a series of 36 morphometric and meristic measurements are made on each fish (Table 1 and Fig. 1). All metric measurements are made, to the nearest millimeter, using a dry point compass. The meristic measurements are made, under a binocular loupe, by means of a count. Sex determination was performed after fish dissection.

Multivariate Statistical Analysis

Principal Component Analysis (PCA): Principal component analysis (PCA) is an exploratory and descriptive method (Dagnélie, 1970, 1986 and 2006; Palm, 1998).

It is used to interpret a matrix of data without a particular structure that, in principle, has no distinction between variables or between individuals.

It aims to replace the initial p variables strongly correlated with each other in p variables called main components or synthetic main axes uncorrelated between them, and of progressively decreasing variance.

The first components can possibly be the subject of a particular interpretation and the latter can generally be neglected (Dagnélie, 1970, 1986 and 2006). It is considered that the principal components having an Eigenvalue equal to or greater than unity.

This method (CPA) was applied to the 14x36 data matrix, with n = 14 row vectors representing the 2 sexes in each of the 7 sites and p = 36 columns representing mean vectors of the variables.

Hierarchical analysis

The search for groups or classes of homogeneous sites can also be done by what is called the hierarchical classification. Several methods are proposed by Dagnélie (1986) to achieve this goal. However, we will only use the one proposed by Bouroche and Saporta (1980), which is taken up by Palm (2000) and Dagnélie (1970 and 2006) and whose algorithm is programmed in the Minitab software (X, 2013).

This method makes it possible to determine the level of similarity or divergence between the individuals (or sites) and gives a distribution of individuals or sites into homogeneous groups or classes.

It is an agglomerative hierarchical method that uses the simple link procedure and the Pearson square distance (X, 2013) to classify the 2 sexes in each of the 7 sites in classes as homogeneous as possible. It was applied to the 14x36 data matrix, with n = 14 row vectors and p = 36 mean column vectors representing the variables.

Results

Principal Component Analysis

Principal component analysis (PCA) was applied to the correlation matrix obtained from the 36 reduced central variables taken in pairs.

The calculations carried out with the Minitab software give the characteristics of the 36 main axes, of which the first 3 have each, an Eigen value superior to the unit. From Table 2, it is clear that the first axis alone accounts for 86.00% of the total variation of the initial variables, the first two axes together explain 91.70% and finally the first three axes account for 94.80%. So these three axes or synthetic indices summarize better the information provided by the 36 variables.

Table 1. Morphometric and Meristic Studied Variables.

Morphometric measurements		
Number	Code	Description
1	Lt	Total length
2	Lf	At fork length
3	Ls	Standard length
4	Lpan	Length pre-anal
5	Lppv	Length pre-pelvic
6	Lppc	Length pre-pectoral
7	Lcep	Cephalic length
8	Lpdo	Length pre-dorsal
9	Dopv	Dorsal / pelvic distance
10	Doan	Dorsal / anal distance
11	Doca	Dorsal / Caudal Distance
12	Lman	Mandible length
13	Lmax	Maxillary length
14	Poor	Distance post-orbitaire
15	Dor	Diameter Orbital
16	Pror	Length Pre-orbital
17	Lpop	Length pre-operculum
18	Lain	Inorbital Width
19	Lcra	Head width
20	Mist	Length mandible / isthmus
21	Lapc	Distance between pectoral insertions
22	Hpc	Pectoral Height
23	Hpv	Pelvic Height
24	Hdo	Dorsal Height
25	Han	Anal Height
26	Hpdc	Peduncle Height
27	Bado	Dorsal Height
28	Baan	Anal Height
29	Dopc	Distance dorsal/pectoral
30	Pcpv	Distance pectoral/pelvic
31	Pvan	Distance pelvic/anal
Meristic counting		
32	Cæc	Cæc Number of pyloric caecum
33	Brin	Number of lower gill rakers of the 1st left branchial arch
34	Brsu	Number of upper gill rakers of the 1st left branchial arch
35	Rypc	Number of rays of the left chest
36	Rypv	Number of left pelvic rays

In Fig. 2 the graph shows the appearance of scree at the foot of an escarpment, showing the existence of an elbow, that is to say a sharp reduction of the slope of the graph thus allows neglecting all the components corresponding to the Eigen values located after this elbow. However, the interpretation of these elements allows the calculation of the correlations between each of the 36 main initials.

