Exploratory analysis of Pacific Ocean data to study "El Niño" phenomenon

Sergio Camiz[a], Jean Jacques Denimal[b] y Wilfredo Sosa[c]
[a]*Dipartimento di Matematica Guido Castelnuovo, Sapienza Università di Roma, Italia,*
e-mail: sergio.camiz@uniroma1.it;
[b]*UFR de Mathématiques Pures et Appliquées, Università des Sciences et Technologies de Lille, Francia,*
e-mail: jean-jacques.denimal@univ-lille.fr;
[c]*Instituto de Matemática y Ciencias Afines - Universidad Nacional de Ingenieria, Perú,*
e-mail: sosa@uni.edu.pe

Recibido el 03 de marzo del 2010; aceptado el 12 de abril del 2010

Aiming at studying the "El Niño" phenomenon on the basis of the available data, we started an exploratory analysis of the set of surface temperature time-series produced from the *USA*'s National Oceanic and Atmospheric Administration (*NOAA*). The first results of Principal Component Analysis and Hierarchical Factor Classification applied on the data set relative to the period 1991-2008 are reported. Together with the regular seasonal fluctuation and the subdivision in 11 classes of the time-series, all spacially connected but two, the occurrence of El Niño in 2007 results from the data as a very strong perturbation of an otherwise very regular pattern.

Keywords: "El Niño/La Niña" South Pacific Oscillation, Time-series, Principal Component Analysis, Hierarchical Factor Classification, Classification of variables.

Con el objetivo de estudiar el Fenómeno del Niño a partir de bases de datos disponibles, nosotros comenzamos un análisis exploratorio de la temperatura superficial de un conjunto de series de tiempo obtenidas de la Administración Nacional Oceánica y Atmosférica (NOAA) de EEUU. Los primeros resultados del Análisis de Componentes Principales y de la clasificación jerárquica de factores aplicada sobre el conjunto de datos relativos al período 1991-2008 son reportados. Conjuntamente con la fluctuación regular estacional y la subdivisión en 11 clases de las series temporales, todas conectadas espacialmente, la ocurrencia de El Niño en 2007 resulta de los datos como una perturbación muy fuerte dentro de un patrón muy regular.

Palabras Claves: La Oscilación del Pacífico Sur El niño/La niña, Series de Tiempo, Análisis de Componentes Principales, Clasificación Jerarquica de Factores, Clasificación de Variables.

1. Introduction

"El Niño" and "La Niña" are part of the climate cycle referred to as the El Niño Southern Oscillation (*ENSO*). During El Niño, warmer than average sea surface temperatures occur in the Equatorial central and eastern Pacific while during La Niña, cooler than average sea surface temperatures predominate. The Southern Oscillation ("*SO*" in *ENSO*) represents the atmospheric component of the cycle in which lower (higher) than normal sea-level pressure occurs near Tahiti and (higher) lower sea-level pressure occurs in Australia during El Niño (La Niña) conditions. *ENSO* is an important component of the climate system since the El Niño/La Niña phases impact the weather on a global scale.

The impact of *ENSO* sea surface temperatures (*SST*s) on the atmosphere is through the tropical response of rain-producing convection and cloud formation, the principal agents for exchanging heat from Earth's surface. Normally, the *SST* is very warm in an area that covers the Equatorial Indian and West Pacific Ocean regions.

During El Niño, among its consequences are the increased rainfall across most of the Americas' Western belt, ranging from South *USA* through North of Chile, which has caused destructive floodings, and drought in the West Pacific, sometimes associated with devastating brush fires in Australia. Observations of conditions in the Tropical Pacific are considered essential for the prediction of short term (a few months to 1 year) climate variations. The opposite occurs during La Niña, with rainfall deficits in the Eastern Equatorial Pacific and the wet conditions confined to the Western Equatorial Pacific.

Our interest is the analysis of the impact of El Niño in Peru and for this task we started collecting the necessary data and analyzing them through exploratory data analysis techniques. In this paper we introduce the data, the exploratory methods used so far, and the first results that we obtained from the first analyses performed.

2. The data

The first data base that we found through Internet of some interest for our purpose is the one provided by the United States' National Oceanic and Atmospheric Administration (*NOAA*). *NOAA* operates an array of 88 buoys almost regularly placed on a regular grid of 8 x 11 nodes in the Equator belt of Pacific Ocean. The buoys measure temperature, currents and winds in the equatorial band and transmit daily the data which are collected, checked for quality, adjusted when necessary, and made available to researchers and forecasters around the world. The distribution of buoys is shown in Figure 1.

For our work, we limited our attention to the Ocean's surface temperature that we downloaded from *NOAA*'s web site (www.pmel.noaa.gov/tao). They were 88 time series of daily surface temperature taken from March 1st, 1980 through December 31st, 2008. Of these, 20 series were empty and we withdrew them. The other 68 time-series are all nearly continuous since 1991, whereas only 27 had data in the previous period. Thus, we decided, as a preliminary task, to limit the study on the 68 time-series and on the period 1991-2008, for a total of 6575 daily observations. To this data base we added, as nominal characters to use as supplemental elements, the year and the month of sampling, together with a combination of year and season, to follow in a medium detail the overall evolution of the temperatures.

