ANALISIS POR COMPONENTES PRINCIPALES DE DATOS PLUVIOMETRICOS
B) APLICACION A LA ELIMINACION DE AUSENCIAS

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

Resumen

La eliminación de ausencias en los bancos de datos es un tema recurrente en todo estudio vinculado tanto a fenómenos naturales como de otras áreas. El presente trabajo fue motivado por la necesidad de completar un banco de datos pluviométricos a ser utilizado en conjunto con un modelo hidrológico. La lluvia media sobre cada subcuenca se calcula según el método de Thiessen, que en principio no requiere de la eliminación total de ausencias. Sin embargo, ese método es muy sensible a los errores si existen pocos registros simultáneos. La detección de errores se presenta en un trabajo adjunto, y aquí se detallan los resultados obtenidos al aplicar varios métodos para eliminar ausencias. Los mismos deben conservar las características principales del banco y ofrecer garantías de no disminuír los niveles de calidad del mismo. Se describen los resultados comparativos obtenidos sobre un banco de 15 años de datos pluviométricos diarios, al que se le aplicaron cuatro criterios: uso de datos medidos en estaciones próximas, datos calculados por interpolación temporal entre registros, interpolación temporal de coeficientes principales y penalización de coeficientes principales. Este último, diseñado específicamente para este trabajo, demostró ser el de mejor desempeño.

Abstract:

Imputation in order to avoid missing values is a common problem in all studies in natural as well as social sciences. The starting point of this work was the need of imputate missing values in a pluviometric data bank, which will be used in the developments of an hydrological model. The mean areal rain-rate is calculated by Thiessen`s method, which not necessarily requires a full data bank. However, that method proves not to be quite robust against errors, if little simultaneous registers exist. The outlier detection phase is presented in a companion paper, and here the differents techniques applied in order to imputate the missing values are described .The imputation technique should preserve the important features of the data, in order to not create outliers by itself. The comparative results obtained with a daily record of 15 years long are presented. Four different criteria were applied: imputation with the nearest neighbor; linear time interpolation within single station records; linear time interpolation using all station records in a multivariate fashion and the newly developed penalty of principal coefficients, which proves to be the most accurate.

1. Antecedentes

En el campo de la Hidrología y la 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.Para el cálculo de lluvia media sobre una región, existen también métodos como el de Thiessen (Jácome Sarmento *et al.*, 1990) que no requieren en principio, de un banco de datos completo.

Ambas situaciones han llevado a que el tema del tratamiento o eliminación de ausencias tenga un interés quizá menor, lo que se refleja en lo escaso de los trabajos específicos en la literatura especializada consultada.

En opinión de los autores en la mayoría de los casos prácticos, el dato ausente es simplemente ignorado, bajo la hipótesis implícita que estas ausencias son al azar, extremo que no necesariamente es verificado.

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

El tema en cambio, es de gran interés en el área de la estadística y las ciencias sociales en general, pudiéndose encontrar en libros específicos (Rubin, 1987) citas a volúmenes producidos por grupos de trabajo dedicados al tópico.
Existen métodos de imputación más o menos sofisticados. Entre éstos últimos, se puede citar el utilizado por la oficina del censo de los EEUU (Rubin, 1987). El mismo consiste en asignar al dato ausente un valor tomado al azar de entre los restantes eventos que tienen idéntica respuesta en el resto del cuestionario. Si eventualmente no existiese otro igual, o bien se relativiza esa exigencia, admitiendo que alguna o algunas respuestas no lo sean, o bien, se introduce una "distancia" entre cuestionarios, y se busca aquel que diste menos.
Otro método también simple, es el de hacer una regresión sobre el conjunto de datos, ajustando un modelo sencillo. Típicamente, se utilizan mínimos cuadrados (total o parcialmente) ó componentes principales, métodos que Stone *et al*., 1990 presenta desde una perspectiva integrada.
Estos métodos, al igual que el que se presentará luego, producen una única alternativa: para una ausencia, una única imputación. Según Rubin, 1987, "..en general, es intuitivamente claro que imputar la predicción 'optima' para cada ausencia subestimará la variabilidad...". Existe, sin embargo, la posibilidad de imputar más de un valor para una misma ausencia. Así Rubin presenta una variedad de técnicas, algunas excesivamente especializadas, para su aplicación en encuestas. Como idea general, se propone crear para cada ausencia, un número m (pequeño) de alternativas, y considerar que se dispone de m conjuntos completos diferentes. Para el caso en que la tasa de ausencias es baja, el método funciona razonablemente bien, requiriéndose sin embargo más espacio (para guardar las múltiples imputaciones) y más tiempo de cálculo (para procesar los diferentes conjuntos completos generados). Por detalles se remite al lector a Rubin, 1987.

