Revue des Energies Renouvelables Vol. 16 N°3 (2013) 413 - 424

An approach to spatio-temporal analysis for climatic data

M. Mesri [1 *], A. Ghilane [1] and N.E.I. Bachari [2]

[1] Laboratoire des Semi-Conducteurs et Matériaux Fonctionnels,
Université Amar Télidji, B.P. 37 G, Laghouat, Algeria
[2] Faculté des Sciences Biologiques, Université des Sciences et de la Technologie
Houari Boumediène, USTHB, B.P. 32, El Alia, Bab Ezzouar, Algiers, Algeria

(reçu le 28 Avril 2012 – accepté le 29 Septembre 2013)

Abstract - *This work considers a large-scale multisite investigation of the effects of sunshine duration on Algerian territory. It also studies the correlation between sunshine duration and commonly used climatic parameters (temperature, water vapour pressure, evaporation, relative humidity and rainfall). Two data sets have served this purpose. The first one consists of sunshine duration measurements of the whole Algerian weather network over the period 1960-2002 while the second set provides climatic parameters collected in Dar El-Beida, Algiers (36° 41'N, 03°13' E) over the period 1950-2008. This station is the best in terms of fulfilling criteria such as long time data series (at least 50 years) as well as reliable measurements. To achieve the expected goal, we use an appropriate objective clustering method, named Principal Component Analysis (PCA) with a coupling of a Hierarchical Ascending Clustering (HAC) algorithm. PCA is often used not only for reducing the data before the actual clustering is carried out but also because it might help identifying the characteristics of the clusters. In this way, we establish a distribution of weather stations and identify the main homogeneous areas of the country that are more distinguishable and useful to our purpose as well as we classify the months of year. Obtained results divide Algeria into three distinct climate regions which daily monthly means of sunshine duration are also studied. Further information can be drawn through maps established by 'MapInfo' Software. According to the correlation matrix, it also turns out that sunshine duration strongly influences climatic parameters.*

Résumé - *Ce travail considère une enquête multi site à grande échelle des effets de l'insolation sur le territoire algérien. Il étudie également la corrélation entre l'insolation et les paramètres climatiques couramment utilisés comme la température, la pression de vapeur d'eau, l'évaporation, l'humidité relative et précipitations). Deux bases de données ont servi cet objectif. La première consiste en des mesures de l'insolation pour l'ensemble du réseau météorologique algérien couvrant la période 1960-2002, tandis que la seconde fournit des paramètres climatiques recueillies à Dar El-Beida, Alger (36°41'N, 03°13'E) sur la période 1950-2008. Cette station convient parfaitement en termes de réalisation des critères, tels que les séries de données de longues durées (au moins 50 ans), ainsi que la fiabilité des mesures. Pour atteindre l'objectif escompté, nous avons utilisé une méthode de classification objective appropriée, appelée Analyse en Composantes Principales (ACP) couplée à un algorithme de Classification Ascendante Hiérarchique (CAH). L'ACP est souvent utilisé non seulement pour réduire les données avant que le regroupement proprement dit n'est effectué, mais aussi parce que cela pourrait aider à identifier les caractéristiques des classes. Ceci a permis d'établir une distribution des stations météorologiques et d'identifier les principales zones homogènes du pays qui sont les plus distinguables et les plus utiles à notre fin. Les résultats obtenus divisèrent l'Algérie en trois régions climatiques distinctes pour lesquelles les moyennes mensuelles par jour de l'insolation sont également étudiées. Des informations complémentaires peuvent être apportées par des cartes géographiques établies par le logiciel MapInfo.*

* meradmesri@yahoo.fr; m.mesri@mail.lagh-univ.dz

Selon la matrice de corrélation, il s'avère également que l'insolation influe fortement sur les paramètres climatiques.

Keywords: Automated clustering - Hierarchical Ascending Clustering 'ACH' - Principal Components Analysis 'PCA' - Climatic parameters - Sunshine duration - Climatic regions - Correlation.

1. INTRODUCTION

Climate is still characterized by nonlinearity and high dimensionality. A challenging task is to find ways to reduce the dimensionality of the system and find the most important patterns explaining the variations. In the last several decades, meteorologists have in fact put major efforts in extracting important patterns from measurements of atmospheric variables. As a result PCA technique has become the most widely used way to do this [1, 2].

