# Principal Component Analysis of pluviometric data
# b) Application to the missing value problem

*English translation of the paper* "Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias" *Estadística (Journal of the Inter-American Statistical Institute) 46, 146-147, pp. 55-83.*

Carlos López, Juan F. González and Rosario Curbelo[1]

## Abstract

The missing value problem is well known in all studies related either to natural phenomena or other areas. The present work has been motivated by the need to fill in the gaps in a daily pluviometric data bank in order to use it with an hydrological model. The spatial mean over the sub catchment area is calculated with the Thiessen's method, which on principle do not require a complete fill in. However, the method is highly sensitive to outliers in the case of few available measurements. The outlier detection problem has been analyzed in a companion paper, and here we will concentrate in reporting results for the missing value problem using some different methodologies. Such methodologies should preserve the main characteristics of the population, as well as the present quality and accuracy levels.
Results for four methods tested using a 15 year, daily pluviometric measurements are presented. The methods were the standard nearest neighbor, linear interpolation of the station time series, linear interpolation of the time series of the Principal Component Scores and Penalty of the Principal Component Scores. The last one were developed for this problem and proved to show the best behavior.

## 1. Introduction

According with Haagenson, 1982, Johnson, 1982, etc. objective analysis of both hydrological and meteorological fields are common practice. They are designed to interpolate an observed quantity using only sparse data. For the spatial mean rain field there exist other methods, like the Thiessen one (see Jácome Sarmento *et al*., 1990 for example) which might produce the required result without the need to imputate all missing values. Both situations led to a comparatively low interest in the research community, which was related with the scarce literature found in the specialized journals reviewed.
In the author's opinion in the overwhelming majority of the practical applications, the missing value is simply ignored, under the implicit assumption that those missing records appear at random, hypothesis which is rarely tested.
On the other hand, the topic is of major interest in statistics and social sciences; working group reports are mentioned in specific books, like the one of Rubin, 1987.
Of course somewhat sophisticated imputation methods do exist. For example, the one used at the US National Bureau of Census is quoted by Rubin, 1987. The idea is to imputate the missing value using a randomly selected one taken from the events which have an identical response in all the other answers (the method is originally designed for surveys). If there is

[1] Centro de Cálculo - Facultad de Ingeniería - Montevideo - Uruguay

no other event in such condition, a distance between surveys is defined, and the nearest is chosen for imputation.
Another typical and simple method is to make a regression over the dataset, fitting a mathematical model. Usually partial or total least squares as well as principal components are used, as presented by Stone *et al*., 1990.
All the abovementioned methods produce a single value for a each missing value. Quoting Rubin, 1987 "... in general is intuitive that imputating using the *optimal* value for each missing value will underestimate variability...". However, the possibility to obtain more than a single value for each missing one can be considered. Rubin, 1987 described a set of techniques (some too much specialized to surveys). The general idea is create, for each missing value, m possible alternative values (with m small) and considering that m different complete sets are available. If the missing value rate is low, the method might be of use, requiring however more space (for saving the multiple imputations) and also more computation time (for processing separately the different sets). For further details please see Rubin, 1987.

# 2. The present work

## *2.1 Motivation*

This work can be considered as a natural extension of the preliminary treatment of the pluviometric data used in the calibration phase of a flow-rain-flow hydrological model for the Río Negro catchment area. Three hydropower dams operate sequentially there, operated by the national electrical utility (UTE[2] ). For further details please refer to Silveira *et al*. (1991, 1992a y 1992b).