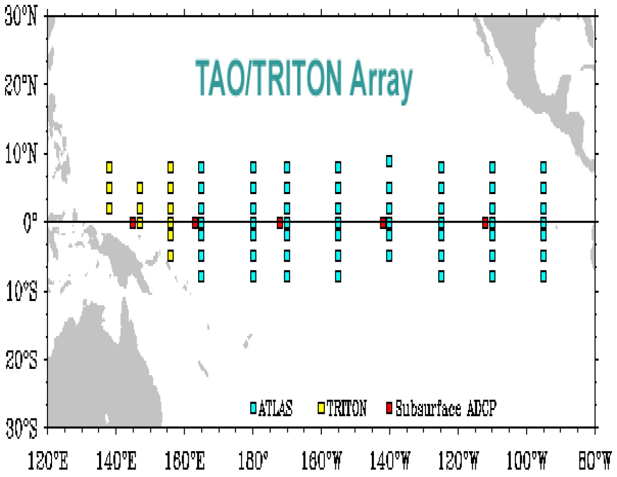


Figure 1. *The geographical position of the TAO/TRITON array of buoys implemented by the NOAA.*

In the following Table, the univariate statistics of all 68 time-series are reported. We submitted this data table to multidimensional data analysis techniques, namely Principal Components Analysis and Hierarchical Factor Classification, two techniques that may well be interlaced among them, in order to have a first picture of the Pacific Ocean surface temperature pattern.

2.1. Principal Component Analysis

Principal Component Analysis (*PCA*, Langrand and Pinzón, 2009; Benzécri et al., 1982; Jolliffe, 1986; Legendre and Legendre, 1998) is a classical exploratory analysis tool, that aims at synthesizing a quantitative (ratio-scale) data table by searching a reduced dimensional representation that summarizes most of the data variation, in the sense of the points inertia around the centroid-origin. In this way, both continuous characters and units may be represented on graphics in which their position reflects respectively the factors' values for the units and the correlation with the factors for the characters.

Table 1. *Univariate statistics of the 68 time-series of temperatures used for our study: Name, Number of non-missing values, mean, standard deviation, minimum, and maximun.*

Site	Freq	Mean	St.Dev.	MIN	MAX
9N140W	6551	27,44	0,87	24,41	29,38
8N137E	2651	29,13	0,62	26,75	30,81
8N156E	4355	29,07	0,54	27,36	30,64
8N156E	4355	29,07	0,54	27,36	30,64
8N165E	6279	28,84	0,59	26,61	30,77
8N180E	4828	28,57	0,64	26,57	30,48
8N170W	5797	28,39	0,67	26,57	30,19
8N155W	5495	27,97	0,76	25,46	29,83
8N125W	5221	27,69	0,63	25,25	29,39
8N110W	5785	27,91	0,75	25,48	30,08
8N095W	4901	27,75	1,04	24,51	30,97
5N137E	3067	29,28	0,59	27,12	30,81
5N147E	5775	29,37	0,49	27,54	30,82
5N156E	5277	29,38	0,47	26,15	30,82
5N165E	6400	29,21	0,54	27,38	30,76
5N180E	5297	28,90	0,67	25,93	30,89
5N170W	5588	28,67	0,76	25,38	30,46
5N155W	6056	28,08	0,86	24,41	30,16
5N140W	6266	27,54	0,94	23,67	29,85
5N125W	5763	27,30	1,00	22,98	29,88
5N110W	5897	27,41	0,98	22,90	29,93
5N095W	5043	27,62	0,90	24,75	30,55
2N137E	5261	29,49	0,56	27,18	30,80
2N147E	3958	29,69	0,41	28,27	31,01
2N156E	5741	29,50	0,52	27,30	30,95
2N165E	6360	29,35	0,67	27,01	31,03
2N180E	5530	28,73	0,99	25,28	30,88
2N170W	5149	28,17	1,13	24,64	31,16
2N155W	5586	27,59	1,12	23,09	30,34
2N140W	6110	26,77	1,32	20,45	30,11
2N125W	5378	26,20	1,49	20,74	29,44
2N110W	6150	26,05	1,68	19,84	30,49
2N095W	4281	26,48	1,49	20,85	30,94
0N137E	948	29,64	0,42	28,36	30,92
0N147E	4911	29,70	0,45	27,63	31,17
0N156E	5507	29,55	0,57	27,25	30,95
0N165E	5469	29,36	0,77	26,23	30,99
0N180E	5363	28,49	1,14	25,12	30,76
0N170W	6460	27,98	1,23	24,09	30,97
0N155W	6324	26,96	1,35	22,56	30,60
0N140W	6051	25,99	1,52	20,57	30,07
0N125W	5692	24,87	1,76	19,84	29,90
0N110W	6061	24,22	2,15	17,51	30,09
0N095W	4717	23,88	2,53	18,16	30,88
2S156E	5841	29,64	0,53	27,11	31,29
2S165E	6076	29,60	0,65	26,58	31,22
2S180E	5750	28,96	0,95	25,27	31,15
2S170W	5758	28,52	1,00	24,85	31,14
2S155W	5513	27,54	1,16	23,45	30,39
2S140W	6186	26,64	1,43	21,13	30,41
2S125W	6235	25,64	1,59	20,92	30,20
2S110W	5905	24,62	2,04	19,35	30,15
2S095W	4930	23,93	2,66	17,43	30,33
5S156E	5828	29,66	0,57	27,06	31,28

Through *PCA* the factors, that are linear combinations of the original characters, may be interpreted on the basis of the characters whose coefficients are higher and of those most correlated with them. As for the units, their position in the factor space reflects their score in the ordination given by each factor, a kind of compromise between the scoring of the characters that contribute or are correlated with them.

3. The methodology

The *PCA*'s rationale is based on the Singular Value Decomposition (*SVD*, Greenacre, 1984) of the data table and its strict relation with the eigendecomposition of the correlation matrix between the data table columns. We start with an $n \times p$ data matrix X, whose n rows $x_i, i = 1, \ldots, n$ represent the values taken by all characters in each of the units and whose p columns $X_j, j = 1, \ldots, p$ represent the values taken by each of the characters on all units. The matrix is first transformed by standardizing the columns, that is by centering them to the respective mean and dividing them by their respective standard deviation, in symbols $x_{ij} \rightarrow z_{ij} = \frac{x_{ij} - \bar{X}_j}{\sigma_j}$ so that each column's mean and variance become 0 and 1 respectively. Then, through *SVD* the so built matrix Z is decomposed as $Z = U\Lambda^{\frac{1}{2}}V'$, where U and V are the symmetric orthogonal matrices of the eigenvectors of $\frac{1}{n}Z'Z$ and $\frac{1}{n}ZZ'$ respectively, with $UU' = I_n$ and $VV' = I_p$, and Λ the diagonal matrix of the corresponding eigenvalues of both, all non-negative, sorted in decreasing order.