One of the major climatic parameters is sunshine duration as being linked to them through the solar activity. The need for a solid and reliable database is of paramount importance in both design and development of solar energy operating systems as well as in the assessment of their performance. This paper uses the historical observations in a few decades to investigate the sunshine duration patterns in Algeria and their connections to some common meteorological parameters (temperature, humidity, precipitation,...).

To achieve this goal, we use an appropriate objective or automated clustering method, named 'PCA' coupled with a HAC algorithm [3-5]. As a matter of fact, the literature is quite confusing when it comes to differences between the Empirical Orthogonal Functions 'EOF' method and the PCA method [6]. Although some authors [5] consider the two methods different, others use the PCA and EOF to describe the same method [7].

Since the literature is more or less in a state of confusion, we could use the terms EOF and PCA interchangeably. This produces new uncorrelated variables for clustering analysis. In this way, a classification of issued results allows to aggregate weather stations where are collected solar data having similar characteristics. The present work studies, by a coupling of PCA with a HAC, the correlation between sunshine duration and other common climatic variables. Accordingly, we need to disaggregate the outputs to find the connections with the local climatic parameters.

The station of Dar El-Beida that has served this purpose provides reliable measurements with little missing data. Basically, the clustering algorithms can be classified into two groups [8].This automated (objective) clustering method using PCA coupled with HAC has advantages including less time and labor. In contrast, there are some disadvantages of the other method called manual or subjective.

This cannot be duplicated by someone else its investigator and requires much time and labor. Some research works have been undertaken in Algeria and proposed in the literature [9, 10] since the automated (objective) methods have become more popular with the advance of computer technology [11, 12].

In the next section authors provide an explanation of the methodology applied to achieve the expected purposes. Results are then discussed and concluding remarks are made on.

2. METHODOLOGY

The first step concerns data preprocessing. This is a necessary stage to make data convenient with the calculation methods to be implemented in order to achieve the expected objectives. So, the stations of the National Office Meteorology with an important number of missing data are not considered (Fig. 1).

The next stage is to find a common period for weather stations. Our study is carried out over the period 1992-2002 that is a common and representative decade for the sunshine duration parameter (**Table 1**).

Another phase of this work concerns the description of the adopted spatiotemporal methods 'PCA' and 'HAC'. We highlight the capabilities of the techniques to characterize Algerian territory with respect to sunshine duration as well as to show the influence of this parameter on climatic parameters.

3. DATA PREPROCESSING

To better guide the clustering, it is wise to well prepare the data.

- Each station provides daily observations. Sometimes, missing values are not limited. This is true for the stations of El Kala, Ghazaouet and Illizi with 6147 and 4018 missing values respectively. However, the stations of Medea, Naâma and Mechria have only one, two and three missing values respectively, (Fig.1).
- The measurement period of sunshine duration parameter differs from one station to another as shown in **Table 1**. To bypass this situation we proceed as:
  - Select a common period for the sunshine duration database, [1992-2002],
  - Identify and inventory missing values for both databases under analysis,
  - Calculate the daily monthly mean for each weather parameter.

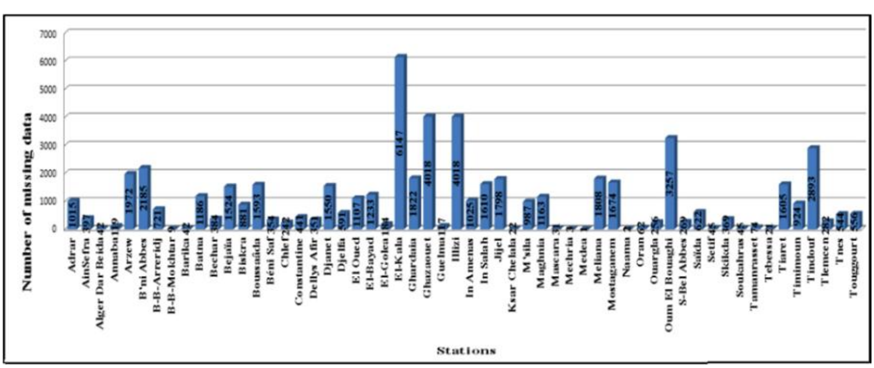

Fig. 1: Total number of missing data of the NOM

Table 1: An example of the measurement period of sunshine duration parameter [1960-2002]