## *2.2 General characteristics of the study area*

a) Geographical
Even though we have analyzed a greater catchment area, we restrict ourselves for this paper to the Río Tacuarembó catchment area, of about 20.000 km$^2$, located at 32° S 55° W, at 400 km from the ocean. The typical landscape is smooth, with heights below 500 m, few canyons and lakes. The typical monthly rain rate is within 74 and 120 mm/month.

b) Measuring net
The national net defined by the DNM[3] is based on a regular grid of size 10 km, sequentially numbered. The identification for the station is the same of the cell which it belongs; from time to time two or more stations might be operating within the same cell, so a letter A, B etc. is appended to the identification code. We coined the term *synonym* for those stations. For our purposes, we will consider all the stations belonging to the same cell as the same station. From the administrative point of view, our net is the sum of four ones, independently operated by different institutions. Those nets have different spatial density and reliability. Its structure has been changing during time, and any station might:

- Start operating in any moment during the period

[2] UTE - Administración de Usinas y Transmisiones Eléctricas
[3] DNM - Dirección Nacional de Meteorología

- Stop operating in any moment during the period
- Start operating to substitute another one which has been withdrawn
- Be replaced by other or not.

For our purposes we will only distinguish those stations that use to be operated by AFE[4], which systematically adds the readings from Sunday and Monday and consider them as belonging to Monday. In the abovementioned catchment area 21 stations are in operation, and we selected 13 for this work.

c) Dataset

As mentioned before, the topology of the net usually suffer from transformations. According to Silveira et al., 1991 it exists at present too many stations. Since many of them have synonyms we join their records and disregard any distinction within a cell.
For this paper we selected a subset of 13 stations, located as shown in fig. 1, which have been carefully checked for typing errors by using some algorithms presented in López *et al*., 1994. We restrict ourselves to records from Jan 1st. 1975 to Dec 2nd. 1989, covering nearly 15 years.

# 3. Methods used for the Missing value problem

## *3.1 Nearest neighbor*

The method assigns a list of alternative stations to the one which is intended to imputate; the missing value will be replaced by a number taken from the first one with measurements for the particular date. On principle the abovementioned list is ordered according to increasing distance to the one intended to imputate; however some confidence considerations are taken into account and the purely geometric order might be altered.

## *3.2 Time series interpolation*

If the value for the day $t_f$ and station j is missing, we search for the nearest previous and following reading available at station j, and a simple linear interpolation is performed.
Let us denote as $t_f$ the date of the missing record and as $p_j(t_f)$ the unknown value for the day $t_f$ and station j.

Let us denote also as $t_{f-m}$ the latest day before $t_f$ with readings, and as $t_{f+r}$ the first day after $t_f$ with readings $\left(t_{f-m} < t_f < t_{f+r}\right)$. The interpolation rule for the missing value is

$$p_j(t_f) = p_j(t_{f-m}) + \frac{t_f - t_{f-m}}{t_{f+r} - t_{f-m}}\left(p_j(t_{f+r}) - p_j(t_{f-m})\right) \qquad (1)$$

[4] AFE - Administración de los Ferrocarriles del Estado

### *3.3 Time interpolation of the Principal Component Scores series (TIPS)*

This method is based upon Principal Component Analysis (PCA), which have been presented in the companion paper by López *et al*., 1994. We will present briefly the notation and refer the reader to the abovementioned reference.

Let us name as $\mathbf{P}_{(n,1)}(t)$ the precipitation vector of the n selected stations for the time t. Let's define a rectangular matrix $\mathbf{M}$ which rows are the vectors $\mathbf{P}(t_{m_j}) - \mathbf{P}_M, j = 1..r$, defined for those days without missing values. $\mathbf{P}_M$ is the mean vector for the considered period.

The eigenvectors of matrix $\mathbf{C}_{(n,n)} = \mathbf{M}^T * \mathbf{M}$ are named principal components or patterns, and will be denoted as $\mathbf{e}_i$. We will assume that the associated eigenvalues are ordered, and decrease with i. The relationship between the pluviometric records $\mathbf{P}_{(n,1)}(t)$ and the scores represented as the vector $\mathbf{A}_{(n,1)}(t)$ is

$$\mathbf{P}(t) = \mathbf{P}_M + \mathbf{E}.\mathbf{A}(t) \tag{2}$$

where $\mathbf{P}_M$ stands for the mean vector for the period and $\mathbf{E}_{(n,n)}$ is the matrix formed by the eigenvectors $\mathbf{e}_i$.