Thanks to the decomposition, the units' coordinates on each factor result as the columns of $U\Lambda^{\frac{1}{2}}$, whose variance equals the corresponding eigenvalue λ and the coordinates of the characters as the columns of V. As they are orthogonal, the coordinates of the units on the factors are uncorrelated among them, and it results that the amount of inertia along each factor equals the corresponding eigenvalue. Thus, its importance may be measured by its share to the total table inertia, given by the ratio of the eigenvalue to *trace* $(\Lambda) = p$. The Eckart and Young (1936) theorem ensures that the best reduced rank reconstruction of the data matrix, in the least-squares sense, is obtained by limiting the data table reconstruction to the first larger eigenelements. It must be reminded that, in decreasing order, the coordinates of the units along each factor are the best approximation of the values of the original characters and that the cosines of the angles among the characters are the best approximation of their correlation in the reduced dimensional spaces.

For the interpretation of the *PCA* results, the contribution given by each character to the linear combination that defines each eigenvector and the correlation between characters and factors are the most important issues. Then, the eigenvalues and their percentage of explained inertia are useful to identify the amount of total information interpreted. This information is currently taken into account to decide the most suitable reduced dimension for the interpretation. Indeed, this is still an issue debated in literature (see Jackson, 1993 and Peres-Neto et al., 2005, for reviews) and we did not take a decision in this sense: in this paper we shall evoke the first four dimensions, interpreting only the first two, just as a provisional examination, with no claim to be exhaustive.

Indeed, on the factor spaces other characters, both continuous and nominal, and other units may be projected as supplemental elements, based on their behaviour in respect to the active elements: the continuous characters are projected on the circle of correlations according to their correlation with the axes; each level of the nominal ones is represented at the centroid of the units that take that level as observed value. As the supplemental characters do not participate to the eigenvectors construction, they are useful as external references in the factors interpretation. We shall take advantage of this feature to include in the graphical representation both the time nominal characters and the *HFC* representative variables of the classes, in order to synthesize the results and ease its interpretation.

3.1. Hierarchical Factor Classification

Hierarchical Factor Classification of continuous characters (*HFC*, Denimal, 2001; Camiz et al., 2006; Camiz and Pillar, 2007) is a new method that aims at combining the classification of characters, a task neither very developped nor very used in literature, with the factorial methods in the same exploratory spirit of *PCA*; it is easy to use, and its results are immediately understandable by a non-particularly specialized user. Originally introduced by Denimal (2001), it combines classification and ordination in a single procedure, so that it outputs at the same time a hierarchy and a set of principal planes associated to the hierarchy's nodes. The association among characters is based on their reciprocal covariance and, for each node, the method provides a principal plane where both characters and units can be represented. This is certainly an advantage for the user accustomed to *PCA* and subsequent hierarchical clustering, in that the interpretation of the groups of characters and of the principal components becomes easy, and the units can as well be classified at each step according to the found differences among the characters. Since the method is based on the same geometric space as *PCA*, the resulting principal components can be represented as supplemental elements in the *PCA* principal planes. This allows an interoperability between the two methods.

In *HFC* one deals with the same set of p quantitative (ratio-scale) standardized characters Z. The method operates as follows:

1. at the beginning each character is considered a representative variable of the singleton group composed by itself. Then the recursive algorithm is based on the following steps:

2. all pairs of existing groups are compared, through their representative variable: each pair of representative variables is submitted to a non-normalized *PCA*, i.e. the *PCA* of their 2×2 covariance matrix. It must be pointed out that in the case of standardized characters this equals the correlation matrix. As a consequence, if the comparison is done

between two original characters an ordinary *PCA* results, whereas for all other comparisons the results will be different: in particular, the trace of the matrix will be larger than 2.

3. Among the pairs of representative variables, the pair is selected whose *PCA* second eigenvalue is minimum. Due to the unpredictable value of the trace of the covariance matrix this is not the same as searching for the highest value of the first one.

4. The two groups of characters corresponding to the selected pair of representative variables are merged in a group, that becomes a new node of the hierarchy.

5. The first principal component of the *PCA* corresponding to this node, i.e. the set of coordinates of units on the first principal axis, is chosen as its *representative variable.*

6. The first eigenvalue equals the variance of the representative variable, thus summarizes the amount of the original characters' variance that it summarizes.

7. The second principal component coefficients measure the distance between each variable in the node and the representative variable. We may call it *node variable*, in the sense that it shows the divergence of the two groups that are gathered in the considered node.

8. The second eigenvalue of this *PCA* is chosen as the hierarchy index of this node.

The steps (2)...(6) are repeated $p-1$ times, obtaining a complete hierarchical classification of the characters.

The idea underlying the method is that the representative variable of a group plays the role of central tendency of the whole set, similar to the first principal component in a *PCA* and the centroid of a group of units. Thus, by non-standardizing the *PCA*, the weight of the groups is in some way given to its representative variable. This is like attributing to the centroid the weight of each group of units in the Ward's (1963) clustering method.

In a scatter diagram of the two representative variables, whose groups joined in a given node, the interpretation of both factors at each step is straightforward: the position of the first factor is within the smallest angle between the two straight lines spanned by the representative variables, since the highest scalar product corresponds to the smallest angle. That is, the first factor summarizes what the representative variables have in common and therefore what have in common also all characters gathered in the node. Instead, the second factor represents what the characters do not have in common, which is minimized at each step. It is then natural to consider the first factor as the variable representative of the new node. Since the *PCA* is not standardized, the first eigenvalue is the difference between the sum of the two variances of the two representative variables minus the second eigenvalue, so that it can take any positive value.