These correlations are useful for specifying the portion of the variance, of a given initial variable, of a report for a particular major component, and of an employee for the graphical representations of the initial variables in the correlation circles.

Table 2. Eigen values and percentages of total variation explained by the three first main axis and cumulative percentages.

Parameter	Axis 1	Axis 2	Axis 3
Own value	30.967	2.045	1.122
Percentage explained in%	86.000	5.700	3.100
Percentage cumulated in%	86.000	91.700	94.800

Indeed, the determination of the part of information contained in any part and relating to any initial variable is given by the square of the calculated correlation coefficient between the two variables in question. If, on the other hand, we take into consideration the first two components, the information, the price in accounts, for a given initial variable, the equivalent in all the information, the price in account of each of the two components (Table 3).

Table 3. Values of the correlations and square correlations of the initial variables with the first 3 main contributions, and price information in account by the first factorial plane represented by axes 1 and 2.

Axis		Axis 1	Axis 2		Axis 3		Factorial 1-2 Plan
Variables	Corr.	Corr. Square (%)	Corr.	Corr. Square %)	Corr.	Corr. Square (%)	Corr. Square (%)
Lt	0.179	3.204	-0.029	0.084	0.014	0.019	3.288
Lf	0.179	3.204	-0.047	0.220	0.016	0.025	3.424
Ls	0.178	3.168	-0.044	0.193	0.001	0.000	3.361
Lpan	0.178	3.168	-0.062	0.384	0.045	0.202	3.552
Lppv	0.178	3.168	-0.027	0.072	-0.055	0.302	3.240
Lppc	0.178	3.168	0.007	0.004	-0.067	0.448	3.172
Lcep	0.178	3.168	0.011	0.012	-0.065	0.422	3.180
Lpdo	0.179	3.204	-0.039	0.152	0.001	0.000	3.356
dopv	0.178	3.168	-0.011	0.012	0.010	0.010	3.180
doan	0.177	3.132	-0.077	0.592	0.051	0.260	3.724
doca	0.176	3.097	-0.118	1.392	0.031	0.096	4.489
Lman	0.176	3.097	0.083	0.688	-0.035	0.122	3.785
Lmax	0.163	2.656	0.131	1.716	-0.070	0.490	4.372
Poor	0.177	3.132	-0.083	0.688	0.034	0.115	3.820
Dor	0.168	2.822	0.067	0.448	-0.178	3.168	3.270
Pror	0.179	3.204	-0.018	0.032	-0.008	0.006	3.236
Lpop	0.179	3.204	0.006	0.003	-0.022	0.048	3.207
Lain	0.176	3.097	-0.077	0.592	0.042	0.176	3.689
Lcra	0.176	3.097	0.092	0.846	0.125	1.562	3.943
Mist	0.177	3.132	0.037	0.136	0.015	0.022	3.268
Lapc	0.174	3.027	-0.150	2.250	0.059	0.348	5.277
Hpc	0.179	3.204	0.002	0.000	0.012	0.014	3.204
Hpv	0.160	2.560	0.215	4.622	-0.062	0.384	7.182
Hdo	0.174	3.027	-0.065	0.422	0.153	2.340	3.449
Han	0.161	2.592	0.198	3.920	0.009	0.008	6.512
Hpdc	0.158	2.496	-0.063	0.396	-0.113	1.276	2.892
Bado	0.177	3.132	-0.102	1.040	0.053	0.280	4.172
Baan	0.177	3.132	-0.057	0.324	0.021	0.044	3.456
dopc	0.178	3.168	-0.068	0.462	-0.038	0.144	3.630
pcpv	0.173	2.992	-0.114	1.299	0.088	0.774	4.291
pvan	0.172	2.958	-0.072	0.518	0.095	0.902	3.476
cæc	0.151	2.280	-0.082	0.672	0.391	15.288	2.952
brin	-	0.004	0.464	21.529	0.647	41.860	21.533
brsu	0.007	0.490	0.564	31.809	-0.002	0.000	32.299
rypc	0.070	1.795	0.199	3.960	-0.487	23.716	5.755
rypv	0.134	0.864	0.429	18.404	-0.225	5.062	19.268
	0.093						