Station	Period of measurement	Station	Period of measurement	Station	Period of measurement
Alger Dar El Beida	1960-2002	El Oued	1964-2002	Constantine	1960-2002
Maghnia	1976-2002	Médéa	1992-2002	Biskra	1965-2002
Bejaia	1968-2002	Jijel	1992-2002	Tlemcen	1980-2002

Table 2: Weather database for the location of Dar El Beïda

Dar El Beïda, Lat. 36°41'N, Lon. 03°13'E, Alt.25 m, Period 1950-2002

	Parameter	Description	Unit	Encoded missing data
Data measurement At NOM	T_{mean}	Average temperature	°C	-99,9
	R_{tot}	Total rainfall	0.1 mm	-9999
	R_{tot}	Sunshine duration	Tenth/h	-9999
	H_{mean}	Average humidity	%	-99.9
	T_{vap}	Average pressure of water vapor	hPa	-99.9
	Evap	Evaporation	0.1 mm	-99.9

The values of total rainfall RR<0.1 mm are encoded by 5555

4. DESCRIPTION OF METHODS

Let us call **Y**, the resulting matrix of individual's × variables. For the first data set, variables are daily monthly means of sunshine duration over the decade of study, arranged by months, whereas individuals are weather stations. For the second database we have six climatic parameters as variables and twelve months as individuals.

4.1 Descriptive statistics 'PCA'

Once the data set selected, we have to choose the type of the dispersion matrix to perform 'PCA'. According to ways that are used for scaling the data, one will determine the dispersion matrix: correlation matrix, covariance matrix or second moment matrix [12].

The correlation matrix suits perfectly in this work since the studied variables are either in different units '**Table 2**' or their variances differ widely '**Table 5**'. Then, we calculate factors that best explain all the observed variations in the database. A fairly detailed description of PC statistical analysis is given in an appendix of this manuscript.

A very good introduction of basic methods of climatic data analyses is emphasized in the text book by Thiebaux (1994) [2].

4.2 Ascending hierarchical clustering

In order to ensure indexed hierarchies [11] we need some way of measuring dissimilarity from a group to another. One of the measures that fit the purpose is the Euclidean distance used in this work.

The class aggregation criterion is that of Ward and the method seeks to minimize intra-class inertia which is to maximize the interclass inertia. Interclass variance measures the dispersion of classes whereas the intra-class variance measures the dispersion of individuals of the same class.

A dendrogram allows visualizing the gradual grouping of data and shows so clearly how the algorithm proceeds to aggregate individuals and subgroups. Detailed steps of HAC method are described in an appendix of this manuscript.

5. RESULTS AND DISCUSSION

5.1 Spatio temporal analysis of sunshine duration

To determine the principal components, derived information are calculated from the daily monthly mean of sunshine duration over the period 1992-2002 as represented in '**Table 3**'. They are used to calculate the associated correlation matrix R '**Table 4**'. This matrix is symmetric and has 12 × 12 elements. The diagonal elements represent variances whereas the others are co-variances.

Table 3: Minimum, maximum, mean and standard deviation of sunshine duration [1992-2002]

Variable	Minimum	Maximum	Mean	Standard deviation
January	4.42	9.30	6.59	1.33
February	4.36	10.13	7.73	1.19
March	5.47	9.61	8.13	0.86
April	6.15	10.75	9.10	**0.85**
May	7.36	11.45	9.65	0.71
June	6.13	11.42	10.39	0.90
July	4.89	11.83	10.81	**1.05**
August	4.83	11.17	9.82	0.94
September	5.41	10.13	8.64	**0.78**
October	4.97	9.74	7.92	0.88
November	3.80	9.52	6.81	1.30
December	4.19	8.90	6.34	**1.36**