Once the process ends a series of coefficients $w_1, \ldots, w_p$ results that can be used to build all the representative variables. Let $Z_j = (Z_{ij}), i = 1, \ldots, n$ the column of Z corresponding to the j-th character. For the k-th node the representative variable is $y_{1,k} = \sum_{j \in k} \left(\frac{w_k}{\sqrt{m_k}} z_j \right)$ with $m_k = \sum w_k^2$, where the sums are extended to the characters that compose the k-th node. The first consequence is that all representative variables are linear combinations of the original ones belonging to the node so that they lie on the same space spanned by the original characters. This allows the projection of the representative variables on the principal spaces of an ordinary *PCA*. Furthermore, for the characters $j = 1, \ldots, p$, let $t_j = \frac{z_j}{m_j}$, and for the representative variables for the nodes $k = p+1, \ldots, 2p-1$ let $t_k = \sum_{j \in k} \frac{m_j}{m_k} t_j$. Then for each node $k = (k1, k2)$ joining the two nodes $k1$ and $k2$, it can be shown that the second eigenvalue takes the form $\nu_k = \frac{m_{k1} m_{k2}}{m_{k1} + m_{k2}} \|t_{k1} - t_{k2}\|^2$. This formula is akin to a within-group variance as in the Ward (1963) clustering criterion, so that we can attribute to the sequence the same properties, in particular that n_k is non-decreasing along the clustering process. Unfortunately the m_js are unknown at the beginning of the process (that would tremendously speed up the computation) but only at the end.

The new feature of this clustering method, an advantage in comparison to the others, consists in the principal planes associated to the nodes of the hierarchy, where both characters and units can be represented, as in a common *PCA*. Indeed, in this case, only the characters belonging to the newly formed group are represented.

All representative variables are linear combinations of the original ones (with zero coefficients of the characters not belonging to the represented group), so that they lay in the same vector space. In particular, all second eigenvectors are orthogonal to each other and to the first eigenvector of the last node. As a consequence, the total inertia of the data table is decomposed according to the sequence of fusion levels plus the first eigenvalue of the last *PCA*. Indeed, the first principal component of the last *PCA* summarizes the similarity of all characters, whereas the dissimilarity is decomposed orthogonally into the $p-1$ second principal components of the previous *PCA*s. These properties would not hold if at each step the representative variables were standardized.

On a principal plane a group of characters may assume the form of a dipole, since the sign of the covariance between the concerned characters has no influence on the principal components, thus on the aggregation process. Therefore, the characters of a node may form a dipole of two groups opposed to each other in the direction of the representative variable of the node, according to the sign of their pair-wise correlation. This is not a drawback of the method, but rather a correct idea of aggregation, since the sign of the correlation depends on how each character is measured and not on its relation with the others.

The method has been recently improved, in order to fit an optimization criterion. Indeed, Denimal (2007) proposed to decompose the hierarchical factorial analysis in two main stages: the first one aims at building an initial hierarchical clustering of variables which is, in a second step, improved through an optimization process. The latter can be interpreted as a k-means type procedure (MacQueen, 1967) defining a convergent series of hierarchies and aims at improving the quality of both clusters and their reprÃ©sentative variables. In this optimization process, the level of the node defining each cluster is also taken into account by allocating increasing weights to the nodes of the hierarchy according to their levels. From this point of view, the optimization process aims at defining a new hierarchy whose significant splittings appear as clearly as possible and are concentrated in a number of upper nodes as small as possible. As a consequence, the two subclusters defined by each of these upper nodes of the optimized hierarchy are as separated as possible and the elements of each of them built up as closely as possible. As a result of this process, the found partition is optimized, together with the upper part of the hierarchy.

Table 2. *The first ten eigenvalues of the PCA of the time-series of surface temperature at the buoys array.*

N	Eigenvalue	Percentage	Cumulate
1	21,4221	31,50	31,50
2	15,1891	22,34	53,84
3	5,8129	08,55	62,39
4	3,6281	05,34	67,72
5	1,7463	02,57	70,29
6	1,5327	02,25	72,55
7	1,2760	01,88	74,42
8	0,9970	01,47	75,89
9	0,9380	01,38	77,27
10	0,7945	01,17	78,44

In order to have a partition of the characters into classes, the composed hierarchy may be cut as usual and one may take as reference the several methods proposed in literature (see Milligan and Cooper, 1985, for a review). Indeed, since the second eigenvalue would represent a share of the original characters' variance, the choice to cut at a level less than one would ensure that no group would be bi-dimensional, that is formed by uncorrelated characters: a choice that we share with Sarle (1983). Indeed, in this work we decided on the basis of a cross-validation, in practice an *a posteriori* check of the goodness of reconstruction of the original characters by the hierarchy's part upper the partition. For this task, the set of the *node variables*, sorted starting from the top node, is taken into account. For each original character, a forward multiple regression is performed, starting with the representative variable of the first node and including in sequence the node variables upper to the class to which the character belongs. As each regression procedure stops when no explanatory character improves significantly the regression, among the many possible stopping rules, we decided here to partition the hiererchy at a level such that all its upper nodes' node variable played a significant role for at least one character.

4. Numerical results

The data table of 68 buoys 6575 daily measures of surface temperature was submitted first to *PCA* adding the months, the years, and the combination of season and years as supplemental characters: 12-, 18-, and 72-nominal level characters respectively. In this way a better understanding of the pattern of the data could be obtained, as it will be shown later.

The examination of the sequence of the first ten eigenvalues of *PCA* reported in Table 2 shows that after the first two of high relevance, two minor follow and then other five all worth around one, whose value suggests that some attention might be deserved to their corresponding eigenvectors. For the moment, we concentrate on the first two, that summarize nearly 54 % of the total inertia, but later the following two might be taken into account, as they may to be tied to some groups or months or of sites. In this case, the explained inertia would raise to 68 %.

