A new technique for imputation of multivariate time series: application to an hourly wind dataset

CARLOS LÓPEZ† and ELÍAS KAPLAN‡
Centro de Cálculo, Faculty of Engineering (11)
CC 30, Montevideo, Uruguay
†internet: carlos@fing.edu.uy
‡internet: elias@fing.edu.uy

Abstract:The techniques employed in the treatment of an hourly surface wind database during the development and calibration phases of an objective wind field interpolator model are presented. The model itself has been applied to estimate the regional wind energy resource creating a layer in a GIS environment.
The outlier detection phase is presented in a companion paper, and here the different techniques applied in order to imputate the missing values are described. The comparative results obtained with an hourly dataset of 15 years long are also presented. Two different problems have been simulated numerically: systematic missing values (i.e. at fixed hours) and non systematic ones.
Five different criteria were applied: imputation with the historical mean value; linear time interpolation within single station records; optimum interpolation (kriging) and the two newly developed **P**enalty **Of** the **P**rincipal **S**cores and linear **T**ime **I**nterpolation of the **P**rincipal **S**cores which considers all station records in a multivariate fashion; they prove to be the most accurate for this particular wind dataset. There is also some evidence of oversampling in time.

1. Introduction

1.1 Presentation of the problem

Since 1988 the Team was involved in evaluating the National Wind Energy Resource. Although it was not an explicit objective of the project, it was necessary to complete the time series to the longest available period. With these goals some algorithms have been implemented for imputation of missing values in the series. Some test have been carried out that confirm the kindness of their performance in controlled cases.

Three different methods have been tested:

a) Time interpolation of the principal scores (TIPS) (includes standard time series interpolation as special case)
b) Penalty of the principal scores (POPS) and
c) Optimum interpolation (Gandin, 1965).

The first two were developed in López *et al.* (1994b). The third is a standard interpolation procedure in the inicialization of mathematical models in meteorology (Johnson, 1982) that allows to find estimates not only in the stations of measuring, but also in another points of the domain of work. In this case their performance has been analyzed in the case in that the point to interpolate is one of the measuring stations. In the work of López *et al.* (1994b) on rain data, the TIPC resulted to be a poor method, while for the case of wind it was the best one.

1.2 Methodological background

Objective analysis methods are very common in meteorology (see Haagenson, 1982, Johnson, 1982, etc.) since they have been designed to produce an interpolated field using only data observed in irregularly distributed networks. This situation gives, on principle, a way to overcome the problem of missing values, because they can be estimated using available information.

For some applications the missing values are not a problem (for example, *find the extreme values, calculate an annual average*, etc.) while for others they are critical. The existence of well established procedures led in the past to a marginal interest for the problem, clearly observed in the scarcity of specific work found in the literature. We believe that, in most cases, missing values are simply ignored, under the implicit hypothesis that those errors appear at random. Such hypothesis is rarely tested nor verified.

Missing values are extremely important in statistics and social sciences, where even books on the topic can be found (Rubin, 1987) mentioned results from international working groups. In such areas there exist both crude and sophisticated imputation methods. For example, the one suggested by the U.S. Census Bureau (Rubin, 1987) assigns a randomly chosen value among those events which other values coincide, or are below a so defined "distance" from the target one.

Another simple method is to make a linear regression using available data. Usually such model is build up least squares criteria, principal component analysis, etc. (Stone et al. 1990) All of the above produce for each missing value, a single candidate. Following Rubin, 1987 "..it is intuitively clear that by imposing an "optimum" value, the variability will be underestimated". The author suggests that more than one candidate might be produced, and he described the techniques typically used for surveys. The general idea is to create, for each missing value, a small number m of candidates, and consider that you have m datasets. The method is workable if there are a low number of missing values; its results are usefull, but require more computing time and also more space (to store the multiple imputed values). We refer to Rubin (1987) for further details.

2 Testing Methodologies

Five stations were selected for the test, all located to the south of the country, (see fig. 1) :Melo (440), Paso de los Toros (460), Treinta y Tres (500), Carrasco (580), Punta del Este (595). They were chosen due to its geographical localization around the automatic stations of the National Department of Energy (DNE) Wind Energy Program. This also conditions the periods to work with, including part of the years 1990-1991 and the year 1984.