Table 4: Correlation Matrix

Varia-	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Jan-	**1**	0.96	091	076	0.54	0.09	-0.12	-0.01	0.26	0.9	0.98	0.97
Feb-	0.96	**1**	0.96	0.86	0.64	0.18	-0.05	0.03	0.4	0.95	0.97	0.94
Mar-	0.91	0.96	**1**	0.90	0.75	0.32	0.09	0.18	0.49	0.95	0.91	0.89
Apr-	0.76	0.86	0.90	**1**	0.89	0.49	0.12	0.25	0.66	0.85	0.8	0.77
May	0.54	0.64	0.75	0.89	**1**	0.73	0.45	0.45	0.77	0.74	0.59	0.57
Jun-	0.09	0.18	0.32	0.49	0.73	**1**	0.86	0.83	0.78	0.31	0.12	0.12
Jul-	-0.12	-0.05	0.09	0.2	0.45	0.86	**1**	0.96	0.69	0.12	-0.09	-0.12
Aug-	-0.01	0.03	0.18	0.25	0.45	0.83	0.96	**1**	0.69	0.17	-0.01	-0.03
Sep-	0.26	0.4	0.49	0.66	0.77	0.78	0.69	0.69	**1**	0.53	0.33	0.27
Oct-	0.9	0.95	0.95	0.85	0.74	0.31	0.12	0.17	0.53	**1**	0.93	0.9
Nov-	0.98	0.97	0.91	0.8	0.59	0.12	-0.09	-0.01	0.33	0.93	**1**	0.98
Dec-	0.97	0.94	0.89	0.77	0.57	0.12	-0.12	-0.03	0.27	0.9	0.98	**1**

Eigenvectors of the correlation matrix are ranked in descending order of their Eigen values. Each eigenvector u_i is a principal axis and each normalized eigenvector $u_i^{'}$ is a principal factor expressed by {Eq. A5}. The explained variance or variability (%) is calculated for the twelve principal components as well as the accumulation percentage of the variance. These results are reported in **Table 5**.

The two first Eigen values (the first factorial plane) account for themselves 91.20 % of the total variance of the grid points, consisting of 53 × 12 solar data. This implies that

the data can be majorly reconstructed using only two patterns. What is remaining '8, 80%' is distributed between the ten other components.

Table 5: Eigen values, variability and accumulation

N° of ordre (i)	F1	F2	F3	F4	F5	F6
Eigen value λ_i	**7.493**	**3.451**	0.495	0.226	0.118	0.086
Variability vari (%)	**62.44**	**28.76**	4.13	1.89	0.99	0.72
Accumulation of variances C_m	**62.44**	**91.2**	95.33	97.21	98.2	98.92
N° of ordre (i)	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**
Eigen value λ_i	0.054	0.032	0.017	0.012	0.008	0.006
Variability vari (%)	0.45	0.27	0.14	0.1	0.07	0.05
Accumulation of variances C_m	99.37	99.64	99.78	99.88	99.95	100

All variables (Fig. 2) are well represented on the factorial plane as they are close to the circumference.

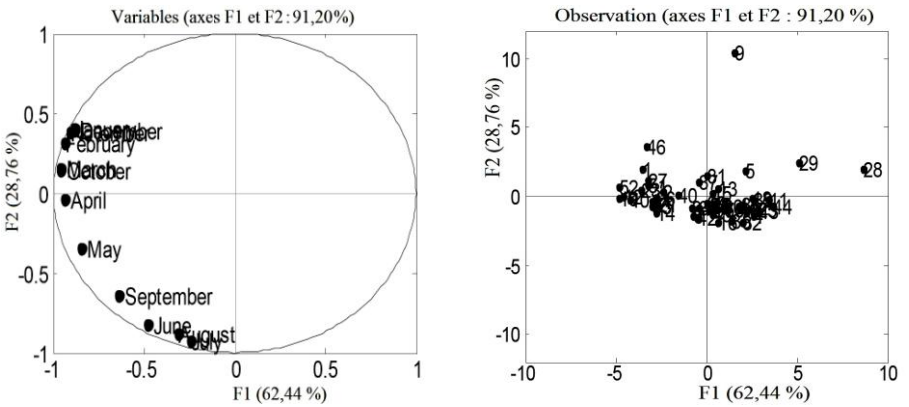

Fig. 2: **a-** Graphical representation of variables
b- Graphical representation of individuals

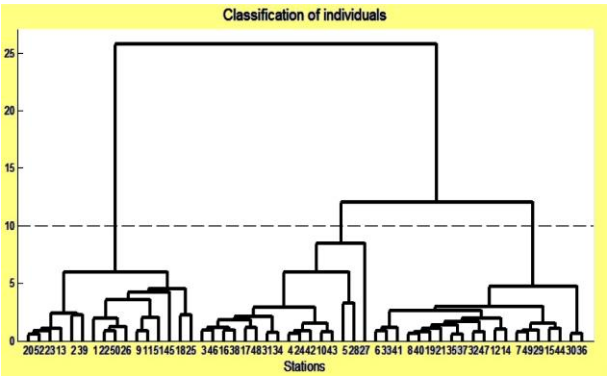

Fig. 3: Dendrogram of the Hierarchical Ascendant Clustering Method

Figure 3 is issued from the implementation of the HAC to the obtained results by PCA.