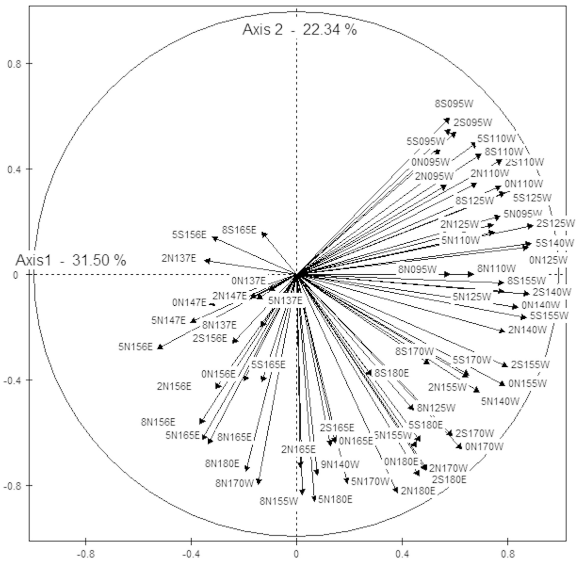


Figure 2. *PCA: Circle of correlations of the temperatures' time-series on the plane spanned by the principal axes 1 and 2.*

Indeed, looking at Figure 2, in which the sites time-series are set according to their correlation with the first two factors of *PCA*, it may be seen that most sites series are rather well represented, with a continuous pattern, that indicates a chain of correlation among them. The interpretation is straightforward, since the arrows point in direction of the warmer days, so that the fourth quadrant represents the warmer periods and the second the cooler ones. Apart from that first remark, the reading of the figure is neither easy nor interpretable, but in the following we shall take advantage of the classification of characters to reduce the amount of displayed items and attempt an interpretation concerning the sites.

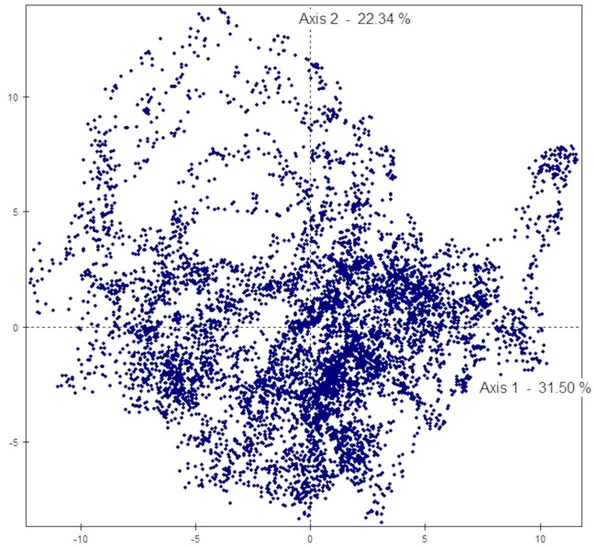


Figure 3. *PCA: The pattern of the daily measures of all buoys on the plane spanned by the principal axes 1 and 2 of the PCA of the time series.*

On Figure 3 the pattern of daily measures may be observed. It is evident that nothing can be said easily concerning the daily variation, due to the too large number of units involved. To understand this pattern, we shall use the projection on the plane of the months, the years, and the combination season-year, that will allow us to draw the trajectories corresponding to the time sequence of these items.

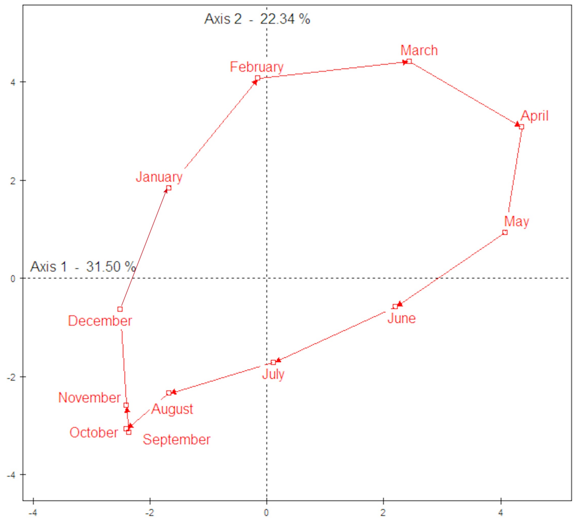


Figure 4. *PCA: The pattern of the months on the plane spanned by the principal axes 1 and 2 of the PCA of the time series.*

In Figure 4 the pattern of the months is represented. Looking at it the meaning of the factor plane becomes evident, since the yearly seasonal variation is regularly represented. Indeed, a regular circular pattern of the month results, with the year’s first season period roughly corresponding to the second quadrant and partially to the first, the second season to the first and partially the second quadrant, with the other two seasons concentrated in the fourth. This circularity suggests to understand the position of the individual series in the sense that the sites should reach their maxima of temperature in the years’ period whose position is in agreement with their direction and the minima in the opposite one. Despite the array is situated regularly around the Equator belt, no true opposition appears among the buoys situated in the opposite hemisphere. Rather, one may state that no correlation should exist between the series situated around the two factors. Indeed, a small group of poorly represented series appears near the negative site of the first axis: their position on the following axes deserves being explored, together with the seasonal pattern, to check if some specific behaviour of these sites may be detected.