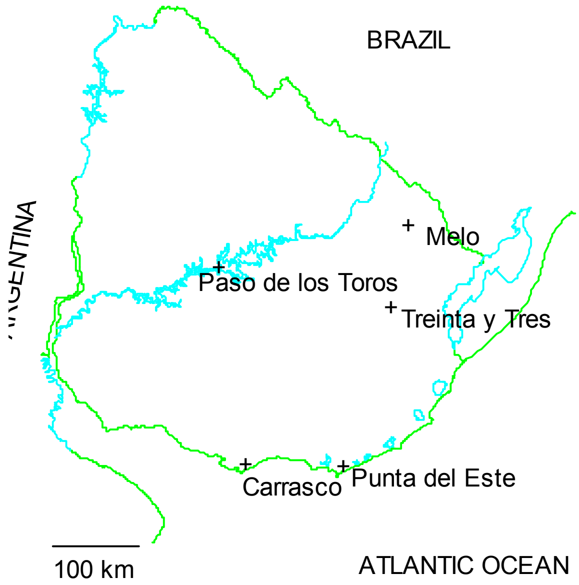

Figure 1 Location of the weather stations

The work carried out consisted in:

a) removing temporarily those values to be imputed

b) for each method

b.1) eliminating all missing values

b.2) calculating RMS and mean between new and true values

The methods were presented in full in López *et al.* (1994a and 1994b) applied to the case of pluviometric data (daily values). A synthesis will be introduced pointing out the differences among rain vs. wind here managed. It will be referred as event the set of data values for a particular hour. As indicated in López *et al.* (1994b), the population could be divided in a) hours with data in all the considerate stations and b) hours with some absence in them. The quality control process requires either measured or estimated data for the considerate event, in all the stations, so an imputation for the missing values is required.

The Principal Components Analysis (PCA) was already used for the same wind dataset in Cisa *et al.* (1990). While considering events without missing values, it is straightforward to find the principal components of the population. They are a property of the group of events, and not of any in particular. Any complete event has associated 2n measurements (components or and v in each one of the n considerate stations), and it can be sought as a vector in the space R^{2n}. The principal components are a base of that space and the principal scores are the projection of the space in that base (Lebart *et al.* 1977).

Since both bases relates each other by means of a linear relationship, it is equivalent to manage the temporal series of the original measurements in the stations or the temporal series of the principal coefficents.

The time series of the principal coefficents are mutually uncorrelated, an important difference with the original values. Such fact enables us to consider each component separately, knowing that there is no redundant information in the others. In figures 2 to 4 some of the distributions of the observed coefficents are introduced as well as its power spectrum.

It should be noticed that except for the first three components ("the most important") the observed distribution is relatively concentrated around the zero, as it was pointed out in Cisa *et al.* 1990. From the analysis of the figures 2 to 3 is deduced that the series of the corresponding scores 1 to 3 varies smoothly, in opposition to the registered in the subsequent figures. The spectrum shows a noisy pattern, and the selfcorrelation decreases more sharply with the lag. It can be that, in the case of existing any missing values, it would be reasonable to perform a linear interpolation in time for the scores of minor index. The other scores are typically of minor or greatly minor importance (compare the dispersion of the figure 4 with the one of the figure 2) and they can be neglected. So after interpolation of the main scores and setting to zero the remaining ones the complete set of scores can be obtained for the event with missing values.

By means of the linear mentioned transformation the tentative registrations are calculated, but only those that were lacking are incorporated. A more precise estimate of the interpolated scores is now possible upon incorporating the values indeed measured corresponding to the event. For more detail, the reader refers to López *et al.* 1994a, 1994b. This procedure of temporal interpolation will be named Time interpolation of the principal scores (TIPS) hereinafter. As a particular case, the standard linear interpolation between registers of the same stations is an special case, when all the principal scores are included in the interpolation.

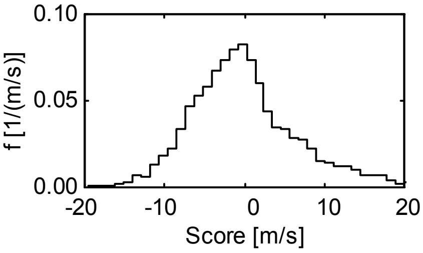

Figure 2 Sketch of some properties of the scores 1(left) and 2(right). On top the probability density function; on the middle the power spectrum and on bottom the self correlation against the lag.

