Comparación de metodologías para la imputación de la lluvia diaria en una pequeña cuenca de latitudes medias

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

Abstract:

This work attempts to compare different methodologies for the missing value problem of daily rain data. A Monte Carlo simulation was designed, randomly choosing both date and place for the missing values and afterwards different imputation procedures were successively applied. We build some statistics which characterize the distribution of the absolute error, namely its expected value, variance and 75, 85 and 95 percentile to compare the results.
Among others, we tested the inverse of distance Cressman's method, Optimum Interpolation, Ordinary Least Squares and Artificial Neural Networks as a nonlinear example. The test region was the Santa Lucía river catchment area of 13000 km^2, at 35°S near the Atlantic; its yearly accumulated precipitation values are around 1000 mm. The dataset has 20 years long and ten stations.
The present results show that it is possible to imputate with a mean error of 2 mm/day and an RMS of 7 mm/day using both linear and nonlinear procedures, while there is still room for improvement in the latter.

Introducción:

En meteorología son práctica corriente métodos de análisis objetivo (ver Haagenson, 1982, Johnson, 1982, etc.), que permiten generar un campo interpolado a partir de datos irregularmente distribuídos. Indirectamente ello proporciona métodos para calcular los valores faltantes a partir de los existentes. Tradicionalmente el volumen de información requerido ha restringido estos métodos a los grandes centros de asimilación de datos (Gandin, 1988), ya que se requiere un banco histórico para el cálculo.
Por otra parte no siempre es necesario que el banco de datos esté completo. Por ejemplo, para el cálculo de lluvia media sobre una región, existen métodos como el de los polígonos de Thiessen (Jácome Sarmento *et al.*, 1990) que no requieren en principio, de la imputación exhaustiva de las ausencias.
Ambas situaciones han llevado a que el tema del tratamiento o eliminación de ausencias haya sido también relegado, lo que se refleja en lo escaso de los trabajos específicos en la literatura especializada consultada.
En la mayoría de los casos prácticos, el dato ausente es simplemente ignorado, (bajo la hipótesis implícita que estas ausencias son al azar) o se aplican técnicas ad-hoc (interpolación lineal, sustitución por el más próximo) que luego no son documentadas en el trabajo final. En cualquier caso, se afecta a la población en forma arbitraria, en base a hipótesis que rara vez son evaluadas.
El tema en cambio es de gran interés en el área de la Estadística y las Ciencias Sociales en general (Rubin, 1987).

Metodologías analizadas:

a) métodos lineales

Por su simplicidad estos métodos son los más corrientes. Se caracterizan por presentar la estimación como una combinación lineal de los datos disponibles. Su forma general es:

$$y_j = \underline{w}.\underline{x} + b \qquad (1)$$

donde y_j es el valor incógnita, $\underline{x}$ son los datos disponibles y tanto los pesos $\underline{w}$ como el número b son propios del método. Típicamente se usan los datos disponibles del día, y tanto $\underline{w}$ como b son constantes.

En esta definición se incluyen genéricamente los métodos de Cressman, Interpolación Optima (Gandin, 1965), mínimos cuadrados ordinarios, etc., así como otros aún más simples, como la sustitución por el vecino más próximo. Se dará a continuación una breve descripción de los mismos:

- **Cressman**

El valor a interpolar es una combinación lineal con pesos que varían en forma inversamente proporcional al cuadrado de la distancia. No se requiere de información histórica sino sólo geométrica.

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

La suma se efectúa sobre el conjunto N de datos disponibles.

- **Interpolación Óptima (Gandin, 1965; Johnson, 1992)**

Este es el método usualmente aplicado en la inicialización de los modelos globales. En lugar de tratar de interpolar el campo deseado, se interpola la anomalía respecto a algún predictor simple, y se estudian las propiedades de correlación en el espacio del mismo. Se asume para ello que la anomalía es isotrópica y homogénea.

Cuando se quiere interpolar en puntos de medida, no es necesario modelar la correlación en el espacio ya que se puede utilizar su valor muestral, lo que hace muy parecido el método al de mínimos cuadrados ordinarios. Esta correlación muestral puede calcularse para períodos disjuntos de tiempo (por ejemplo, invierno y verano) o asumirse válida para todo el período, que fue lo que se hizo en este caso. También es posible considerar información disponible tanto del día como del día precedente.