Examination of Table 3 shows that axis 1, which on average alone contains 86% of the information provided by all the initial variables, consists essentially of the 31 morphometric variables that have almost equal proportions. The proportion of variance taken into consideration by axis 1 for each of these 31 initial variables is generally close to 3%. While axis 2, which contains only 5.7% of the information is formed by the variables strand, brsu, rypv and whose proportion of each is greater than 18.4%. As for the third one, which contains 3.1% of information, it is formed by the cæc, brin, brsu, rypc variables which are positively correlated to it and whose proportions are respectively 15.588% and 23.716% and 41.860%.

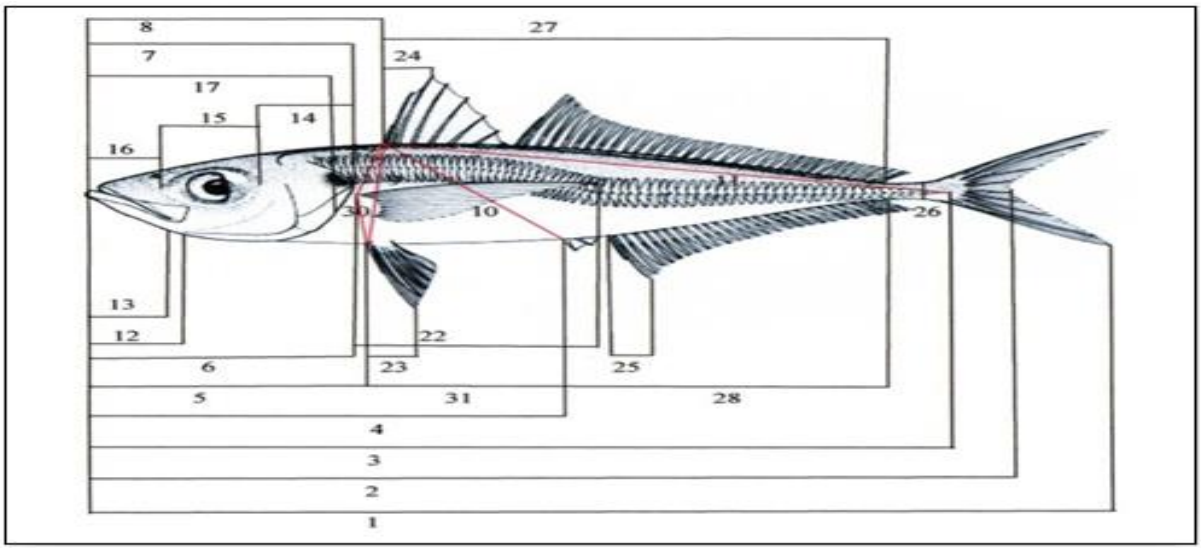

Fig. 1. Morphometric measurements taken on each fish.

