ANÁLISIS POR COMPONENTES PRINCIPALES DE DATOS PLUVIOMETRICOS
A) APLICACIÓN A LA DETECCIÓN DE DATOS ANÓMALOS

Carlos López [1] , Elizabeth González [2]
y Jorge Goyret [1]
[1] Centro de Cálculo, Facultad de Ingeniería, CC 30, Montevideo, Uruguay
[2]Instituto de mecánica de los fluidos e Ingeniería ambiental, Facultad de Ingeniería, CC 30, Montevideo, Uruguay

Resumen

Se describen las técnicas aplicadas en el tratamiento de un banco de datos pluviométricos, utilizado en la fase de desarrollo y calibración de un modelo hidrológico del tipo caudal-precipitación, caudal.
Dado que la calibración de este tipo de modelos es muy afectada por errores en los registros, es que se hace necesario eliminar aquellos que sean inconsistentes. Se han aplicado en este caso, una variedad de estrategias. De entre ellas, el Análisis por Componentes Principales (ACP) es el que dio mejores resultados.
La metodología desarrollada permite además realizar control en tiempo real de los nuevos datos recogidos, con un mínimo de recursos informáticos, lo que habilita a su aplicación en forma independiente de grandes equipos de cómputo. Para la etapa del trabajo que se describe, se han considerado como errores solamente aquellos registros digitales que no coinciden con el valor asentado por el observador en la planilla.
Sin embargo, se entiende que el ACP permite detectar también errores aleatorios del observador y aún ciertos tipos de errores sistemáticos, lo que es aún tema en estudio.

Abstract:

The techniques employed in the treatment of a pluviometric data base used during the development and calibration phases of a Flow-Rain, Flow hydrological model are presented.
It's well known that such a model is strongly affected by errors (outliers) in the data, both random and systematic. So it is needed to remove them prior to use the data bank.
For this case, some different methodologies have been applied. From them, the most successful was the one based in Principal Component Analysis (PCA).
The methodology is liable to be used in real time, involving minimum computer resources. For the stages described here, only errors coming from manually digitizing are considered. However, it is suggested that PCA may help in detecting random errors from the observer himself, and also some kind of systematic errors, all of which is still in an investigation phase.

1. Introducción

1.1 Planteo del problema

En todo banco de datos existen dos fuentes de errores: aquellas inherentes a la operación de medida, y las generadas en tiempo de ingreso o proceso de la información. Ambos tipos de error pueden tener un efecto más o menos importante según el problema en estudio. Según Husain, 1989, "...el fracaso de muchos proyectos de abultado presupuesto puede ser atribuido en parte, a la imprecisión de la información hidrológica manejada...". En el caso de los modelos hidrológicos, los errores se propagan en el tiempo, y dependiendo de las características de la cuenca, su efecto es más o menos persistente.
En la operación diaria de este tipo de modelos, es muy fácil para el usuario notar errores importantes, ya que evalúa al día siguiente la bondad de la predicción.
En cambio, en la etapa de calibración o ajuste de parámetros empíricos del modelo, no es posible analizar una secuencia de miles de parejas caudal medido vs. caudal calculado si no es a través de estimadores como la desviación estándar, etc.
Ello oculta en el conjunto, tanto aquellos eventos nítidamente erróneos como otros más sutiles, sesgando el valor de los parámetros en forma descontrolada. A los efectos de la depuración del banco, se han tomado los datos asentados en papel por el observador como totalmente ciertos, y se detectan errores en el proceso de digitación. Como se verá es posible extender el método a la identificación de errores tanto

aleatorios como sistemáticos, producto de una inadecuada exposición los últimos, y de falta de celo los primeros.
El presente trabajo es una extensión del realizado durante la calibración de un modelo hidrológico de tipo caudal-precipitación, caudal, para la cuenca del Río Negro. Por detalles de ese trabajo, ver Silveira *et al.* (1992a y 1992b).