En este trabajo se utilizaron diferentes definiciones del campo a interpolar y de la anomalía recogidas en la tabla 1

	Anomalía respecto a:	*Variable a interpolar*	*Con datos del día*	
			t	*t-dt*
gandin	*media histórica*	*lluvia*	*X*	-
gandintrans	*media histórica*	*f(lluvia)*	*X*	-
gandin6	*media histórica*	*lluvia*	*X*	*X*
gandin7	*media histórica*	$\sqrt{lluvia}$	*X*	-
Inicializando el campo con el valor cero				
gandin_diario	*0*	*lluvia-media diaria*	*X*	*X*
gandin4	*0*	*lluvia*	*X*	*X*
gandin5	*0*	*lluvia*	*X*	-
Despreciando el error instrumental				
gandin20	*media histórica*	*lluvia*	*X*	-
gandin3a	*media histórica*	*lluvia-media diaria*	*X*	-

Tabla 1 Descripción de los métodos basados en la interpolación con funciones climatológicas. f(lluvia) indica una transformación que logra una función de densidad de probabilidad casi uniforme (ver texto). t y t-dt indican los datos del día y del día anterior.

Por ejemplo, el método codificado como "gandin7" imputa los valores de la variable $x_i = \sqrt{lluvia\ diaria}$, tomando la anomalía respecto a la media histórica de la misma. En ese caso, de los términos de la ecuación (1), $\underline{w}$ se determina como se indica en Johnson, 1992 y $b = \overline{x_j}$ (la barra indica promedio en el tiempo). El método clásico de Interpolación Optima sería el aquí denominado "gandin20".

Como la lluvia diaria tiene una distribución de probabilidad muy irregular, se diseñó una transformación $x = f(lluvia)$ que hace uniforme la densidad de probabilidad de x, excepto para $lluvia = 0$. La transformación es invertible y x varía en el intervalo $[0,1]$.

- **mínimos cuadrados ordinarios**

Este método es completamente estándar y su teoría puede encontrarse en cualquier libro de texto. Los pesos $\underline{w}$ se eligen de forma de minimizar la norma 2 del vector $M^{(j)}\underline{w} - m^{(j)}$. (número proporcional al error cuadrático medio), siendo $M^{(j)}$ la matriz con los datos disponibles (tantas filas como fechas, tantas columnas como estaciones, excepto la j-ésima) y $m^{(j)}$ un vector columna con los datos de la estación j-ésima. El caso que se implementó asume que los datos no tienen error, por lo que la expresión que define el vector $\underline{w}$ es (omitiendo el superíndice j):

$$M^T M\underline{w} = M^T . m$$

El superíndice T indica traspuesta. La constante b de la ecuación (1) es cero.

- **mínimo promedio**

Aquí se determinan los pesos $\underline{w}$ por la vía de minimizar la norma 1[1] del vector $M^{(j)}\underline{w} - m^{(j)}$. A diferencia del método anterior, los pesos $\underline{w}$ así definidos no son fáciles de calcular: el problema no se traduce en un sistema lineal de ecuaciones, sino que se trata de una optimización no lineal. El costo en tiempo de cálculo es sensiblemente mayor.

- **mínimo percentil 95**

Dado que la población puede estar afectada por algunos pocos errores groseros, es razonable tratar de minimizar algún estadístico robusto, como el percentil 95 de la distribución. Al igual que en el caso anterior, el problema de hallar los pesos es fuertemente costoso en términos de tiempo de cálculo.

- **vecinos más cercanos**

Aquí se analizaron dos criterios: vecinos *geográficamente* más cercanos, y vecinos *cualitativamente* más próximos. En ambos casos la ausencia se imputa tomando uno de los datos disponibles con una precedencia predeterminada, en el primer caso por la mínima distancia, y en el segundo caso por el criterio de un experto meteorólogo. Los pesos aquí son todos ceros, excepto uno, y la constante b es cero.