$$\mathbf{P}(t) = \begin{bmatrix} \mathrm{p}_1(t) \\ . \\ . \\ . \\ . \\ \mathrm{p}_n(t) \end{bmatrix} ; \mathbf{P}_M = \begin{bmatrix} \overline{\mathrm{p}_1} \\ . \\ . \\ . \\ . \\ \overline{\mathrm{p}_n} \end{bmatrix} ; \mathbf{A}(t) = \begin{bmatrix} \mathrm{a}_1(t) \\ . \\ . \\ . \\ . \\ \mathrm{a}_n(t) \end{bmatrix} . ; \mathbf{E} = \begin{bmatrix} \mathbf{e}_1 \, \mathbf{e}_2 \, .. \mathbf{e}_n \\ \\ \\ \end{bmatrix}$$

If $\mathbf{C}_{(n,n)}$ is non-singular then $\mathbf{E}_{(n,n)}$ is invertible, so given the readings $\mathbf{P}(t_{f-m})$ y $\mathbf{P}(t_{f+r})$ it is possible to obtain vectors $\mathbf{A}(t_{f-m})$ and $\mathbf{A}(t_{f+r})$ from (2).

Eq. (2) can be expressed as well as

$$\mathbf{P}(t) = \mathbf{P}_M + \sum_{\mathrm{i}=1}^{\mathrm{i}=\mathrm{n}} \mathbf{a}_\mathrm{i}(t).\mathbf{e}_i \tag{3}$$

For any intermediate time $t_l, l \in (f-m+l, f+r-1)$ the precipitation is calculated by linear interpolation of vector $\mathbf{A}(t)$ . On principle, all the readings for time t can be obtained from eq. (2).

By analyzing the scores $a_i$ it is clear that the standard deviation of $a_i$ decreases as i increases making typically a minor contribution to the summation.

It is then a natural conclusion that for the reconstruction of vector $\mathbf{P}(t)$ all terms with i>q can be neglected for some q, without substantial loss of information. Then the equation (3) can be substituted by the following approximate expression

$$\mathbf{P}(t) = \mathbf{P}_M + \sum_{i=1}^{i=q} \mathbf{a}_i(t).\mathbf{e}_i \qquad (4)$$

Summing up, if there is at least one missing value of vector $\mathbf{P}(t_f)$ corresponding to time $t_f$ we search for the nearest previous and following without missing values. It should be stressed that, in opposition with the standard linear interpolation, this method uses all n stations, and not each one independently.

Let denote as $t_f$ the day to be imputated. Let also be $t_{f-m}$ the nearest previous day without missing values and $t_{f+r}$ the nearest following $(t_{f-m} < t_f < t_{f+r})$. Since both $t_{f-m}$ and $t_{f+r}$ are complete days the scores $\mathbf{A}(t_{f-m})$ and $\mathbf{A}(t_{f+r})$ corresponding with vectors $\mathbf{P}(t_{f-m})$ y $\mathbf{P}(t_{f+r})$ can be easily calculated using eq. (2).

Then, for time $t_{f-m+l}$, the vector of scores $\mathbf{A}(t_{f-m+l})$ can be calculated by linear interpolation of both vectors $\mathbf{A}$ mentioned before. The first guess for the precipitation of time $t_{f-m+l}$ can be obtained from eq. (4).

However at time $t_{f-m+l}$ there are some values recorded. The missing ones can be taken from the first guess vector, and then we can complete all elements vector $\mathbf{P}(t_f)$ using as much available information as possible.

Once completed the time $t_f$ we are in position for a new interpolation, using now vectors $\mathbf{P}(t_{f-m+l})$ and $\mathbf{P}(t_{f+r})$ as starting point; this step can be repeated as many times as necessary, in order to fill all the gaps.

The performance of this approximation is heavily connected with the autocorrelation properties of the scores $a_i$. For meteorological variables this autocorrelation properties are very different for different $a_i$, which is another argument to limit the number of terms in eq. (4).

