Application of ANN to the prediction of missing daily precipitation records, and comparison against linear methodologies[1]

Carlos López-Vázquez
Facultad de Ingeniería, Centro de Cálculo
CC 30, Montevideo, Uruguay
e-mail: carlos@fing.edu.uy

Abstract

Depending upon the user, weather records can be used as they are, or they need to be imputated prior its use. Despite the fact that general methods for meteorological variables exist, they are difficult to apply for daily rain. A specially difficult feature is that the overwhelming majority of the records (>80%) are of zero rain, leading to a very non-gaussian distribution. Other characteristic is the low autocorrelation of the time series.
The test region was the Santa Lucia river catchment area of 13000 km^2, at 35°S near the Atlantic; its yearly accumulated precipitation values are around 1000 mm, without a clear dry or wet season. The selected subset has 20 years long and 10 stations; 30% of the events show missing values.
A Monte Carlo simulation was designed, randomly choosing both date and station for the missing values and afterwards different imputation procedures were successively applied. Some statistics which characterize the distribution of the absolute error, namely its expected value, variance and 75, 85 and 95 percentile have been derived in order to compare the results.
Both traditional linear meteorological interpolation procedures as well as a suite of Backpropagation Artificial Neural Networks(ANN) has been compared. The present results are not very good, and show that is possible to imputate with a mean error of 2 mm/day and a RMS of 7 mm/day using both linear and nonlinear procedures, while ANN seems to be more robust against outliers.

Introduction

The problem of interpolate a field using sparse data is typical in many areas. In meteorology the objective analysis method (Haagenson, 1982; Johnson, 1982) is commonly applied because of its simplicity. It provides indirectly a way for calculating missing values using available data. However, the significant amount of information required by this method usually restricts its use to Global Data Assimilation Centers (Gandin, 1988), because they require historical records for the calculations.
Ideally the availability of all records is preferred, but there are meteorological problems which do not require a full dataset. For example, to calculate the areal mean value of rain the Thiessen-Voronoi polygons method (Jácome *et al*., 1990) can be applied, without requiring extensive imputation of missing values.

[1] This work was funded by the Uruguayan CONICYT, under contract 51/94
Presented at the International Conference on Engineering Applications of Neural Networks. Stockholm, 16-18 June, 1997, pp. 337-340

Both situations led to a low interest on the topic reflected on the scarce meteorological papers on it. In most practical cases, either the missing value is ignored (assuming implicitly that the missing mechanism is random) or some ad-hoc technique is applied (linear interpolation, nearest neighbor, etc.) without further test or documentation. In any case, the population is clearly affected in an arbitrary way, under some hypothesis of unknown validity. However, it should be noticed that the missing value problem is of great interest in the Statistics and Social Sciences in general (Rubin, 1987).

Considered methods: a) linear

Due to its simplicity, this methods are widely used. They can be grouped together since the estimated quantity is a linear combination of the available data. Its general expression is $y_j = \underline{w}.\underline{x} + b$ being y_j the unknown quantity, $\underline{x}$ a vector with the available data and both the weight vector $\underline{w}$ and the number b are depending on the method. Typically vector $\underline{x}$ holds the values of the same day, and both $\underline{w}$ and b are constants for the whole dataset. This definition covers the methods of Cressman, Optimum interpolation (Gandin, 1965), Ordinary least squares, as well as other more simple ones, as the nearest neighbor. For the sake of completeness a brief description of them will follow:

- **Cressman**

The requested number is obtained after a linear combination with weights which are the inverse of square distance. The method does not require historical information, but only the station coordinates.

$$y_j = \frac{\sum_{i \in N} y_i / d_{ij}^2}{\sum_{k \in N} 1/d_{kj}^2} \text{ being vector } w_i = \frac{1/d_{ij}^2}{\sum_{k \in N} 1/d_{kj}^2}, \text{ and } b = 0$$

- **Optimum interpolation (Gandin, 1965; Johnson, 1992)**

This method is routinely applied for the initialization of global weather codes. Instead of interpolate the desired field, it interpolates the anomaly or difference with a simple predictor, and the spatial correlation properties of the anomaly field are analyzed. Usually it is assumed both isotropic and homogeneous, and it should be modelled in the general case. However, if the point where the prediction is required is one of the measuring point, its covariance with the other stations is available, and it looks very similar to the Ordinary Least Squares. The covariance might be calculated separately for winter and summer, for example, or used all together as we did. This procedure allows using information from the day before.