2. Introducción

2.1 Origen del trabajo

El presente trabajo es una extensión de lo realizado en el tratamiento de los datos pluviométricos utilizados para la calibración de un modelo numérico de tipo Caudal-Precipitación, Caudal, para la cuenca del Río Negro. En la misma operan tres centrales hidráulicas en cascada, administradas por el ente eléctrico nacional UTE[2] . Por detalles de ese trabajo, ver Silveira *et al*. (1991, 1992a y 1992b).

2.2 Características generales de la zona en estudio

a) Geografía

Si bien el trabajo incluyó una superficie mucho mayor, se restringió para este análisis a la cuenca del Río Tacuarembó, con una extensión de aproximadamente 20.000 km , ubicada en 32 latitud sur y 55 longitud oeste, a unos 400 km de la costa oceánica. La zona se caracteriza por un suave relieve con alturas que no superan los 500 m, pocos valles abruptos y sin grandes espejos de agua. La media pluviométrica mensual típica para esa zona oscila entre 74 y 120 mm/mes.

b) Red Pluviométrica.

La red definida por la DNM[3] se basa en una grilla con elementos cuadrados de 10 por 10 km, numerados correlativamente. Las estaciones reciben su nombre del cuadro en que están ubicadas y si en algún período en un cuadro opera más de una estación, se le agrega al nombre una letra, (A, B, C), y a los efectos de este

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

trabajo serán consideradas sinónimas. Administrativamente la red está formada por la superposición de otras cuatro, a cargo de distintos organismos, con diferente densidad espacial y nivel de fiabilidad. La estructura de la red ha ido cambiando con el correr de los años y cada estación ha tenido particularidades como ser:

- Haber entrado en funcionamiento en cualquier momento del período.
- Haber salido de funcionamiento en cualquier momento del período.
- Haber sido creada para reemplazar a otra que salía de servicio.
- Puede haber sido reemplazada por una sinónima o no.

A los efectos del presente trabajo sólo se distinguen las estaciones que corresponden a AFE [4], cuyas mediciones acumuladas de los días domingo y lunes no han de ser consideradas. En la subcuenca seleccionada operan 21 estaciones, de las que se seleccionaron 13 para este trabajo.

c) Banco de Datos.

Como se ha dicho, la topología de la red ha sufrido diversas transformaciones. Según las conclusiones de Silveira *et al*. (1991) hay en la actualidad un exceso de estaciones. Muchas de ellas presentaron sinónimos, por lo cual su historia se conformó con la unión de historias de sus sinónimos.
Para este trabajo se seleccionó un conjunto de 13 estaciones, ubicadas según la fig.1, las cuales fueron cuidadosamente depuradas de errores de digitación mediante la aplicación de diversos algoritmos detallados en López *et al*. (1994). El período en estudio comprende casi 15 años, del 1/1/75 al 2/12/89.

3. Métodos utilizados en la eliminación de ausencias

3.1 Por Proximidad

Consiste en asignar a cada estación que se desea completar, una lista de estaciones alternativas, de las que se extraerán los datos faltantes en la original. La lista en principio está en orden creciente de distancia a la estación original, pero también se tiene en cuenta el nivel de fiabilidad, alterando un poco ese ordenamiento.

3.2 Por interpolación temporal entre registros

Cuando falte el dato correspondiente al día t_f t en la estación j se buscan los días anterior y posterior más próximos, en los que se tenga dato medido en esa estación, y se interpola linealmente.
Sean t_f el día a imputar y $p_j(t_f)$ el registro desconocido para el día t en la estación j.
Sean t_{f-m} el último día anterior a t_f con dato, y t_{f+r} el primer día posterior a t_f con dato $\left(t_{f-m} < t_f < t_{f+r}\right)$. La interpolación para el dato buscado es