For example, in the work of Cisa *et al*. (1990) it is shown that for the hourly surface wind in southern Uruguay the time lag $T_i$ required for the autocorrelation of the i-th score time series to take for the first time the value 0.5 is (25,9,5,3,3,...,1.2,1) for i=1...15, being the bigger values for the most important PC. $T_i$ is measured in hours.

Such situation is not the case for daily rain, because all scores have a dramatic drop for $T_i$=l day, being 1 day the sampling period (see figs. 2 and 3). This fact explains in part the poor results obtained with this method, despite it is better than the one obtained with the standard linear interpolation of station time series.

### *3.4 Penalty of the Principal Component Scores*

If we analyze the histogram of the scores $a_i$ it can be observed that for the main PC it is heavy skewed (asymmetric?), or has a significant dispersion around zero. On the other hand, for the weak PC the histogram is symmetric and the dispersion around zero is very low. As an example see figs. 4 and 5. They have been obtained from a slightly modified population, because all events with null precipitation *in all the stations* have been removed for this plot. Any imputation procedure should preserve this properties, and then it should produce scores $a_i$ consistent with this histograms, i.e. very near zero for all weak PC. Such property might be imposed as a condition, choosing for any given date all missing components of vector $\mathbf{P}(t_f)$ in order to minimize some penalty function, like

$$S(\mathbf{P}) = \sum_{i=k}^{i=n} w_i . a_i^2(\mathbf{P}) \tag{5}$$

being the scores $a_i(\mathbf{P})$ corresponding to the vector $\mathbf{P}$ (now complete) and the weights $w_i$ selected in order to consider the different absolute value of each score $a_i$. Vector $\mathbf{P}$ is only partially known, and it is assumed that it has q unknowns (or missing values). The optimum of S can be obtained making its partial derivatives null for all unknowns

$$\frac{\partial S}{\partial p_{m(j)}} = 0, \ \mathrm{j} = 1..\mathrm{q}$$

being $p_{m(j)}$ the missing records for this time. The so defined linear system can be easily solved by standard procedures.

## 4. Experimental procedure

We generate at random pairs of time-station which will be regarded as fictitious missing values. Valid pairs are those which have been measured; all methods should calculate a value to *imputate* it. We will denote as *real value* the measured one, and *calculated value* as the one obtained by means of an imputation procedure. If the pair is not valid, it is simply discarded.

Once a prescribed number of valid pairs have been processed the standard deviation of the difference between real and calculated values for each method is calculated; it will be the main statistics to compare within methods.

The total number of pairs is taken as a percentage of the days in the analyzed period (5450 days from Jan 1st. 1975 to Dec. 2nd. 1989) and we restrict the number of missing values per day to one. We varied the percentage from 20% through 80% and no significant difference in the results were noticed.

For all methods we made runs considering all pairs; for the nearest neighbor and Penalty we made also runs disregarding those events with zero rain in all stations, which account for 80% of the cases. We did it as an attempt to avoid the negative impact of such significant amount of constants in the estimators. The results in terms of standard deviation increased two times approximately, but the relative values for different methods remained the same.
We depurated as much as possible the data bank against the original paper records. All 13 stations show less than 5% missing values for the period, which we considered enough for the purposes of this work.
We also made some sensitivity checks for the parameters of the different methods: for both methods using scores, we varied the number of terms to consider; for the nearest neighbor method we varied the list of alternatives (by removing the nearest ones).

# 5. Results

All figures correspond with an experiment with 2091 days with missing values, implying 53% of the analyzed days approximately.