The method of Penalty of the principal scores (POPS) was also used, as introduced in López *et al.* 1994b. It is based upon the fact that the weakest patterns have associated usually very small scores; so the functional (suggested for the first time by Hawkins, 1974)

$$\mathrm{S}(u_1, v_1, u_2, v_2, \ldots, u_n, v_n) = \sum_{\mathrm{i} \in \mathrm{p}} \frac{\mathrm{a}_\mathrm{i}^2}{\mathrm{w}_\mathrm{i}} \qquad \mathrm{i} = k..2n$$

will be also small. a_i is the i-th score, and w_i is a weighting score. The key idea is that any missing value u_j or v_j can be estimated directly by minimizing the S functional.

$$\frac{\partial \mathrm{S}}{\partial u_\mathrm{j}} = \frac{\partial S}{\partial v_\mathrm{j}} = 0 \qquad \mathrm{j} \in r$$

where $\mathrm{u}_{\mathrm{j},}$ v_j denotes the missing component u and v in the j-th station and being r the set of stations with missing values for the event. The system of equations is linear in $\mathrm{u}_{\mathrm{j},}$ v_j and of moderate size. It may happend that the optimum values still are unacceptable, by application of the criteria documented in López *et al.* 1994a, but such constraint has not been imposed in our experiments. The particular case k=1 corresponds to minimize the Mahalanobis distance to the mean value.

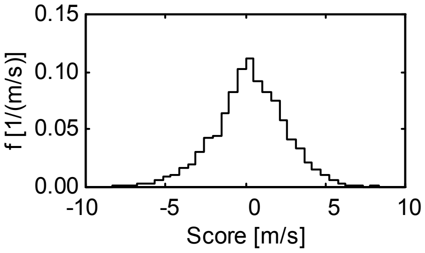

Figure 3 Sketch of some properties of the scores 3 (left) and 4 (right) as presented in fig. 2. Notice the change in scale for the power spectrum. The peak near 0.04 is due to daily variations.

As a final point, it should be stressed that this procedure neglects the information of the temporal self correlation of the coefficents. For details, again the reader refers to López *et al.* 1994b.

3 Results

Two experiments were carried out, depending on the systematic nature or not of the missing values. The dataset corresponds to the years 1990-91 after removing errors

3.1 Systematic location of missing values

Typically in the northern zone of the country most stations take readings at 8, 14 and 20 hours, as some stations from Southern Brazil and the Argentina do. Only the stations of Artigas, Rivera, Salto and Paso de los Toros in Northern Uruguay take hourly wind values.

It has been evaluated the possibility of imputate such systematic holes. In order to have a frame with what compare, we have analyzed again the five stations of the southern zone: four of them were considered with data only at 8, 14 and 20 hours local time, while Melo was taken as fully hourly. The three above mentioned methods were applied to this case, being the results presented in Table 1. As a reference, it has been evaluated the error when every component x_j was imputed simply with its mean value for the period, and by application of the Optimum interpolation, which renders in this case for each hour, the same value for every station.

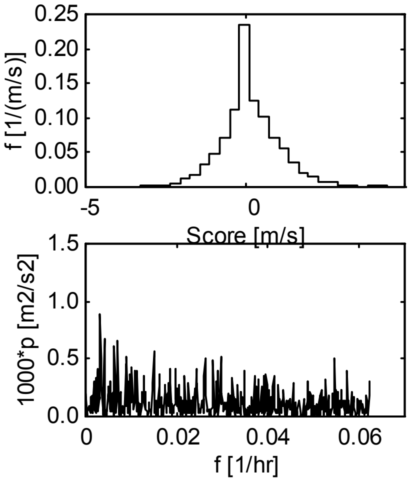

Figure 4 Sketch of some properties of the scores 9 (left) and 10 (right) as presented in fig. 2. Notice again the change in scale for the power spectrum, and also in the x-scale for the pdf.

The outputs are consistent in that the TIPC performs the better, without a clear difference between interpolating the 4 first scores or all of them.

The experiment simulates a population of 60924 absences in 4 of the 5 considerate stations, leaving only the hours 8, 14 20 like well-known data, being the total usable population in the calculation of the eigenvectors of 83724 values (corresponding to the hours with mensurements in all the stations). The critera adopted imputated the missing values with a mean error of 0.10 m/s and a RMS of 2.05 m/s, calculated in relation with the original data (Table 1). The procedures were also evaluated in that the absences are at random. The results also support as a suitable choice interpolating all of the 10 coefficents, and will be presented in the next section

Table 1 Calculation of the root mean square (RMS) and the mean of the error (measured value - calculated value) upon imputation assuming complete Melo and the another 4 stations with registrations in the hours 8, 14 20 only. Data from the year 1990-91. In boldface the most significant outputs.