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

[4] AFE - Administración de los Ferrocarriles del Estado

3.3 Por interpolación temporal de coeficientes principales

Este método se basa en el Análisis de Componentes Principales (ACP), y que ha sido tratado en *López et al*.(1994). Aquí sólo se describe brevemente la notación, y se remite al lector a la referencia citada.
Sea $\mathbf{P}_{(n,1)}(t)$ el vector de precipitaciones de las n estaciones elegidas, para el instante t. Se considera la matriz $\mathbf{M}$ cuyas filas son los vectores $\mathbf{P}(t_{m_j}) - \mathbf{P}_M$, $j = 1..r$, definidas para aquellos días en que no faltan datos. $\mathbf{P}_M$ es el vector de precipitaciones medias en el período.
Los vectores propios de $\mathbf{C}_{(n,n)} = \mathbf{M}^T * \mathbf{M}$ serán denominados patrones, y se denotan como $\mathbf{e}_i$. Se supondrá que los valores propios asociados son decrecientes con i. La relación entre los registros pluviométricos $\mathbf{P}_{(n,1)}(t)$ y el vector de coeficientes $\mathbf{A}_{(n,1)}(t)$ está dada por

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

donde $\mathbf{P}_M$ es el vector de precipitaciones medias en el período, y $\mathbf{E}_{(n,n)}$ la matriz formada por los vectores propios $\mathbf{e}_i$.

$$\mathbf{P}(t) = \begin{bmatrix} p_1(t) \\ . \\ . \\ . \\ . \\ p_n(t) \end{bmatrix} ; \mathbf{P}_M = \begin{bmatrix} \overline{p_1} \\ . \\ . \\ . \\ . \\ \overline{p_n} \end{bmatrix} ; \mathbf{A}(t) = \begin{bmatrix} a_1(t) \\ . \\ . \\ . \\ . \\ a_n(t) \end{bmatrix} . ; \mathbf{E} = \begin{bmatrix} \mathbf{e}_1 \, \mathbf{e}_2 \, .. \mathbf{e}_n \\ \\ \\ \\ \\ \end{bmatrix}$$

La matriz $\mathbf{E}_{(n,n)}$ es invertible, por lo que dados los datos $\mathbf{P}(t_{f-m})$ y $\mathbf{P}(t_{f+r})$ es posible obtener los vectores $\mathbf{A}(t_{f-m})$ y $\mathbf{A}(t_{f+r})$ correspondientes.
La ecuación (2) también se puede expresar como

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

Para el tiempo intermedio $t_l, l \in (f-m+l, f+r-1)$ la lluvia se calcula mediante interpolación lineal el vector $\mathbf{A}(t)$. Todos los valores de la precipitación para ese día, se pueden obtener en principio de la ec. (2).
Del análisis de los coeficientes a_i surge que cuanto mayor es el índice i el coeficiente a_i tiene una desviación estándar menor por lo que su aporte a la suma también es menor típicamente.
Lo anterior justifica que en la reconstrucción del vector $\mathbf{P}(t)$ se desprecien los términos para i>q, para algún q, sin perder información esencial, sustituyéndose la fórmula (3) por:

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

En resumen, para un día t_f en que falte algun dato del vector $\mathbf{P}(t_f)$ se buscan los días más próximos, anterior y posterior, en los cuales se tenga dato medido en todas las estaciones. Se hace notar que en este método se trabaja con el conjunto de las n estaciones, no con cada una por separado.
Sea t_f el día a imputar. Sean t_{f-m} el último día anterior a t_f con datos completos y t_{f+r} el primer día posterior a t_f con datos completos $\left(t_{f-m} < t_f < t_{f+r}\right)$. Se calculan los coeficientes $\mathbf{A}(t_{f-m})$ y $\mathbf{A}(t_{f+r})$ correspondientes a los vectores $\mathbf{P}(t_{f-m})$ y $\mathbf{P}(t_{f+r})$ con la ecuación (2).
Para el momento t_{f-m+l}, se calcula el vector $\mathbf{A}(t_{f-m+l})$ interpolando linealmente los vectores $\mathbf{A}$ anteriormente mencionados. El valor tentativo de la precipitación para
ese día, $\mathbf{P}(t_f)$, se calcula con la ec (4).