It provides insight into the data by assembling all the objects into a dendrogram so that each sub-cluster is a node of the dendrogram and the combinations of sub-clusters create a hierarchy that is more informative.

To better understand and facilitate interpretation, representations in the form of maps have been established using the georeferenced software Mapinfo and the daily monthly mean of sunshine duration for each class is then represented.

- Class N°1 extends over eastern and central parts where most of which are coastal (35° < latitude < 37°) (Fig. 4.a-). The daily monthly average of sunshine duration varies between 3.8 and 11.5 hours and its distribution differs during the year relatively to winter and summer (Fig. 4.b-).

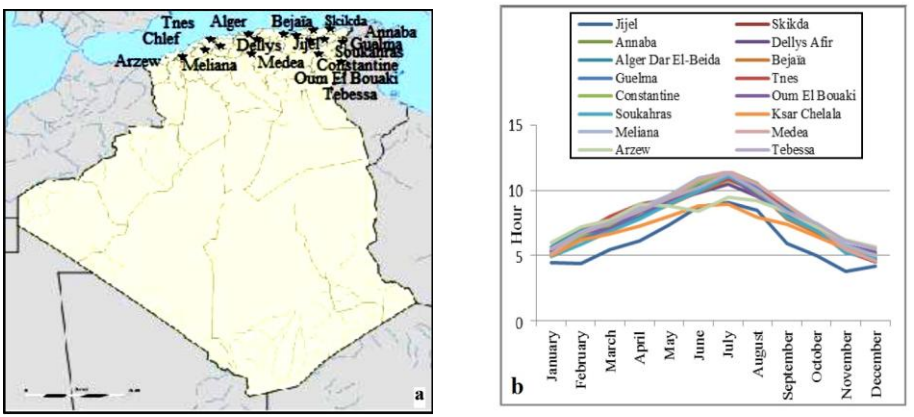

Fig. 4: **a-** Map of weather stations of the 1st class;
b- Monthly mean of sunshine duration of the 1st class [1992 - 2002]

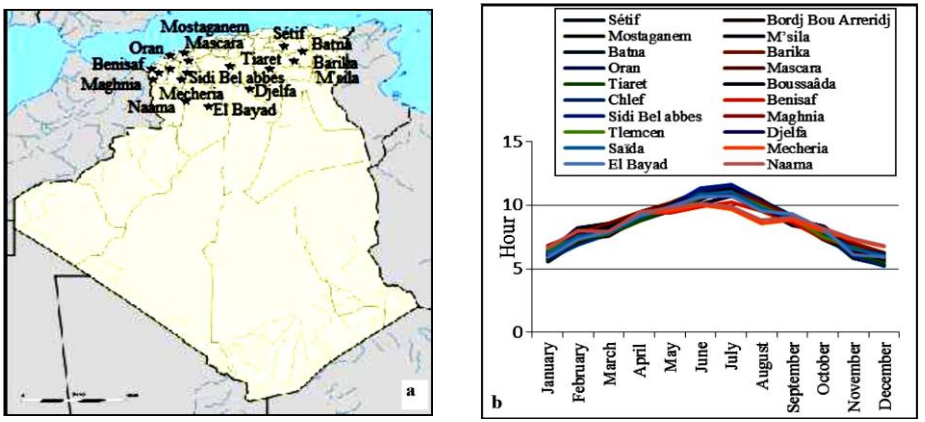

Fig. 5: **a-** Map of weather stations of the 2nd class;
b- Monthly mean of sunshine duration of the 2nd class [1992 - 2002]

- Class N°2 contains the coastal regions of the West and all the highlands as shown in (Fig. 5.a-). It is characterize d by a daily monthly sunshine duration average which varies from 5.2 to 11.6 hours. Summer is the sunniest season as in the class N°1 (Fig. 5.b-).
- Class N°3 represents the south of the country (Sahara), (21° < latitude < 35°) as shown in Fig. 6.a-. The daily monthly mean of sunshine duration is very high, almost uniformly distributed throughout the year (Fig. 6.b-).