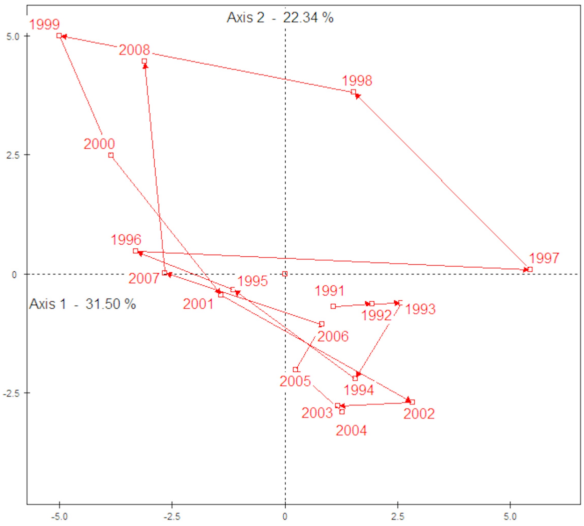


Figure 5. *PCA: The pattern of the years on the plane spanned by the principal axes 1 and 2 of the PCA of the time series.*

In Figure 5 the trajectory of the years is reported on the same plane. In this case, a rather limited fluctuation is visible in the years 1991-1994, that are concentrated in the fourth quadrant. Then in 1995-1996 an evident displacement in direction of the second quadrant is followed by a dramatic shift along the first axis, so that 1997 (when the last registered El Niño event occurred) is set in a position corresponding to maximum heath. Then, the following two years result progressively more cold, to return two years later to the previous average situation, so that the years 2002-2006 are again in the fourth quadrant. Indeed, the last two years show a pattern reversed in respect to the El Niño previous one, with a shift towards the maximum cold.

In Figure 6 the time pattern is given more complex by combining the year and the season. Thus is may be seen that the El Niño maximum that occurs in winter 1997 is preceded by a change in the circular fluctuation that may be noticed in winter 1996, in which the tendence is inverted towards colder temperatures than usual, so that the increase until winter 1997 is horizontal along the first axis. After winter 1997 the regular pattern is reestablished, but at colder levels, until winter 2000, after which a stability period reappears with the same temperature levels as before 1996, until a new pattern change, this time in direction of the cold, reappears starting in autumn 2007.

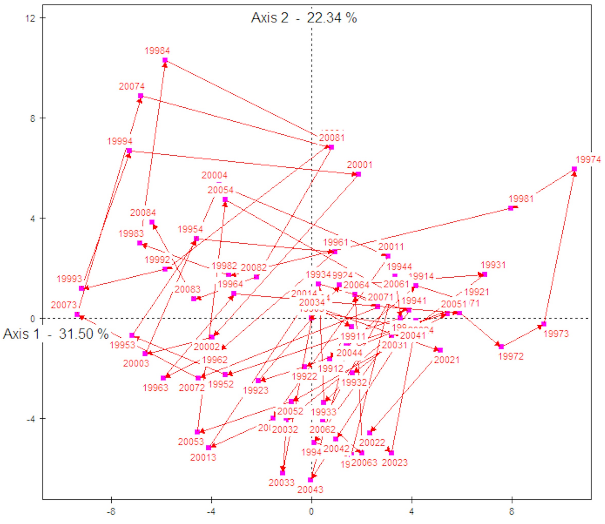
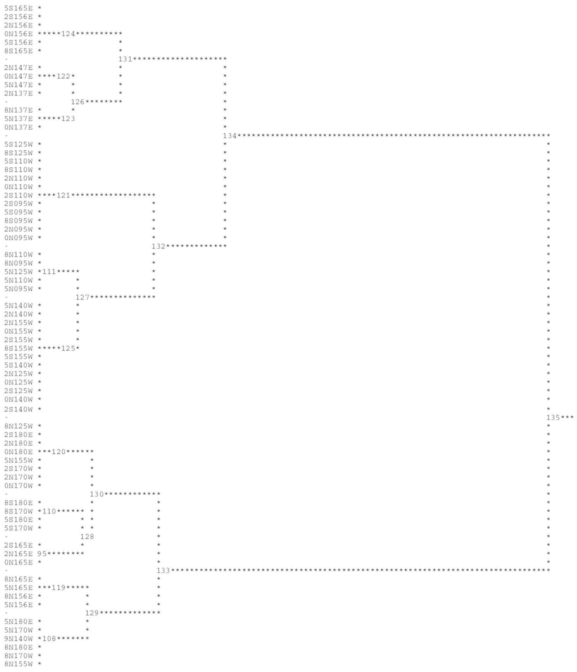


Figure 6. *PCA: The seasons pattern from 1991 to 2008 on the plane spanned by the principal axes 1 and 2 of the PCA of the time series.*

On this basis, we are now able to understand the meaning of the "finger" that appears on the representation of the daily measures: it should be the period of maximum effects of El Niño.

We study now the results of *HFC* applied to our data table. Indeed, due to the large number of units and of missing values, we could not take advantage of the offered possibility, to represent the units on factor spaces and to partition them according to their relative position. Nevertheless, we could obtain the classes and the representative variables, that we further projected as supplemental on the first factor plane of *PCA*.