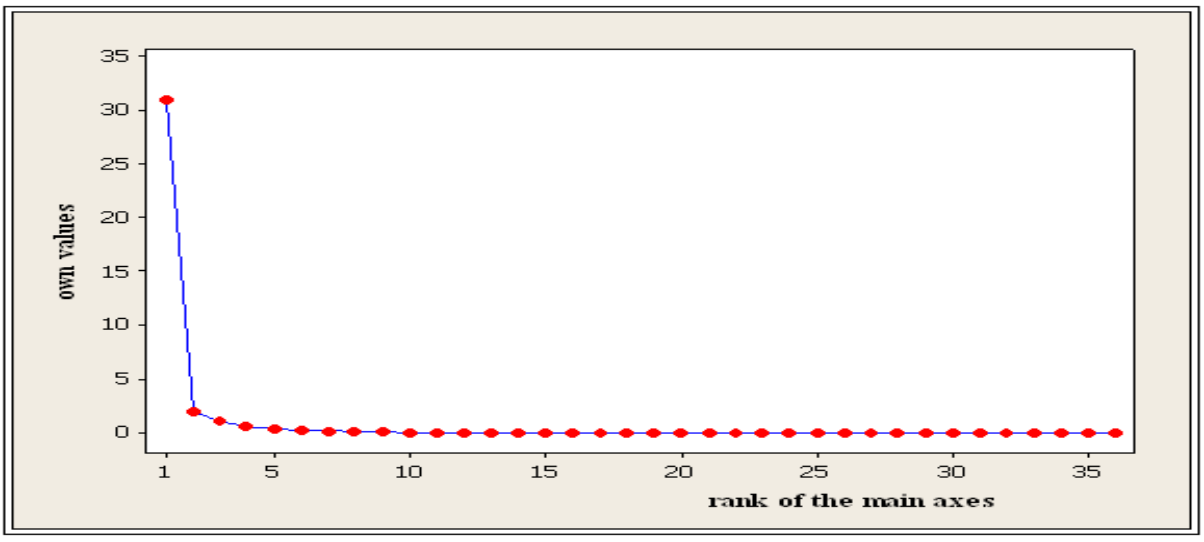

Fig. 2. Graph of Eigen values by rank of principal components.

Analysis of the cloud of variable points: circles of correlations: The correlation circles are graphs that geometrically represent the initial variables in the new coordinate system. Thus, the representation of the 36 initial variables in the plane formed by the axes 1 and 2 and called the first factorial plane is useful, given the importance of these two axes in the reconstitution of the initial variables (ie 91.7% of the variation total) (Fig. 3). The coordinates of the initial variables on the axis 1 are the correlations of these same variables with the axis 1, and same thing for the axis 2 (Fig.3).

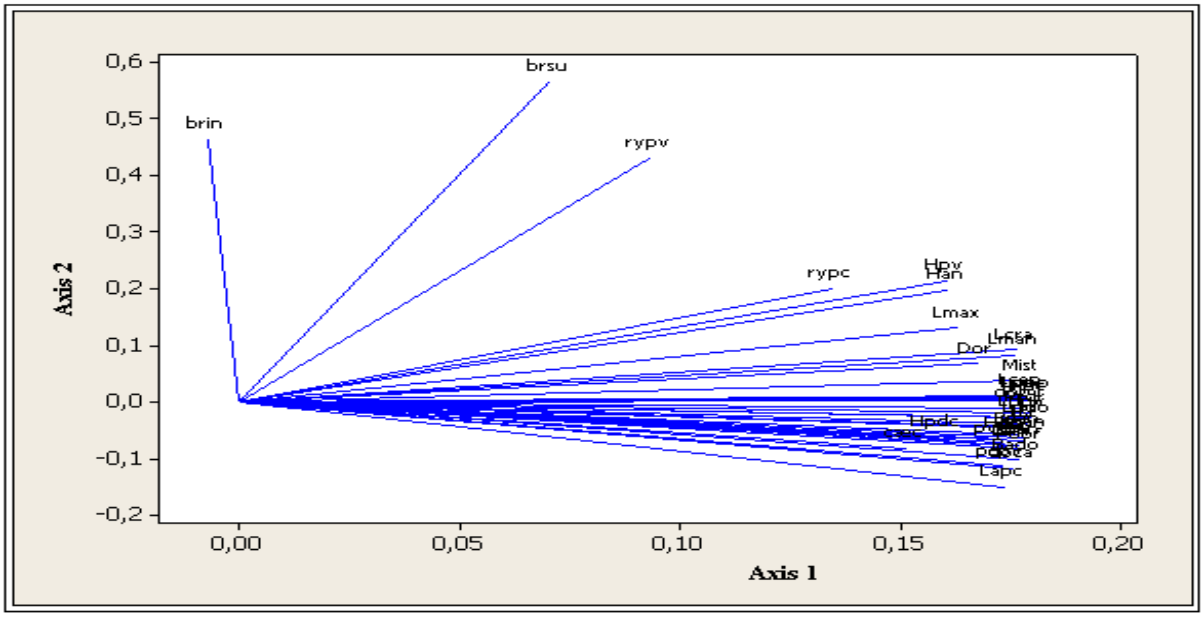

Fig. 3. Graphical representation of the 36 variables within the circle of correlations of the factorial plan 1-2.