Las faltantes de la base de datos correspondientes a componentes del vector $\mathbf{P}(t_f)$ se toman de los valores del vector tentativo.
Una vez completado el día t_f, se reinicia la interpolación, utilizando los vectores $\mathbf{P}(t_{f-m+l})$ y $\mathbf{P}(t_{f+r})$ como puntos de partida, hasta completar todos los faltantes.

El mejor o peor desempeño de esta aproximación, está vinculado a las características de la función de autocorrelación de los a .
Usualmente, para otras variables meteorológicas, las propiedades de autocorrelación de la serie temporal de a son muy diferentes entre sí. Esto es otra justificación para limitar el número de sumandos.
Se muestra en Cisa et al. (1990) que para el parámetro viento de superficie, el tiempo T en horas en el cual la autocorrelación de la serie de los a toma por primera vez el valor 0.5 es (25,9,5,3,3,...,1.2,1) para i=1...15, correspondiendo los valores mayores para los componentes más importantes. El dato es observado en forma horaria.
Tal situación no se da en la lluvia, ya que todos los patrones exhiben una dramática caída de la autocorrelación al cabo de un día, siendo éste el período de muestreo (ver figs. 2 y 3). Ello explica el pobre resultado obtenido, aún siendo mejor que el derivado de la Interpolación temporal entre registros.

3.4 Por penalización de coeficientes principales

Si se realiza un histograma de los valores de cada a_i se observa que para los patrones más importantes, el mismo es fuertemente asimétrico, o tiene una dispersión respecto a cero muy grande. En cambio, para los más débiles, la dispersión alrededor de cero es típicamente muy pequeña, y la distribución simétrica. En las figs. 4 y 5 se presentan a modo de ejemplo esos histogramas para el caso en que se han removido los días con precipitación nula en todas las estaciones dato. Cualquiera sea el esquema elegido para eliminar ausencias, se supone que producirá valores de los a_i coherentes con estos histogramas, o sea, típicamente muy pequeños para los patrones débiles. Ello puede ser impuesto como condición eligiendo, para una fecha determinada, las componentes del vector de lluvias $\mathbf{P}(t_f)$ cuyos valores se desconocen de forma de hacer mínima la expresión

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

donde los $a_i(\mathbf{P})$ son los coeficientes correspondientes al vector $\mathbf{P}$ completado y los w_i coeficientes de ponderación que tienen en cuenta el diferente valor absoluto de cada a_i. El vector $\mathbf{P}$ se supone tiene q incógnitas (o ausencias). El mínimo de S se obtiene haciendo nulas las derivadas parciales

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

siendo $p_{m(j)}$ los registros faltantes para esa fecha. El sistema lineal así definido se puede resolver por técnicas estándar.

4. Metodología del análisis

Se generan al azar parejas de fecha - estación, que serán consideradas como ausencias ficticias. Luego se verifica que para cada fecha - estación exista un dato: si existe un "Valor Real", los métodos calcularán un valor para sustituírlo, que se llamará "Valor Calculado"; si no existe dato, la pareja no se analiza.
Una vez que se procesan todas las parejas se evalúa la desviación estándar de la diferencia entre los valores real y calculado para cada método, siendo éste el estimador con que se comparará el desempeño de los mismos.
Las ausencias ficticias que se dan como entrada a los diferentes métodos se crean como un porcentaje del total de días del período de análisis, 5450 días (del 01/01/75 al 02/12/89), y sin considerar más que una ausencia por día. Se comenzó con un 20 % y se finalizó con un 80 % y no se encontró que esto cambiara cualitativamente los resultados obtenidos.
Para todos los métodos se hicieron corridas considerando todas las ausencias; para los de proximidad y Penalización se realizaron además cálculos ignorando las ausencias con valor pluviométrico igual a cero, que son un 80% del total de los datos. Con ello se procuraba evaluar el impacto negativo en los estimadores, ya que se eliminaba una importante masa de constantes. Esta última forma de análisis multiplicó aproximadamente por dos los valores de las desviaciones estándares correspondientes obtenidas analizando todos los casos, pero se mantuvieron los valores relativos entre los diferentes métodos.
Los datos de la subcuenca elegida para realizar los análisis habían sido debidamente revisados y depurados y las trece estaciones elegidas tienen en promedio un 95% de los datos correspondientes a ese período, lo que da las garantías necesarias para la realización del trabajo.
Se intentó estudiar la sensibilidad de los diferentes métodos a los parámetros impuestos: para los modelos de Interpolación Temporal de Coeficientes Principales y Penalización de Coeficientes Principales, se hicieron corridas variando la cantidad de términos a considerar; para el modelo de Proximidad se varió la distancia media entre la estación y las que le rellenaban, eliminándole alternativas.