- **imputar un valor constante**

Este es un método muy simple, que imputa con algún estadístico de la distribución de cada estación. Se ensayó el valor más probable (moda) y el valor esperado (promedio). En esta cuenca el primer número es siempre 0 mm/día y el segundo es próximo a 3 mm/día.

b) métodos no lineales

Estos métodos son de reciente desarrollo, y están basados en modelaciones simples de las redes neuronales biológicas. Se han utilizado por ejemplo para la predicción a corto plazo de la concentración de SO_2 (Boznar et. al., 1993), de la demanda eléctrica (Park et. al., 1991), etc. Las redes neuronales se organizan en capas, la primera de las cuales recibe directamente como estímulo los datos observados; cada neurona de la capa siguiente recibe como estímulo una combinación lineal de los outputs de las capas anteriores obtenidos via una función de transferencia simple, por ejemplo la logsig (Demuth et. al., 1994) dada por: $output = \left(1 + \exp(-\sum a_i * input_i)\right)^{-1}$, siendo los parámetros a_i a determinar para cada neurona. La red neuronal requiere, al igual que su equivalente biológico, un proceso de aprendizaje, que aquí está simulado por el ajuste de los parámetros a_i para

---

[1] Se recuerda que la norma p de un vector se define como $\|\underline{d}\|_p = \sqrt[p]{\sum_{k=1}^{n}\left(|d_k|\right)^p}$

cada neurona de la red. En el trabajo utilizamos una red de 2 capas, con 6 neuronas logsig en la primera y una neurona lineal en la última, y una red de tres capas, con 8 neuronas lineales, 4 logsig y una logsig a la salida. A ambas se las entrenó con un tercio de los datos disponibles, de forma que los valores de a_i minimizaran el error cuadrático cometido. En el segundo caso, al ser la última capa logsig, la salida toma valores únicamente en el intervalo $[0,1]$; se usa la inversa de $x = f(lluvia)$ para calcular la lluvia. En el primer caso, al ser la última capa lineal, este problema no aparece (Demuth et. al., 1994).

Resultados y conclusiones

Se presentan en la tabla 2 los resultados obtenidos luego de 250 simulaciones. Se destaca la mejor performance de los métodos que usan información del día anterior (gandin4, gandin6 y gandin_diario). De los que sólo usan información del día, los que dan mejor son los correspondientes al mínimos percentil 95, seguido muy de cerca por el método de mínimos cuadrados.

	Promedio	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
mincuadrados	**2.34**	1.33	**4.09**	13.13	**7.01**
minpercentil95	**2.34**	1.34	**4.10**	**13.07**	**7.01**
minpromedio	**2.26**	0.86	3.60	**13.23**	7.21
valor modal	2.79	**0.00**	**1.78**	19.04	10.26
promedio histórico	4.73	2.96	**3.02**	16.25	9.88
vecino por distancia	2.76	**0.02**	4.22	17.37	9.13
vecino por experto	2.82	**0.01**	4.31	17.74	9.33

Tabla 2 Resultados preliminares en mm/día para los diferentes métodos de imputación. Se presenta el valor esperado, los percentiles 75,85 y 95, y la raíz del error cuadrático medio para la distribución del valor absoluto de la diferencia entre el dato imputado y el disponible. En negrita los cinco mejores resultados obtenidos

Nótese que dado que la base de datos contiene errores, es posible que los métodos sugieran valores razonables y sin embargo esos outliers estén afectando los estadísticos considerados. Ello no sucede para el percentil 95, 85, etc., y allí aparece con interés la red neuronal de dos capas ocultas bp7. Estas redes fueron entrenadas con el algoritmo de *backpropagation* (Rumelhart et. al., 1986) al que se le limitó (por razones prácticas) el número de iteraciones, por lo que sería posible seguir entrenándolas para mejorar su performance. El costo del entrenamiento en término de uso de CPU de estas redes es muy alto: aproximadamente 10 horas de SUN 20 para cada estación meteorológica.
Como conclusiones: a) las metodologías basadas en la simple sustitución por un vecino o por una constante dieron resultados pobres b) las basadas en la interpolación óptima dieron resultados casi óptimos en RMS como era de esperar, junto con los mínimos cuadrados y mínimo percentil. c) las metodologías no lineales, si bien muy costosas en la fase de entrenamiento, dieron resultados muy interesantes por lo que se seguirá trabajando en su perfeccionamiento.

Referencias

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., V27B, N 2, pp. 221-230

Demuth, H. and Beale, M. 1994 "Neural Network User's guide (Toolbox for MATLAB)" The MathWorks, Inc. 226 páginas, 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., Vol 116, pp 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.; Martins, P.R., 1990. "Cálculo dos coeficientes de Thiessen em microcomputador". En 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.

Park, D.C. et. al., 1991 "Electric load forecasting using an artificial neural network" IEEE Transactions on Power Systems, N 2, pp. 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, pp 533-536