The examination of Fig.3 shows two positive correlations with axis 1 of these variables, the first correlation is of variable rypc, Hpv, Han and Lmax, and the variables Lt, Lf, Ls, Lpan, Lppv, Lppc, Lcep, Lpdo, dopv, doan, doca, Lman, Poor, Dor, Pror, Lpop, Lain, Lcra, Mist, Lapc, Hpc, Hdo, Hpdc, Bado, baan, dopc, pcpv, pvan, cæc, representing the second correlation.

For the interpretation of the second axis, we find that there are two correlations that are clearly marked. This is positive strand correlations and the second variable correlation brsu and the other rypv. There are therefore 4 groups of variables: first group consists of the variable strand, second group consists of two variables brsu and rypv; third group includes the following four variables: rypc, Hpv, Han, Lmax, and fourth group includes the rest of the variables which are: Lt, Lf, Ls, Lpan, Lppv, Lppc, Lcep, Lpdo, dopv, doan, doca, Lman, Poor, Dor, Pror, Lpop, Lain, Lcra, Mist, Lapc, Hpc, Hdo, Hpdc, Bado, baan, dopc, pcpv, pvan, cæc.

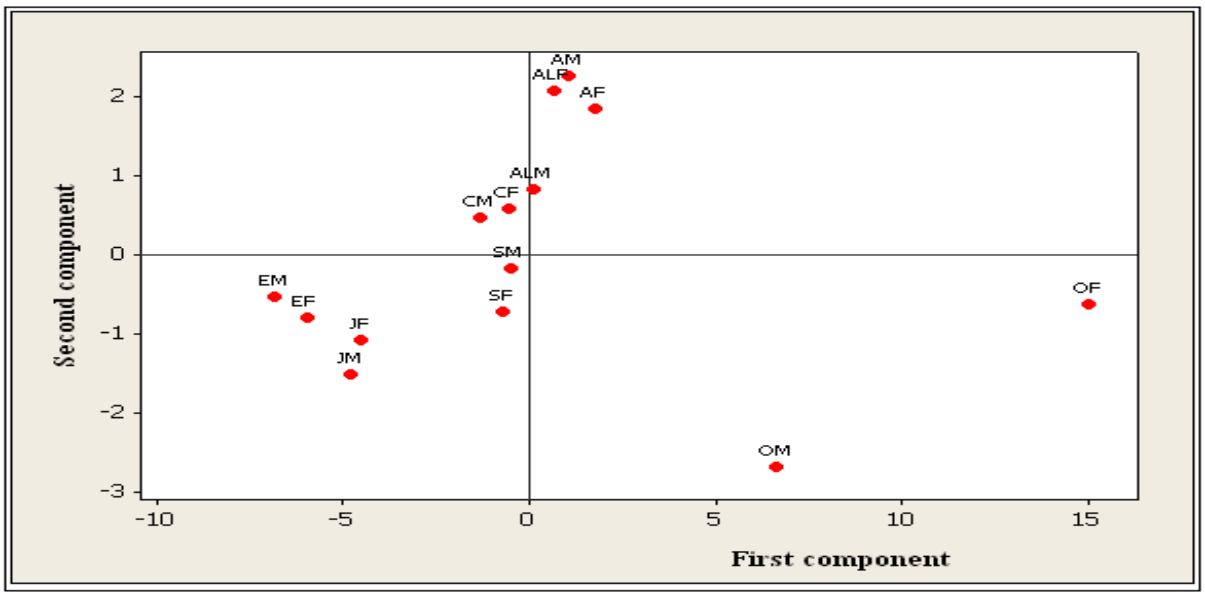

Fig. 4. Graphical Representation of Individual Points (Sites) in Factorial Design1-2.