We used different anomaly fields and transformations for the variable to be interpolated which are summarized in table 1. For example, the method coded as "gandin7" assigns values for the variable $x_i = \sqrt{rain}$, taking the anomaly respect to its historical mean. In this case, $\underline{w}$ is fixed (following Johnson, 1992); $b = \bar{x}_j$ (the overbar stands for average over time). The classic Optimum Interpolation procedure is coded as "gandin20". Because daily rain has a very irregular probability density function (pdf) we designed a transformation $x = f(rain)$ which makes pdf(x) nearly uniform, except for $rain = 0$.

The transformation based on the cumulated density function is invertible and assures that x belong to the interval $[0,1]$.

Our coded name	*Anomaly respect to:*	*Variable to interpolate*	*Using data from days*	
			t	*t-dt*
gandin	*historical mean*	*rain*	*X*	-
gandintrans	*historical mean*	*f(rain)*	*X*	-
gandin6	*historical mean*	*rain*	*X*	*X*
gandin7	*historical mean*	$\sqrt{rain}$	*X*	-
Initial value for the field chosen as zero				
gandin_diario	*0*	*rain-daily mean*	*X*	*X*
gandin4	*0*	*rain*	*X*	*X*
gandin5	*0*	*rain*	*X*	-
Neglecting instrumental error				
gandin20	*historical mean*	*rain*	*X*	-
gandin3a	*historical mean*	*rain-daily mean*	*X*	-

Table 1 Brief information about the different methods based on climatological functions. f(rain) denotes the transformation which renders a nearly uniform probability density function(see text). t and t-dt denotes values from the day and the day before

- **Ordinary Least Squares**

This method is completely standard and its theory can be found elsewhere. The weights $\underline{w}$ are chosen in order to minimize the 2-norm of the vector $M^{(j)}\underline{w} - m^{(j)}$ (a scalar proportional to the RMS) being $M^{(j)}$ the matrix of the available data (as many rows as dates, as many columns as stations but without the j-th one) and $m^{(j)}$ is a column vector with the j-th stations values. The implemented version assumes that the data is error free, so $\underline{w}$ can be derived from (dropping the index j) $M^T M\underline{w} = M^T.m$. The T stands for transpose. The number b is 0.

- **Least average (Least 1-norm)**

Here the weights $\underline{w}$ are chosen in order to minimize the 1-norm (sum of absolute values) of the vector $M^{(j)}\underline{w} - m^{(j)}$. This is a much more difficult problem because it does not lead to a linear system of equations and has to be solved as a non-linear optimization task. Also it requires substantially more CPU time than all previous methods.

- **Least 95 percentile**

Since the population might be affected by a small set of gross errors, it is fit to minimize a robust statistic, as the 95 percentile of the distribution of errors. As before, this problem requires significant CPU time.

- **Nearest Neighbor**

We considered two criteria for the distance: euclidean and qualitative similarity. In both cases the missing value is taken directly from another station following a given order. In the

first case, the order is due to geometrical distance, and in the second we used the expertise from a meteorologist. All weights are zero, except one which is 1, and the number b is 0.

- **Assign a constant value**

This is a simple method, which disregard any other information. We applied it using the modal value and the expected value. For our dataset, the former is 0 mm/day and the latter is near to 3 mm/day.