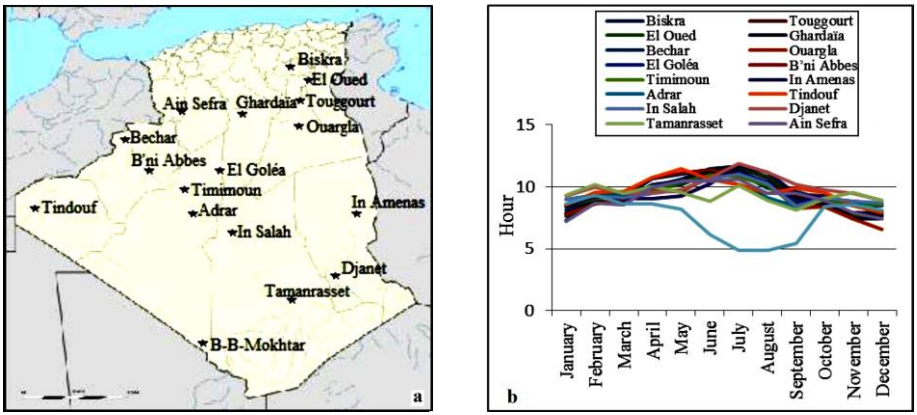

Fig. 6: **a-** Map of weather stations of the 3rd class;
b- Monthly mean of sunshine duration of the 3rd class [1992 – 2002]

5.2 Spatio temporal analysis of weather parameters

Let us remind that this study concerns six climatic parameters collected in the station of Dar El-Beida, Algiers over the period 1950 - 2008. On the basis of the resulting standard deviations σ in **Table 6**, one can see that the two variables R_{tot} and Evap (σR_{tot} = 17,770, σ Evap = 10,822) are responsible of the dispersion of the grid points.

Table 6: Minimum, maximum, mean and standard deviation of weather parameters [1950-2008]

Variable	Minimum	Maximum	Mean	Standard deviation
T_{mean}	10.503	26.023	17.949	5.566
R_{tot}	0.645	51.567	21.010	**17.770**
Inso	4.481	11.956	8.348	2.255
H_{mean}	62.000	74.517	69.115	3.445
T_{vap}	10.063	22.465	15.245	4.437
Evap	16.633	48.000	34.458	**10.822**

As shown in **Table 7**, the first factorial plane (1-2) representing 93.222 % of the inertia is sufficient for a better representation of individuals and variables since there is a loss of information (6.778 %) that could be neglected.

Table 7: Eigen values, variability and accumulation

	F1	F2	F3	F4	F5	F6
Eigen value	**4.928**	**0.665**	0.278	0.063	0.047	0.019
Variability (%)	82.139	11.083	4.630	1.047	0.785	0.320
Accumulation %	82.139	**93.222**	97.848	98.895	99.680	100.000

Table 8: Correlation between variables and principal components

	F1	F2	F3	F4	F5	F6
T_{mean}	0.958	0.033	0.215	0.065	0.105	-0.081
R_{tot}	-0.939	-0.195	0.224	-0.02	0.134	0.061
Inso	0.815	0.544	-0.149	-0.041	0.089	0.053
H_{mean}	-0.869	0.436	0.202	0.131	-0.032	0.047
T_{vap}	0.925	0.066	0.343	-0.021	-0.095	-0.061
Evap	0.923	-0.369	-0.018	0.197	0.014	0.006

Analysis of the results given in **Table 8** reveals that:

- The first principal component is characterized by:
  - A significant positive correlation between the variables T_{mean}, Evap, T_{vap} and Inso,
  - A strong negative correlation between the two variables R_{tot} and H_{mean}.
- The second principal component is characterized by:
  - A medium positive correlation between the variables Inso and H_{mean},
  - A small positive correlation between the variables and T_{mean} and T_{vap},
  - A small negative correlation between the variables R_{tot} and Evap.