Cloud analysis of point-sites: graphs of individuals: Knowing that the first factorial plan 1-2 alone accounts for a high percentage of 91.7% of the total inertia, we will naturally graphically represent the individual sex (sites) in this plan. The coordinates of each sex point in site are the values of axis 1 and axis 2 (Fig. 4).

On the basis of their position in the first factorial plane 1-2, there are 5 groups of sites which are: first group is represented by the females of Oran, second group consists of the males of Oran, third group is related to Annaba males, Annaba females and Algiers females, and fourth group includes:

Algiers males, Collo males, Collo females, Skikda males, Skikda females, and finally, the fifth group consists of the males of Jijel, females of Jijel, males of El-kala and females of El-kala.

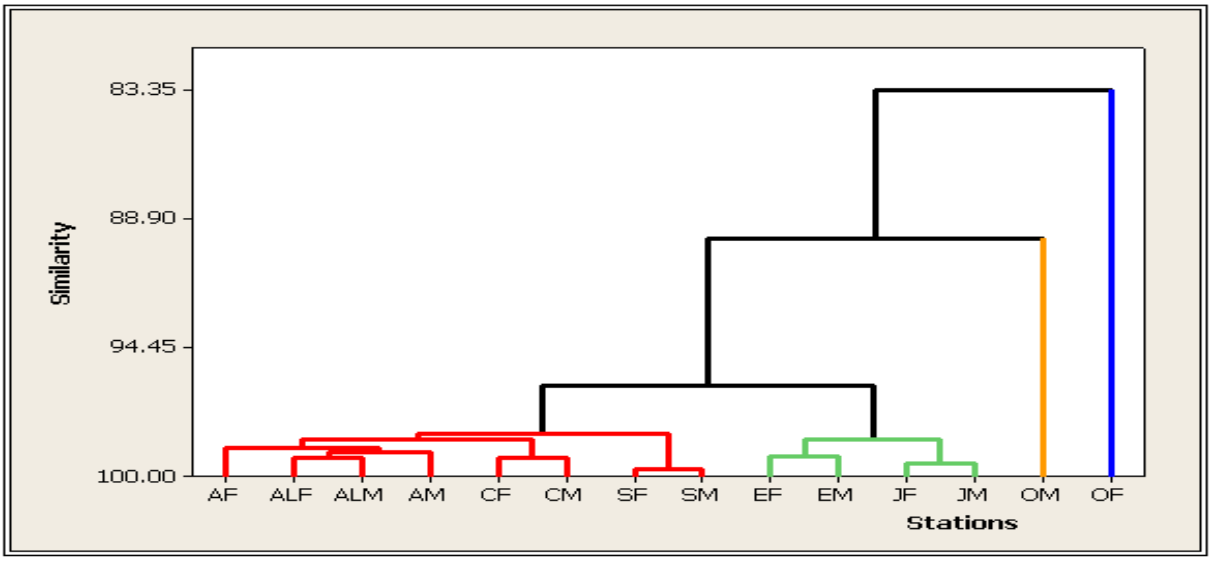

Fig. 5. Dendrogram of the grouping of both sexes in each of the 7 sites sampled, obtained using the simple link and the square distance of Pearson.

Hierarchical Analysis

The use of numerical classification methods, in addition to variance analysis, is uncommon (Dagnélie, 2006).

The bibliography concerning this approach is relatively limited (Cox and Cowpertwait, 1992; El Kherrak, 1993; Bautista *et al.*,1997;Di Rienzo *et al.*, 2002; Mezedjri, 2008).

The grouping of the two sexes in each of the 7 sampling sites according to the 36 morphometric and meristic variables, using a dendrogram using the simple link and the Pearson square distance and for a minimum level of similarity of 98.16%, distinguishes four distinct homogeneous groups that are (Fig. 5): first group composed by the females of Oran, second group composed by the males of Oran, third group relating to Jijel males, Jijel females, El-kala males and El-kala females, and Finally, a last group with a similarity level of 98.16%, is composed by: Annaba males, Annaba females, Algiers females, Algiers males, Collo males, females Collo, Skikda males and Skikda females.