1.2 Antecedentes metodológicos

Para la detección de datos anómalos, el único antecedente nacional registrado consiste en las recomendaciones realizadas por la Dirección de Climatología y Documentación de la Dirección Nacional de Meteorología (DNM, 1988). Las vinculadas a datos anómalos en pluviometría son sumamente laxas, y se basan en un control por rango admisible.
A nivel internacional, existen trabajos (Sevruk, 1982) que proponen procedimientos para corregir errores sistemáticos en cada estación. Se requiere conocer, entre otros, la velocidad del viento, la intensidad de la lluvia, la temperatura y humedad de aire, etc.
Con respecto a los errores aleatorios, la tendencia es a cruzar las medidas con un modelo del fenómeno (p. ej.: Francis, 1986; Hollingsworth *et al.*, 1986). Este último asevera que para el caso del viento, las diferencias entre observaciones y predicciones tienen aproximadamente una distribución normal. En ese caso, es relativamente fácil detectar los datos anómalos y separarlos para un análisis a posteriori. Como desventaja debe señalarse el importante volumen de información requerido, así como los altos costos computacionales involucrados.
Si se prescinde o se desconoce la relación física que debería ligar a las variables, los métodos puramente estadísticos son una alternativa a evaluar. Barnett *et al.*, 1984 efectúa una síntesis de distintas técnicas aplicables para el abordaje de este problema. En el caso del análisis multivariado de datos, distingue dos grandes líneas metodológicas, según que la función de distribución de la muestra es supuesta conocida, o no.
El primer grupo corresponde a los llamados Tests de discordancia, que agrupa una serie de técnicas aplicables según la forma en que se distribuyen los datos muestreados, y requieren conocer -o poder estimar- además los parámetros de la distribución. Existen también antecedentes vinculados al caso en que la distribución teórica responda a un tipo de ley y los datos muestreados a otra, como es el caso del planteo de O'Hagan, 1990, en que el hecho de que una de ambas distribuciones sea normal y la otra de tipo t habilita al uso de cierta metodología para poner en evidencia los datos anómalos. El caso aquí tratado no es abordable a partir de este tipo de métodos, puesto que al analizar la distribución de datos diarios de lluvia, la misma no encuadra en las leyes allí tratadas.
La segunda línea identificada por Barnett corresponde a lo que se ha dado en llamar Métodos informales. Estos prescinden de los aspectos formales de la distribución de los datos, y apuntan a explotar ciertas propiedades de los mismos. En este grupo se encuentran los métodos de detección de marginales, fijando un rango de probabilidad; los métodos gráficos, basados en la búsqueda de puntos alejados de la nube de datos; la aplicación de métodos de correlación (Gnanadesikan *et al.*, 1972); la búsqueda de distancias generalizadas representativas, técnicas asociadas con el análisis de agrupamientos (cluster) (ver por ejemplo, Fernau *et al.*, 1990) y análisis de componentes principales (ACP), entre otros.
Un antecedente muy específico respecto al ACP lo presenta el trabajo de Hawkins, 1974. En él se comparan cuatro indicadores o estadísticos, diseñados para resaltar datos anómalos. Hawkins asume que cada observación tiene distribución normal, por lo que su hipótesis no es aplicable al caso de lluvia; sin embargo, los conceptos por él vertidos son similares a los aquí manejados, así como los resultados obtenidos por él en muestras de carbón.

2. Breve presentación del ACP

El ACP es una de las técnicas de análisis multivariado de mayor aplicación en los últimos tiempos (Richman, 1986 como revisión general; Pio *et al.*, 1989 para contaminantes en atmósfera; White, 1991, para precipitación, etc.). Ella permite transformar una serie de registros correlacionados, en otra serie de variables con correlación mutua nula, lo que permite tratar a cada una como independiente.
A su vez, estas nuevas coordenadas tienen la propiedad de minimizar la varianza remanente, como se verá, lo que permite detectar y diferenciar entre cada una de ellas, el ruido de la física. No se ha intentado en este trabajo rotar las componentes obtenidas, como sugieren Richman, 1986, White, 1991, entre otros, proceso que según estos autores, mejora la interpretación física de los patrones.

2.1 Planteo teórico

En lo que sigue, se denomina $p_i(\tau_k)$ al valor de la precipitación correspondiente al instante τ_k (k=1..r) en la estación i (i=1..n). Se denominará como $\overline{p}_i$ a la media temporal de $p_i(\tau_k)$, k=1..r.