5. Resultados obtenidos

Los resultados aquí presentados corresponden a un total de 2091 días con ausencias simuladas, que es aproximadamente un 53% del total de días analizados.

5.1 Asignación por proximidad

Se realizó una serie de cálculos con las trece estaciones seleccionadas para el análisis y otra con una lista de estaciones, clasificadas estrictamente por proximidad, que incluía 86 estaciones, incluso alguna fuera de la cuenca en estudio.
El hecho de que se disponga de una densa red de estaciones, ubicadas en una geografía muy regular, hace que este método dé buenos resultados. Al aumentar la distancia media entre la estación a completar y las lternativas se ve (fig. 6) que la desviación estándar tiende a aumentar.
Lo regular del fenómeno en el área hace que aún con distancias medias de más de 150 km se obtengan buenos resultados. Esta distancia media se define como el valor esperado de la distancia geométrica, ponderado por la frecuencia con que cada estación aportó un dato. La distancia máxima entre dos estaciones (de las trece seleccionadas) es de 201 km.
El punto con distancia media = 33.7 km en la fig. 6 fue encontrado utilizando como lista de alternativas únicamente el conjunto de las 13 estaciones, resultando σ = 5.55 mm/día, por lo que es importante a la hora de comparar el desempeño de los diferentes métodos.

En general, las estaciones utilizadas como alternativa no fueron sistemáticamente depuradas de errores de digitación; ello se refleja en que aún con una distancia media menor, el imputar con tales estaciones arroja resultados con desviación estándar superior a la que se obtiene con las trece estaciones seleccionadas.
Es importante resaltar que, para la aplicación de este método, se utilizó en general una red mucho más densa que la formada por las trece estaciones en estudio.

5.2 Asignación por interpolación temporal entre registros

Como se ha dicho, se pueden esperar buenos resultados cuando el fenómeno presenta cambios muy lentos con respecto a la frecuencia de muestreo, o la cantidad de ausencias consecutivas es menor a los tiempos típicos de variación del fenómeno.
En el Uruguay, la lluvia tiene características de gran irregularidad en el tiempo y generalmente se da en forma de tormentas más o menos cortas que pueden tener una duración de unas pocas horas hasta dos o tres días. Por ello, con mediciones diarias no es posible que este método dé buenos resultados.
Los resultados del análisis para las 13 estaciones con este método dieron una desviación estándar en el orden de los 12 mm/día.

5.3 Asignación por interpolación temporal de coeficientes principales

Los resultados de este método son comparables con los de la Interpolación Temporal entre Registros, dado que responde a las mismas condicionantes: autocorrelación temporal y frecuencia de muestreo del fenómeno, y si bien insume un costo computacional mayor en su aplicación, los resultados son apenas mejores, oscilando el error entre 11.3 y 11.83 mm/día, según la cantidad de términos empleados en los cálculos.
La poca variación entre emplear uno o más términos en el cálculo está vinculada a que el fenómeno tiene autocorrelación temporal muy débil. En la fig. 7 se observa la evolución del estimador al variar el índice q (ver ec. (iv))
Se espera que este método arroje resultados sensiblemente mejores para otras variables, -tipo temperatura, viento, presión, etc-, pero al presente no ha sido posible verificarlo.