Interpolation			Penalization		
Interpolated terms	RMS (m/s)	Mean (m/s)	Penalized terms	RMS (m/s)	Mean (m/s)
1:10	2.06	0.09625	10:10	3.41	0.10094
1:9	2.06	0.09669	9:10	3.41	0.10274
1:8	2.05	0.09613	8:10	3.39	0.10452
1:7	2.06	0.09671	7:10	3.28	0.06608
1:6	***2.05***	***0.08151***	6:10	3.26	0.06191
1:5	2.06	0.09585	5:10	3.23	0.04485
1:4	2.05	0.09541	4:10	3.21	0.01852
1:3	2.05	0.08414	3:10	3.40	0.01686
1:2	2.11	0.08763	2:10	2.97	0.00177
1:1	2.73	0.07331	***1:10***	***2.84***	***0.05171***

Results obtained assigning the mean value	3.24	0.28839
Results obtained with the Gandin´s method	2.84	0.05353

3.2 Non systematic location of missing values

15197 absences in the population of 83728 were created at random dates and stations, being approximately 20% of the total. In this case the selected criterion of interpolate with all the terms (terms 1:10) renders a mean error of 0.03 m/s and RMS of 1.67 m/s between the calculated data and the original values (Table 2).

The results from Table 2 are valid for a *single* random set. Other independent runs revealed that in some cases a minimum in k = 7 could be noticed, more in accordance with the previous results for rain. However, the corresponding optimum RMS error was not very different from the value for k = 10. The objective (goal) function was the RMS of the population of differences between the calculated value and the one indeed measured. The optimum value is approximately 2 m/s, which is acceptable for wind speed. It should be noticed that interpolating with all 10 terms is equivalent to an independent interpolation of each station's time series.

4 Conclusions

From the performed experiments it could be inferred that:

a) The algorithms employed in order to imputate values performed very satisfactorily. They outperform the standard procedures.

b) The near optimum value obtained using the standard interpolation of the time series suggests that, at least in Uruguay, the wind records are oversampled in time. When an artificial undersample is introduced the optimum number of interpolated terms decreases. We provided a physical explanation of the phenomena based upon the spectrum of the score's time serie.

Table 2 Calculation of the root mean square (RMS) and the mean of the error (measured value - calculated value) upon imputation assuming 20% of the data with missing values. Data from the year 1990-91. In boldface the most significant outputs.

Interpolation			Penalization		
Interpolated terms	RMS (m/s)	Mean (m/s)	Penalized terms	RMS (m/s)	Mean (m/s)
1:10	**1.67**	***0.03193***	10:10	6.83	0.19072
1:9	1.67	0.03092	9:10	7.78	0.09796
1:8	1.68	0.03401	8:10	7.99	0.13297
1:7	1.68	0.03286	7:10	8.27	0.14827
1:6	1.70	0.04416	6:10	7.02	0.02573
1:5	1.73	0.04740	5:10	5.34	0.03002
1:4	1.76	0.02857	4:10	3.33	0.10016
1:3	1.79	0.03594	3:10	2.73	0.06495
1:2	1.89	0.03005	2:10	2.35	0.04825
1:1	2.57	0.00417	***1:10***	**2.33**	***0.07813***

Results obtained assigning the mean value	2.76	0.03141
Results obtained with the Gandin´s method	2.37	0.07596

5 References

Cisa, A.; Guarga, R.; Briozzo, C.; López, C.; Alonso, J; Cataldo, J.; Canetti, R.; Acosta, A.; Penza, E.; Xavier, V.;Tozzo, A.; Estrada, J.; Bevc, A.;

Maggiolo, G.; Chaer, R.; Rosenblatt, R.; Lamas, R.; Martínez, F. y Cabrera, R., 1990. "Proyecto de Evaluación del Potencial Eólico Nacional: Informe Final" Facultad de Ingeniería, Instituto de Mecánica de los Fluídos e Ingeniería Ambiental e Instituto de Ingeniería Eléctrica, Montevideo, Uruguay. 1000 pp. (in spanish)

Gandin, L. M., 1965. "Objective analysis of Meteorological Fields". Israel Program for Scientific Translations, 242 pp.

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.

Hawkins, D.M., 1974. "The detection of errors in multivariate data, using Principal Components" Journal of the American Statistical Association, V 69, 346, 340-344.

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

López, C.; González, E.; Goyret, J., 1994a. "Análisis por componentes principales de datos pluviométricos. a) Aplicación a la detección de datos anómalos" Estadística (Journal of the Inter-American Statistical Institute) 1994, 46, 146,-147, pp. 25-54.

López, C.; González, J. F.; Curbelo, R., 1994b. "Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias". Estadística (Journal of the Inter-American Statistical Institute) 1994, 46, 146,-147, pp. 55-83.

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 (in spanish). Part 2, Cap. 4. 83 pp.

Stone, M.; Brooks, R.J., 1990: "Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression" J. R. Statist. Soc. B, 52, N 2, pp 237-269.