In addition, the grouping of the variables using the simple link and the distance based on the correlation coefficient gives ten homogeneous groups, for a level of similarity of 98.31%, which are the following (Fig. 6). The first group consists of the following 27 variables: Lt, Lf, Ls, Lpan, Lppv, Lppc, Lcep, Lpdo, dopv, doan, doca, Lman, Poor, Dor, Pror, Lpop, Lain, Lcra, Mist, Lapc, Hpc, Hdo, Bado, baan, dopc, pcpv, pvan, with similarity levels close to 100%.

A second group composed by the variable Lmax.
A third group constituted by the variabe Hpv.
A fourth group composed by the variable Han.
A fifth group composed of the variable Hpdc.
A sixth group composed of the variable cæc.
A seventh group composed of the variable rypc.
An eighth group composed by the variable rypv.
A ninth group composed by the variable brsu.
And finally, a tenth group composed by the variable strand.

It should be noted here that this method illustrates well and highlights the relationships between the studied variables thanks to the correlation coefficient.

Discussion

Significant differences between the seven sites depend on several factors, which can be summarized as follows: The Algerian coastline is characterized by differences in climate, due to the orientation of the Algerian coast, which shows that the eastern region is more advanced towards the north than in the western region, which gives the advantage at the point of rainfall. And lower temperatures (Gulf of Collo) compared to the western region where this area characterized by low rainfall (Godbani, 2001), with the highest temperatures of the Algerian coast (Saada, 1997). What says rainfall, says abundance of nutrients, so phytoplankton and zooplanktonic populations and nutrient richness.

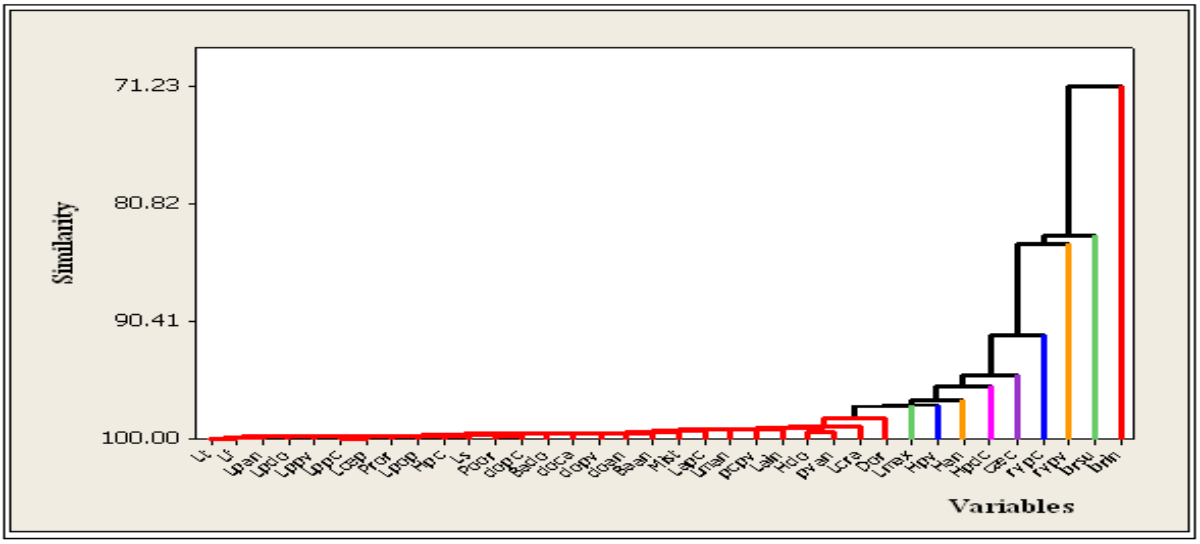

Fig. 6. Dendrogram of the grouping of the 36 variables, obtained with the help of the simple link and the distance of the correlation coefficient.