5.4 Asignación por penalización de coeficientes principales

Este método obtuvo los mejores valores en términos de la desviación estándar, alcanzando un mínimo de 4 mm/día. Como medida comparativa, la serie de las lecturas en una estación particular cualquiera, tiene una desviación típica superior a 15 mm/día, tomada a lo largo de todo el período.
Este valor de 4 mm/día se obtiene al penalizar con k comprendido entre 8 y 10 (ver ec. (iv)), lo que implica un número de términos entre 5 y 3.Por ejemplo, cuando k es 1, todos los coeficientes son penalizados, por lo que se obliga a la superficie de lluvias a parecerse a la correspondiente al valor medio. Para k próximo a n, sólo están afectados en la suma S el ruido correspondiente a los patrones más débiles, pero se deja libres otros a que son también ruido. Ello hace que el algoritmo sea muy inestable, produciendo valores de peor calidad.
El problema de determinar el k óptimo para cada situación, puede ser abordado mediante un experimento similar a éste, o a través de un análisis subjetivo de los propios patrones $\mathbf{e}_i$. La forma de las isoyetas que los mismos representan permiten distinguir fácilmente los que constituyen ruido de los datos, de los que representan el comportamiento físico del fenómeno. Los patrones débiles, son además muy sensibles a valores erróneos aislados en la base de datos (ver Silveira et al. 1991).
Obsérvese en la fig. 8 la evolución del error vs. el número de términos utilizados. El criterio se muestra robusto frente al índice k.
Los pesos w_i fueron especificados durante el trabajo, de forma que los términos w_i .a fueran de un orden comparable. Ello se hizo inicialmente, adoptando para w_i el recíproco de la varianza de la serie a_i . Ello, si bien razonable, se reveló inadecuado para el tratamiento de los primeros patrones, en que la distribución es marcadamente asimétrica.

Por ello, se tomó $w_i = 1/\alpha_i^2$, donde α_i es tal que

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

siendo f la función de distribución del coeficiente a , para i=1..n. Con ello, para menos del 4% de los casos, se hace $w_i . a_i^2(t) \geq 1.0$

6. Conclusiones y recomendaciones

De lo analizado se desprende que los métodos basados en el comportamiento temporal del fenómeno dan, en este caso, resultados que podrían distorsionar en forma significativa las características generales del banco de datos, en particular si la cantidad de ausencias es un porcentaje importante del total de información.
Los métodos de Proximidad y Penalización de Coeficientes Principales, que consideran el comportamiento espacial del fenómeno, se comportan significativamente mejor que los que consideran el comportamiento temporal. Ello se explica especialmente por las características del fenómeno precipitación frente a la frecuencia de los muestreos y por una cuenca de superficie pequeña con una geografía que no tiene grandes accidentes.
El método aquí presentado de Penalización de Componentes Principales exhibe un error 28% menor que el de Proximidad (4.01 vs 5.55 mm/día), valor que corresponde a una lista de alternativas tomadas del conjunto de las 13 estaciones.
Si se consideran otras estaciones, el error es mayor, incluso para distancias medias menores, lo cual es explicable debido a que las otras estaciones no fueron depuradas.
Otra ventaja no desdeñable es que la propuesta de un valor para completar los datos, puede ser calculada en tiempo real en un computador de porte mínimo. Sólo se requiere del mismo que conserve una matriz de n*n, el vector de precipitaciones medias, y que sea capaz de resolver un sistema de ecuaciones lineal. En el caso de la operación rutinaria de un modelo hidrológico, esta posibilidad no debe dejar de ser considerada.
Como una futura mejora al procedimiento, se plantea el maximizar la probabilidad conjunta de los a_i , lo que implica la solución de un problema no lineal para cada evento con ausencias. Ello incrementará sensiblemente el costo en términos de tiempo de máquina, pero es teóricamente más justificable.

7. Reconocimientos

Juan González implementó y realizó los cálculos de la asignación por proximidad e interpolación temporal de registros, y Rosario Curbelo la asignación por interpolación temporal de coeficientes. Carlos López desarrolló la asignación por penalización de coeficientes principales. El diseño de la metodología, así como del experimento también estuvo a cargo de Carlos López. El análisis de los resultados estuvo a cargo de los tres autores.

8. Agradecimientos

Se deja constancia que este trabajo es una extensión de las tareas realizadas en el marco del convenio "Desarrollo de un modelo matemático-hidrológico de la cuenca del Río Negro" por encargo de UTE. Se agradece la autorización para utilizar la información disponible y publicar los resultados.

9. Referencias

González, E.; Morales, C., 1991. "Depuración de la base de datos pluviométricos de la cuenca del Río Tacuarembó". Informe interno preparado para el Departamento de Hidrología del Instituto de Mecánica de los Fluidos e Ingeniería Ambiental. 11 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.

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.

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.

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" Informe final. Parte 2, Cap. 4. 83 pp.

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" . Publicación Interna del Dpto. Hidrología, IMFIA 1/92, Instituto de Mecánica de los Fluidos, 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" Publicación Interna del Dpto. Hidrología, IMFIA 2/92, Instituto de Mecánica de los Fluidos, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.