## *5.1 Nearest neighbor method*

Some calculations using only the list of 13 stations were performed; also we enlarged the set by using up to 86 stations located in the catchment area as well as in the vicinity. The precedence order in this case was strictly geometric distance.
We confirmed that the availability of a dense network of stations in a regular topography like this improves the results. If we delete alternatives from the list increasing the mean distance between the station to be imputated and its alternatives is clear (see fig. 6) that the standard deviation increases.
The regularity characteristics of the rain phenomena in this region even up to distances of 150 km apart leads to good results. This mean distance is defined here as the expected value of the geometric distance weighted by the frequency of substitutions by an alternative. The maximum distance between any of the 13 stations is 201 km.
The point of fig. 6 with mean distance 33.7 km results from choosing alternatives from the set of 13 stations. The resulting standard deviation is $\sigma$ = 5.55 mm/day which has to be taken into account when comparing the different procedures.
In general the alternative stations have not been corrected at all; one consequence is that even with a lower mean distance, the alternative set renders a bigger standard deviation than the abovementioned case of the 13 stations.

## *5.2 Linear interpolation of the station time series*

As mentioned before, good results can be expected for this method if the considered phenomenon evolves slowly in relation with the sampling interval, or if any contiguous gap is shorter than the typical time scale of the problem.
The rain in Uruguay shows significant variation in time, and generally the meteorological event occurs under the form of storms more or less concentrated in time (from hours up to three or four days). That's why using daily sampling this method is unlikely to produce good results.
The results for the set of 13 stations show a standard deviation of 12 mm/day.

### 5.3 Linear interpolation of the time series of the Principal Component Scores

As expected this method renders similar results as the standard interpolation of the time series. Despite it requires more CPU time, the results are not significantly better: the standard deviation of the error varies from 11.3 and 11.83 mm/day depending on the number of terms using in the calculation.

The weak effect in using more or less terms in the calculations is not surprising, and it is related with the extremely low time autocorrelation of the phenomena. In fig. 7 the evolution of the standard deviation respect to the index q (see eq. 4) is shown.

It is expected that this method will produce significantly better results for other meteorological variables - like surface temperature, surface wind, etc. - but this extreme have not been tested yet.

### 5.4 Penalty of the Principal Component Scores

This method renders the best values in terms of the standard deviation, with a minimum value of 4 mm/day. To give an indication of the variability of the time series itself we calculated its standard deviation. For any station its value is over 15 mm/day for the whole period.

The value of 4 mm/day has been obtained by using k in the range 8 to 10 (see eq. 5) which implies 3 to 5 terms in the equation. For example, if k is equal 1 all the scores are taken into account, so the surface is forced to resemble the historical mean value. For k near n, only the weakest PC are affected in the summation S, but some others which also explain noise are left uncontrolled. This fact explains why the results are of poor quality.

The task of fixing the optimum k for each data bank can be accomplished by means of an experiment like this, or by a subjective analysis of the PC. The shape of the isolines might easily distinguish from those which are related with noise from those which are related with the physics. The weakest PC are also very sensitive to outliers (see Silveira *et al*. 1991).

From fig. 8 it can be analyzed the dependence of the standard deviation of the error in relation to the number of terms analyzed. It is clear that the optimum is robust in relation with k.

The weights $w_i$ were chosen in order that all terms $w_i a_i^2$ are of comparable order. In early stages we adopted as $w_i$ the inverse of the RMS. of the time series of the scores $a_i$. Despite being reasonable this rule revealed soon unsuitable for the main patterns because they are clearly asymmetric. Thus, we defined the weights as $w_i = 1/\alpha_i^2$, being the limit $\alpha_i$ defined in order to

$$\int_{-\alpha_i}^{\alpha_i} a_i^2 . f_i(a) . da > 0.96$$

being $f_i$ the probability distribution function for the score $a_i$, i=1..n. It is automatically verified that for less than 4% of the events, $w_i . a_i^2(t) \geq 1.0$

## 6. Conclusions and recommendations

From the results presented it is clear that all methods which rely on the temporal behavior of the rain might distort significantly the general characteristics of the population, in particular if the proportion of missing value is important.

Both the Nearest neighbor and Penalty of principal component scores methods which take into account the spatial behavior of the rain render significantly better results than the ones based on temporal properties. The main reasons are the characteristics of the rain in terms of the time sampling strategy and the smooth topography of a small catchment area.