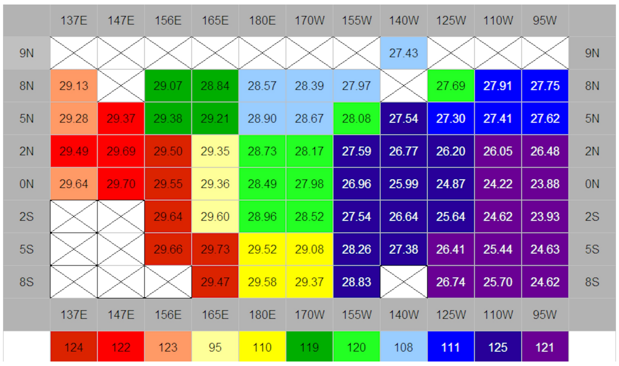


```
5S165E *
2S156E *
2N156E *
0N156E *****124************
5S156E *                  *
8S165E *                  *
-                         131***********************
2N147E *                  *                        *
0N147E ****122*           *                        *
5N147E *      *           *                        *
2N137E *      *           *                        *
-             126**********                        *
8N137E *      *                                    *
5N137E *****123                                    *
0N137E *                                           *
-                                                  134*************************************************************************
5S125W *                                           *                                                                          *
8S125W *                                           *                                                                          *
5S110W *                                           *                                                                          *
8S110W *                                           *                                                                          *
2N110W *                                           *                                                                          *
0N110W *                                           *                                                                          *
2S110W ****121*********************                *                                                                          *
2S095W *                          *                *                                                                          *
5S095W *                          *                *                                                                          *
8S095W *                          *                *                                                                          *
2N095W *                          *                *                                                                          *
0N095W *                          *                *                                                                          *
-                                 132***************                                                                          *
8N110W *                          *                                                                                           *
8N095W *                          *                                                                                           *
5N125W *111*****                  *                                                                                           *
5N110W *       *                  *                                                                                           *
5N095W *       *                  *                                                                                           *
-              127*****************                                                                                           *
5N140W *       *                                                                                                              *
2N140W *       *                                                                                                              *
2N155W *       *                                                                                                              *
0N155W *       *                                                                                                              *
2S155W *       *                                                                                                              *
8S155W *****125*                                                                                                              *
5S155W *                                                                                                                      *
5S140W *                                                                                                                      *
2N125W *                                                                                                                      *
0N125W *                                                                                                                      *
2S125W *                                                                                                                      *
0N140W *                                                                                                                      *
2S140W *                                                                                                                      *
-                                                                                                                             135***
8N125W *                                                                                                                      *
2S180E *                                                                                                                      *
2N180E *                                                                                                                      *
0N180E ***120*******                                                                                                          *
5N155W *           *                                                                                                          *
2S170W *           *                                                                                                          *
2N170W *           *                                                                                                          *
0N170W *           *                                                                                                          *
-                  130**************                                                                                          *
8S180E *           *               *                                                                                          *
8S170W *110******  *               *                                                                                          *
5S180E *        *  *               *                                                                                          *
5S170W *        *  *               *                                                                                          *
-               128                *                                                                                          *
2S165E *        *                  *                                                                                          *
2N165E 95********                  *                                                                                          *
0N165E *                           *                                                                                          *
-                                  133*****************************************************************************************
8N165E *                           *
5N165E ***119******                *
8N156E *          *                *
5N156E *          *                *
-                 129***************
5N180E *          *
5N170W *          *
9N140W *108********
8N180E *
8N170W *
8N155W *
```

Figure 7. *HFC: The dendrogram representing the upper part of the hierarchy, with the 10 nodes dominating the 11 classes partition.*

The inspection of the hierarchy structure obtained after optimization gives eleven groups as the partition of higher interest. The upper part of the dendrogram is shown on Figure 7: here, it may be seen that the eleven groups gather to form an important partition in three classes, with two of them relatively more similar than the other.

The partition is represented in Figure 8, in which the time-series are represented in a schematic way according to their geographical position. Here the different colours represent the different groups, as identified in the bottom row by the number of the corresponding node of the hierarchy. On each series the mean temperature along the whole period is reported. It is interesting to observe that, apart from the two classes 120 and 123, all others result spacially connected, an important sign of continuity, that may concern the two said classes too, considering that one is on the border of the array and within the other a missing buoy results, so that even these may be somehow connected in the reality.



	137E	147E	156E	165E	180E	170W	155W	140W	125W	110W	95W	
9N	☒	☒	☒	☒	☒	☒	☒	27.43	☒	☒	☒	9N
8N	29.13	☒	29.07	28.84	28.57	28.39	27.97	☒	27.69	27.91	27.75	8N
5N	29.28	29.37	29.38	29.21	28.90	28.67	28.08	27.54	27.30	27.41	27.62	5N
2N	29.49	29.69	29.50	29.35	28.73	28.17	27.59	26.77	26.20	26.05	26.48	2N
0N	29.64	29.70	29.55	29.36	28.49	27.98	26.96	25.99	24.87	24.22	23.88	0N
2S	☒	☒	29.64	29.60	28.96	28.52	27.54	26.64	25.64	24.62	23.93	2S
5S	☒	☒	29.66	29.73	29.52	29.08	28.26	27.38	26.41	25.44	24.63	5S
8S	☒	☒	☒	29.47	29.58	29.37	28.83	☒	26.74	25.70	24.62	8S
	137E	147E	156E	165E	180E	170W	155W	140W	125W	110W	95W	
	124	122	123	95	110	119	120	108	111	125	121	

Figure 8. *HFC: The composition of the classes on a schematic reproduction of the buoys geographical position. The number of class is given in the coloured legenda below. In each cell, the average temperature of the corresponding time-series is reported.*

Looking at the table, one may notice that in the Eastern site of the Pacific a larger homogeneity results, as it may be reflected by the size of the two classes 121 and 125 that summarize 25 out of 68 time-series, more than one third of the total. With class 111 they identify the coldest site of the Pacific Ocean and constitute a group of a higher partition into three classes of the dendrogram. The other classes are much smaller: the classes on the Easter site 122, 123, and 124 are situated North and North-East of Indonesia and are the warmest. Their behaviour seems alternative to the others, so that the big class they form seems tied to the previous one mainly in the dipole sense, that is based on their negative correlation. The other classes are situated in the Central Pacific in an intermediate position, with temperatures in-between the others, thus reflecting an independence from them. This could explain the fact that they gather with the other dipole at the very last level. On the opposite, it is not easy to understand the differences among the series in the Central and in the Western Pacific, since this may depend on variations that may not be so easily visible.

5. Final remarks

At a very first sight, the El Niño fluctuation results very well depicted by the first few graphics that we showed here and it appears as a very important variation in the otherwise stable fluctuation of the temperature's regime in Pacific Ocean's Equatorial belt. From the graphics it results that the corresponding raise in temperature is preceeded, around a year earlier, by a decrease. This appears in the only complete El Niño fluctuation that was registered by the collected data, namely the 1996-2000 one, but maybe a new El Niño cycle started in autumn 2007, this time apparently with an important lowering until end 2008, the end of the downloaded data.