Nutrient inputs due to the modified Atlantic water would penetrate the Algerian coastline (Gulf of Oran especially) which gives rise to turbulences associated with upwelling, with all the consequences that engender on the biological productivity, thus the increase of trophic capacities of the environment (Millot, 1993). Thus, the characteristic hydrographic network in each region of the Algerian littoral plays a role in the nutrient abundance of the environment. The differences thus depend on the nature of the diet of these fish in each environment. The pollution factor and its impact on the marine environment, which concerns Skikda, Annaba, Algiers much more.

Conclusion

The multivariate statistical methods (PCA and hierarchical analysis) applied to the Saurels obtained from the seven sites of the Algerian littoral show that the principal component analysis (PCA) on the one hand, allowed to group together in the case of individuals (sites), five homogeneous groups, on the other hand to group the 36 variables selected into 4 homogeneous groups. The hierarchical classification methods make it possible to distinguish, for the

grouping of the 7 sites according to the 36 variables, four distinct homogeneous groups for a minimum level of similarity of 98.16%, and to group the studied variables into ten homogeneous groups.

References

Bautista MG, Smith DW, Steiner RL. 1997. A cluster-based approach to means separation. J. Agric. Biol. Environ. Stat. **2(2),** 198-211.

Bouroche JM, Saporata G. 1980. L'analyse des données. 1èr édition. Presses Académique de France. Paris, 127 p.

Boutarcha F. 2012. L'apport de la politique environnementale algérienne pour l'activité de la pêche : Cas de la Wilaya de Bejaia.p 181-182.

Cox TF, Cowpertwait PSP. 1992. Clustering population means under heterogeneity of variance. Statistician **41(5),** 591-598.

Dagnélie P. 1970. Théorie et méthodes statistiques : applications agronomiques **(2),** Gembloux, Pesses agronomiques, 451 p.

Dagnélie P. 2000. Statistique théorique et appliquée. Tomme 2 : Inférences à une et à deux dimensions. Bruxelles-université DE BOECK et LARCIER, 659 p.

Dagnélie P. 1986. Analyse statistique à plusieurs variables. Gembloux, Pesses agronomiques, p 362.

Dagnélie P. 2006. Statistique théorique et appliquée. Tomme 2 : Inférences à une et à deux dimensions. Bruxelles-université DE BOECK et LARCIER : p 659.

Di Rienzo JA, Guzman AW, Casanoves F. 2002. A multiple-comparison method based on the distribution of the root node distance of binary tree. J. Agric. Biol. Environ. Stat. **7(2),** 129-142.

El Kherrak H. 1993. Etude des méthodes de comparaisons multiples et de classification des moyennes en expérimentation agronomique. Gembloux, faculté des sciences agronomiques,196 p.

Food and Agriculture Organization. 2013.Species Fact Sheets Trachurus trachurus (Linnaeus, 1758), Food and Agriculture organization of the United Nations. Fisheries and Aquaculture Department.

Ghodbani T. 2001. Extension Urbanistique dans le Littoral d'Oran et ses Effets sur l'Environnement. Thèse de Magistère. Univsité d'Oran. Département de Géographie et de l'aménagement du territoire. p 378.

Lacaze JC. 1993. Les biocénoses marines et littorales de Méditerranée. Synthése, menace et perspective. Bellan-Santini D., Lacaze.J C., poizat C. eds. Secrétariat de la Faune et de la Flore. Muséum National d'Histoire Naturelle, Paris.

Mézédjri L. 2008. Modélisation de l'Impact de la Pollution Industrielle Hydrique dans le Golfe de Skikda (Littoral Est algérien). Thèse de Doctorat, Université Badji Mokhtar, Annaba.

Millot C. 1993. La circulation générale en méditerranée occidentale. Workshop sur la circulation des eaux et pollution des côtes méditerranéennes du Maghreb, 9-11, Novembre 1992. Doc. 29-30.

Palm R. 2000. L'analyse de la variance multivariée et l'analyse canonique discriminante principes et applications. Notes stat. Inform. (Gembloux) 2000 **(1),** p 40.

Palm R. 1998. L'analyse en composantes principales: Principes et applications. Notes stat.Inform. (Gembloux) **98(2),** p 33.

X. 2013. Minitab software statistique, version 16 pour windows.