Considered methods: b) Non linear methods (ANN)

Such methods are very new and they are based upon simple models of the biological neural networks. They have been used for the short term prediction of SO_2 concentration (Boznar *et al.,* 1993), electrical load (Park*.,* 1991), etc. The ANN is organized in layers, being the first one stimulated directly by the observed values; each neuron of the next layer is stimulated by a linear combination of the outputs of the previous layers by means of a simple transfer function, like the logsig (Demuth *et al.,* 1994) given by:

$out_j = \left\{1 + \exp\left[-\sum_i \left(a_{ij} * input_i\right)\right]\right\}^{-1}$, with parameters a_{ij} to be fixed for each neuron. The ANN requires, as its biological counterpart, a training process which is simulated here by means of adjusting the a_i parameters. In this work we compared one two-layer net, with 6 logsig neurons in the hidden and 1 linear neuron in the output layer, and one three-layer one, with 8 linear neurons, 4 logsig and one logsig for the output. Both were trained using one third of the available values trying to minimize the RMS of the error. The error is defined as the difference between ANN output and true value. For the first case we substracted rain values in mm/day while for the second case something different has to be done, since the last neuron has an output belonging to the interval $[0,1]$. We trained the net in order to minimize the error with the transformed rain $x = f(rain)$. All nets were trained using *backpropagation* (Rumelhart *et al*., 1986) and due to practical reasons the number of iterations was kept low, so its performance might be improved with more iterations. Its training cost in CPU time is high: over 10 hours of SUN 20 for each meteorological station.

Conclusions and results

After 250 simulations, the results are summarized in table 2. It should be noticed the improved results for those methods using information from the day before (gandin4, gandin6 and gandin_diario). Among those which use only information of a single day, the best results are obtaining by the minimum 95 percentile, closely followed by the Ordinary Least Squares method.

It should be stressed that, since the database still has errors, it is possible that the methods suggest suitable values and the outliers affect some of the considered statistics. This is unlikely to occur for the 85, 95, etc. percentile, and then the importance of the ANN denoted bp7.

As final conclusions:

a) common methods based upon mere substitution by a neighbor or by a constant gave poor results.

b) as expected, optimum interpolation methods outperforms the others in terms of RMS, fairly close to the ordinary least squares and least 95 percentile.
c) non linear methods (very expensive in the training phase) led to slightly more robust results, but renders similar figures in terms of average and RMS.

	Average	75%	85%	95%	RMS
bp1	2.65	1.92	4.53	**13.03**	**7.15**
bp7	2.51	1.28	3.64	**12.54**	7.71
cressman	2.63	**0.80**	4.58	15.75	8.20
gandin	2.64	1.48	4.20	13.57	7.24
gandin3a	2.60	1.83	4.73	13.96	7.42
gandin20	2.68	1.56	4.21	13.42	7.21
gandin4	2.53	1.92	4.59	13.28	**7.02**
gandin5	2.39	1.25	4.14	13.64	7.25
gandin6	2.71	2.06	4.72	13.39	**7.05**
gandin7	**2.23**	**0.50**	**3.11**	13.39	7.48
gandin_diario	**2.04**	0.89	**2.99**	**11.01**	7.66
gandintrans	3.06	**0.80**	4.52	13.46	8.11
least squares	**2.34**	1.33	**4.09**	13.13	**7.01**
least 95's percentile	**2.34**	1.34	**4.10**	**13.07**	**7.01**
least average	**2.26**	0.86	3.60	**13.23**	7.21
modal value	2.79	**0.00**	**1.78**	19.04	10.26
expected value	4.73	2.96	**3.02**	16.25	9.88
geometrical distance	2.76	**0.02**	4.22	17.37	9.13
expert distance	2.82	**0.01**	4.31	17.74	9.33

Table 2 Preliminary results in mm/day for the different imputation methods. The expected value and the 75, 85 and 95 percentile of the distribution of the absolute error, and its RMS are presented and compared. In bold the five best results for each estimator.

References

Boznar, M.; Lesjak, M. and Mlakar, P., 1993 "A neural Network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain" Atmos. Environ., V 27B, N 2, 221-230

Demuth, H. and Beale, M. 1994 "Neural Network User's guide (Toolbox for MATLAB)" The MathWorks, Inc. 226 pp., http://www.mathworks.com

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

Gandin, L. M., 1988 "Complex Quality Control of Meteorological Observations". Mon. Wea. Rev, V 116, 1137-1156

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. and 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 (in portuguese)

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

Park, D. C., 1991 "Electric load forecasting using an artificial neural network" IEEE Transactions on Power Systems, N 2, 442-449

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

Rumelhart, D. E.; Hinton, G. E. and Williams, R. J. 1986 "Learning representations by Back-Propagating errors", Nature, V 323, 533-536