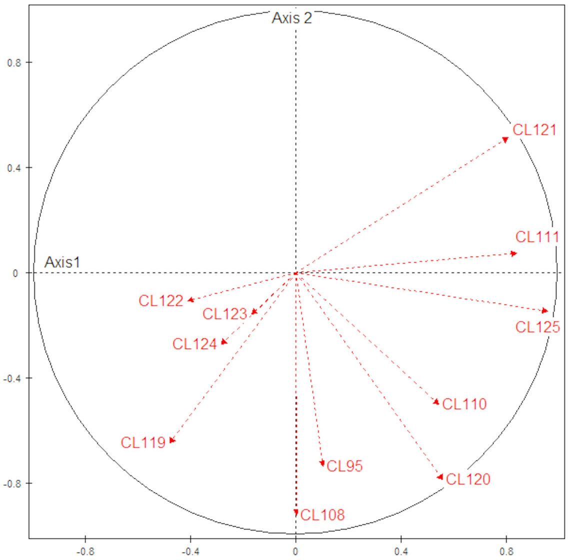


Figure 9. *The variables representative of the 11 classes obtained from the HFC projected as supplemental on the circle of correlations on the factor plane spanned by the first two factors of PCA.*

Unlike *PCA,* in the first results of *HFC* we are not able to detect the El Niño fluctuation, since we could not take full advantage of the units representation. Nevertheless, the results gave us interesting information concerning the relative homogeneity of the classes and the relative difference among them. Indeed, the higher fragmentation of the Central and Western Pacific, in respect to the Eastern side, maybe could be interpreted either on a morphological basis, such as the larger presence of islands, or on the presence of different streams. From the analyses other interesting results derive, in particular the different behaviour detected in the different groups of buoys along time, that results from their different position on the first factor plane, as well as the special situation of two groups of buoys, whose main variations are along the third and fourth factors respectively.

From the unified representation given by projecting the representative variables on the factor planes (shown in Figure 9) we may observe the three classes of time-series of the Eastern Pacific on the right side of the first axis, the five of the Central Pacific around the lower part of the second axis, and the third group of three situated in the Western Pacific, poorly represented on this factor plane, in the third quadrant. Reminding the Figure 4 we may say that the maximum temperature occurs for the first class in se second fourth of the year, for the second in the third and for the third in the fourth, so that the first years period appears as the coldes all the monitored area round. The El Niño anomaly, that was increasing progressively until the end of 1997, appears then very well, with an indication that it interests in particular the most Eastern site of the Ocean, close to the American coasts.

Indeed, a deeper comprehension of the Pacific Ocean temperature pattern could derive by studying the correlation among series at some time-interval lag. Maybe this could explain the difference among the small classes obtained.

It must be observed the contrast between the regular pattern of the seasonal variation during the normal years and the important deviation due to El Niño effects in the studied period: with more historical data one may evaluate the different deviation during the various manifestation of the fluctuation and try to derive any sistematic pattern. This could be the subject of a further investigation.

Acknowledgment

This work is being developed in the framework of the bilateral agreement, of which the first and third authors are the promoters, between the Dipartimento di Matematica of the Università di Roma La Sapienza and the Facultad de Ciencias / Instituto de Matemáticas y Ciencias Afines of the Universidad Nacional de Ingenieria of Lima. The paper was written during their reciprocal visits, granted by both institutions. The support of the Istituto Italiano di Cultura is also gratefully acknowledged, as it contributed to the first author's stage in Lima in 2010. Also the *TAO* Project Office of *NOOA/PMEL* is gratefully acknowledged for providing the time-series that were used in this study.

1. J.P. Benzécri et al. *L'Analyse des données*, Tome 1, 2nd Ed., Paris, Dunod. (1982).

2. S. Camiz, J.J. Denimal and V.D.P Pillar. Hierarchical Factor Classification of Variables in Ecology, *Community Ecology* **7(2)** (2006) 165–179.

3. S. Camiz, and V.D.P Pillar. Comparison of Single and Complete Linkage Clustering with the Hierarchical Factor Classification of Variables, *Community Ecology* **8(1)** (2007) 25–30.

4. J.J. Denimal. Hierarchical Factorial Analysis, *Actes du 10th International Symposium on Applied Stochastic Models and Data Analysis*, (2001).

5. J.J. Denimal. Classification Hiérarchique Optimisée d'un tableau de mesures. *Revue de Statistique Appliquée*

148(2) (2007) 29-61.

6. C. Eckart and G. Young. Approximation of one matrix by another of lower rank, *Psychometrika* **1** (1936) 211–218.
7. M.J. Greenacre. *Theory and Application of Correspondence Analysis*, London, Academic Press, (1984).
8. D.A. Jackson D.A.. Stopping Rules in Principal Components Analysis: A Comparison of Validation of stopping rules in eigendecomposition methods Heuristical and Statistical Approaches, *Ecology*, **74(8)** (1993) 2204-2214.
9. I.T. Jolliffe. *Principal Components Analysis.* Springer, Berlin (1986).
10. C. Lagrand and L.M. Pinzón. *Análisis De Datos. Métodos y ejemplos.* Escuela Colombiana de Ingenieria Julio Garavito, Bogotà (2009).
11. P. Legendre and L. Legendre. *Numerical Ecology.* 2nd Ed., Elsevier, Amsterdam (1998).
12. J.B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, **1** (1967) 281-297
13. G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50(2)** (1985) 159-179.
14. P.R. Peres-Neto, D.A. Jackson, and K.M. Somers. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, **49** (2005) 974-997.
15. W.S. Sarle. *Cubic clustering criterion.* Technical Report A-108. Cary, N.C.: SAS Institute (1983).