By applying a HAC, months are classified as reported in Fig. 7. They are represented by their corresponding numbers

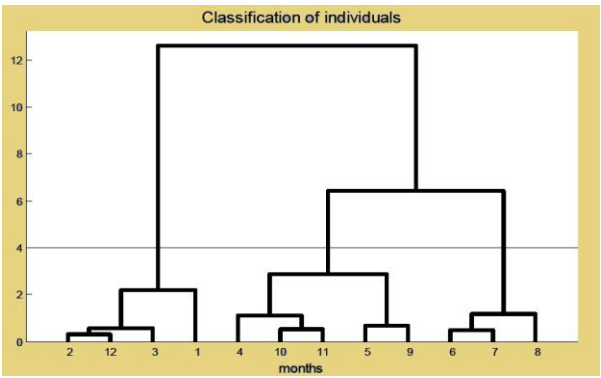

Fig. 7: Classification of individuals

- Class 1: includes the months of winter (December, January, February and March)
- Class 2: includes the months of Spring and Autumn,
- Class 3: includes the months of summer (June, July and August)

6. CONCLUSION

Actually, the application of solar energy to a given site requires comprehensive and detailed knowledge of the sunshine duration parameter. This is generally easier when the site has a radiometric station running regularly for several years.

Fifty-eight years of historical daily climatic data of the northern station of Dar El Beida, as well as eleven years of sunshine duration data all over Algeria have been analyzed through a PCA coupled with HAC.

The obtained results in the present work are convincing, as they reveal a good analogy and homogeneity of extracted energy areas and turn out that sunshine duration strongly influences climatic parameters.

It should be stressed that even though the present study is based on practical results, a number of questions are raised and may provide opportunities to pursue other work.

They may be:

▪ PCA technique might be used to reconstruct the scope of sunshine duration at any point of the territory covered by the meteorological network. Data produced by PCA can be used to more easily generate hourly monthly means variations of global solar radiation flux on a tilted collector [13].

▪ Weather forecast can be modeled adequately using Markov model [14].

REFERENCES

[1] S. Tolwinski, '*Statistical Methods for the Geosciences and Beyond*', University of Arizona, RTG Project, Fall 2007.

[2] H.J. Thiebaux, '*Statistical Data Analyses for Ocean and Atmospheric Sciences*', Academic Press, 1994.

[3] J. Key and R.G. Crane, '*A Comparison of Synoptic Classifications Schemes Based on "Objective Procedures*', Journal of Climatology, Vol. 6, pp. 375 – 386, 1986.

[4] I.T. Jolliffe, '*Principal Component Analysis*', Second Edition, Springer, 2002.

[5] M.B. Richman, '*Rotation of Principal Components*', Journal of Climatology, Vol. 6, N°3, pp. 293 – 335, 1985.

[6] A. Hannachi, '*A Primer for EOF Analysis of Climate Data*', Reading RG6 6BB, U.K, 2004.

[7] R.W. Preisendorfer, '*Principal Component Analyses in Meteorology and Oceanography*', Elsevier, 1988

[8] A. Mefti and M.Y. Bouroubi, '*Estimation et Cartographie de la Composante Global du Rayonnement Solaire*', Revue des Energies Renouvelables, 219-224, 1999.

[9] A. Mefti, '*Contribution à la Détermination du Gisement Solaire par Traitement de Données Solaires au Sol et d'Images Météosat*', Doctorant Thesis, USTHB, 2007

[10] J. Wu, H. Xiong and J. Chen, '*Towards Understanding Hierarchical Clustering: A Data Distribution Perspective*', Journal Neurocomputing, Vol. 72, N°10-12, pp. 2319 - 2330, 2009.

[11] B. Yarnal, D.A. White and D.J. Leathers, '*Subjectivity in a Computer-Assisted Synoptic Climatology II: Relationship to Surface Climate*', Journal of Climatology, Vol. 8, N°3, pp. 227 - 239, 1988.

[12] A. Sfetsos and A.H. Coonick, '*Univariate and Multivariate Forecasting of Hourly Solar Radiation with Artificial Intelligence Techniques*', Solar Energy, Vol. 68, N°2, pp. 169 - 178, 2000.

[13] A. Maafi A., '*Markov-Models in discrete time for solar radiation*', In Proceedings of Multi-Conference on Computational Engineering in Systems Applications, IMACS-IEEE, Hammamet, Tunisia, Vol. 2, pp. 319 – 322, 1998.