The method named Penalty of Principal Component scores shows a standard deviation 28% less than the Nearest neighbor one (4.01 vs. 5.55 mm/day), a value which has been obtained using only measurements of 13 stations.

If one includes more stations, the standard deviation is even bigger, even for lower mean distances, which can be partly explained by the unknown quality of the extra stations.

Another advantage for this proposed method is the modest computer resources involved. For instance, even a hand held computer might be enough, since the only requirements is that it can hold a matrix of size n, a vector of size n for the mean averages, and the capabilities to solve a linear system of equations. This feature should be considered for routine operation of any hydrologic model.

As a further improvement of the procedure, we plan to minimize the joint probability of the scores $a_i$ which in turn implies the solution of a non lineal problem for each event with missing values. Despite a (significantly) increased computer time cost this procedure is theoretically more sound.

## 7. Acknowledgments

Juan González implemented and performed the calculations for both the nearest neighbor and temporal interpolation of the station series, and Rosario Curbelo programmed the linear interpolation of the time series of the Principal Component Scores. Carlos López developed and coded the Penalty of the Principal Component Scores program. The methodological and experimental design was due to Carlos López. All three authors took part in the discussion of the results. It should be mentioned that this work is an extension of the task accomplished under the contract "Development of an hydrological model for the Río Negro catchment area" funded by UTE. The permission for using the information and publish the results is gratefully acknowledge.

# 8. References

González, E.; Morales, C., 1991. "Depuración de la base de datos pluviométricos de la cuenca del Río Tacuarembó". Internal report prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos e Ingeniería Ambiental. 11 pp. (in spanish)

Haagenson, P.L, 1982. "Review and evaluation of methods for objective analysis of meteorological variables" Papers in Meteorological Research, V 5, N 2, 113-133.

Jácome Sarmento, F.; Sávio, E.; Martins, P.R., 1990. "Cálculo dos coeficientes de Thiessen em microcomputador". In Memorias del XIV Congreso Latinoamericano de Hidráulica, Montevideo, Uruguay (6-10 Nov., 1990). V 2, 715-724.

Johnson, G.T. 1982. "Climatological Interpolation Functions for Mesoscale Wind Fields". Journal of Applied Meteorology, V 21, N 8, 1130-1136.

Lebart, L.; Morineau, A.; Tabard, N. 1977. "Techniques de la Description Statistique: Méthodes et logiciels pour l'analyse des grands tableaux". Ed. Dunod, París. 344 pp.

López, C.; González, E.; Goyret, J., 1994. "Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos" Estadística, V 6, N 146-147, 55-83.

Richman, M.B., 1986. "Review article: Rotation of principal components" Journal of Climatology, V 6, 293-335.

Rubin, D. B., 1987. "Multiple imputation for nonresponse in surveys". John Wiley and Sons, 253 pp.

Silveira, L.; López, C.; Genta, J.L.; Curbelo, R.; Anido, C.; Goyret, J.; de los Santos, J.; González, J.; Cabral, A.; Cajelli, A., Curcio, A., 1991. "Modelo matemático hidrológico de la cuenca del Río Negro" Final report. Part 2, Chapter 4. 83 pp. (in spanish)

Silveira, L.; Genta, J.L.; Anido Labadie, C., 1992a. "HIDRO URFING - Modelo hidrológico para previsión de caudales en tiempo real- Parte I: Simulación de los procesos hidrológicos en el suelo". Internal report 1/92 prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos e Ingeniería Ambiental, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

Silveira, L.; Genta, J.L.; Anido Labadie, C., 1992b. "HIDRO URFING - Modelo hidrológico para previsión de caudales en tiempo real- Parte II: Transformación en cuenca, ruteo y criterios de calibración y verificación" Internal report 2/92 prepared for the Hydrology Department of the Instituto de Mecánica de los Fluidos, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

# Figures