APPENDIX A (Steps of performing PCA method)

Let's Y be the matrix of individuals

$$Y = \begin{bmatrix} y_1^1 & \cdots & y_1^2 & \cdots & y_1^p \\ y_2^1 & \cdots & y_2^2 & \cdots & y_2^p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_n^1 & \cdots & y_n^1 & \cdots & y_n^p \end{bmatrix} \quad \text{(A1)}$$

$\tilde{Y}^T$ is the transposed of the centered and reduced matrix $\tilde{Y}$ given by

$$\tilde{Y} = \begin{bmatrix} (y_1^1 - \overline{Y}_1)/\sigma(Y_1) & \cdots & (y2 - \overline{Y}_2)/\sigma(Y_2) & \cdots & (y_1^p - \overline{Y}_p)/\sigma(Y_p) \\ (y_2^1 - \overline{Y}_1)/\sigma(Y_1) & \cdots & (y_2^2 - \overline{Y}_2)/\sigma(Y_2) & \cdots & (y_2^p - \overline{Y}_p)/\sigma(Y_p) \\ \cdots & & \cdots \cdots & & \cdots \cdots \\ (y_n^1 - \overline{Y}_1)/\sigma(Y_1) & \cdots & (y_n^2 - \overline{Y}_2)/\sigma(Y_2) & \cdots & (y_n^p - \overline{Y}_p)/\sigma(Y_1) \end{bmatrix} \quad \text{(A2)}$$

Where, $\sigma(Y_i)$ represents the standard deviation of column i, it is calculated by the equation (A3),

$$\sigma(Y_i) = \sqrt{\frac{1}{n} \times \sum_{j=1}^{n} \left(y_j^i - \overline{Y}_i\right)^2} \quad \text{(A3)}$$

And, $\overline{Y}_i$ the mean of column i, is calculated by equation (A4),

$$\overline{Y}_i = \frac{1}{n} \times \sum_{j=1}^{n} y_j^i \quad \text{(A4)}$$

The correlation matrix R is obtained as follows

$$R = \tilde{Y}^T \tilde{Y} = \begin{bmatrix} \tilde{y}_1^1 & \cdots & \tilde{y}_2^1 & \cdots & \tilde{y}_n^1 \\ \tilde{y}_1^1 & \cdots & \tilde{y}_2^2 & \cdots & \tilde{y}_n^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \tilde{y}_1^{p1} & \cdots & \tilde{y}_{2\cdots}^{p} & & \tilde{y}_n^p \end{bmatrix} \times \begin{bmatrix} \tilde{y}_1^1 & \cdots & \tilde{y}_2^1 & \cdots & \tilde{y}_n^1 \\ \tilde{y}_1^1 & \cdots & \tilde{y}_2^2 & \cdots & \tilde{y}_n^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \tilde{y}_1^{p1} & \cdots & \tilde{y}_{2\cdots}^{p} & & \tilde{y}_n^p \end{bmatrix} \quad \text{(A5)}$$

Principal axes are the eigenvectors u_i of the correlation matrices. Let us considered I, the identify matrix, the eigenvalues are solutions of the equation (A6).

$$\det(R - \lambda \times I) = 0 \quad \text{(A6)}$$

Variability Var_i and accumulation Cu_i in % are calculated using the eigen value λ_i as given by Equation (A6).

$$Var_i = \frac{\lambda_i}{\sum_i \lambda_i} \quad \text{(A7)}$$

$$Cu_i = Cu_{i-1} + Var_i \quad \text{(A8)}$$

The principal factor u_i' associated with the principal axis u_i is the normalized eigenvector,

$$u_i' = \frac{u_i}{\| u_i \|} \quad \text{(A9)}$$

where, $\| u_i \|$ is the square norm of the eigenvector.

Correlation between variables and principal components is noted Cv_i and is obtained by equation (A10).

$$Cv_i = \sqrt{\lambda_i} \times u_i' \quad \text{(A10)}$$

Principal components are the variables $C^i \in R^n$ defined by the principal factors by the relation,

$$C^i = \tilde{Y} \times u_i' \quad \text{(A11)}$$

$$Var(C^1) = \lambda_i \quad \text{(A12)}$$

One of the expected objectives of PCA is graphical representation of individuals. It is used to represent individuals on a map in two dimensions (Fig. 2) and thus to identify trends, coordinates of individuals are obtained according to equation (A11).