Dado un conjunto de registros $p_i(\tau_k)$ se les puede representar mediante un vector $\mathrm{P}_{(n,1)}(\tau_k)$ en el espacio R^n (fig. 1). Cada punto k de la nube, corresponde a una fecha τ_k. El origen de coordenadas se toma en el baricentro de la nube, que tiene componentes $\overline{p}_i$ y se denotará como $\mathrm{P_M}$.

Es posible demostrar que existe una dirección $\vec{e}_1$ (en general, única) que minimiza la suma de cuadrados S_1

$$S_1 = \sum_{k=1}^{r} \overline{M_k H_k}^2$$

según se esquematiza en la fig. 1. $\vec{e}_1$ no depende del tiempo τ_k. Se denominará como $a_1(\tau_k)$ a la proyección OH_k. Cada sumando en S_1 puede interpretarse como la norma L^2 del vector

$$\mathrm{P}(\tau_k) - \mathrm{P_M} - a_1(\tau_k).\vec{e}_1$$

Obsérvese que el vector de datos de lluvia para cualquier τ_k se ajusta con un vector constante, más un múltiplo de un vector constante. El término $\frac{S_1}{r}$ es interpretable como la varianza no explicada por la aproximación con un único término.

A continuación puede definirse $\vec{e}_2$ como el vector que minimiza la varianza remanente

$$S_2 = \sum_{k=1}^{r} \left| P(\tau_k) - P_M - a_1(\tau_k).\vec{e}_1 - a_2(\tau_k).\vec{e}_2 \right|^2$$

siendo $a_2(\tau_k)$ la proyección según la dirección $\vec{e}_2$ del segmento $\mathrm{OM_k}$. Incluso geométricamente es posible ver que $\vec{e}_1.\vec{e}_2 = 0$.

Análogamente se procede hasta $\mathrm{S_n}$. En Lebart *et al.*(1977) se demuestra que los $\vec{e}_\mathrm{i}$ son los vectores propios de la matriz de covarianza $C = \left\{ c_{ij} : c_{ij} = \sum_k \left(p_i(\tau_k) - \overline{p}_i \right).\left(p_j(\tau_k) - \overline{p}_j \right) \right\}$ y que los valores propios λ_i están directamente vinculados con los $\mathrm{S_i}$. Se puede ver que las variables $a_i(\tau)$ y $a_j(\tau), i \neq j$, tienen correlación cruzada nula. Si se denomina D a la matriz cuya diagonal está formada por los λ_i , y E a la matriz formada por los vectores propios $\vec{e}_\mathrm{i}$, entonces resulta:

$$C = E.D.E^T$$

En lo que sigue, se denominará como componentes principales, a los vectores unitarios $\vec{e}_\mathrm{i}$, y como coeficientes principales, a la serie de los $a_i(\tau)$ correspondientes. Nótese que el índice i no está asociado con una estación pluviométrica.

En resumen, existe una transformación lineal que vincula las series de registros $p_i(\tau), i = 1..n$, con los coeficientes principales $a_i(\tau)$ mediante

$$P(\tau) = P_M + E.A(\tau)$$

donde P_M es el vector de precipitaciones medias en el período.

$$P(\tau)=\begin{bmatrix} p_1(\tau) \\ \vdots \\ \vdots \\ \vdots \\ p_n(\tau) \end{bmatrix}; P_M=\begin{bmatrix} \overline{p}_1 \\ \vdots \\ \vdots \\ \vdots \\ \overline{p}_n \end{bmatrix}; A(\tau)=\begin{bmatrix} a_1(\tau) \\ \vdots \\ \vdots \\ \vdots \\ a_n(\tau) \end{bmatrix}; E=\begin{bmatrix} \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ \vec{e}_1 & \vec{e}_n & \cdots & \vec{e}_{n-1} & \vec{e}_n \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \end{bmatrix}$$

La matriz E es en general invertible, por lo que dados los datos $p_i(\tau), i=1..n$ es posible obtener los coeficientes $A(\tau)$ correspondientes con la siguiente expresión:

$$A(\tau)=E^{-1}.\left(P(\tau)-P_M\right)$$

La ecuación (iv) también se puede expresar como

$$P(\tau)=P_M+\sum_{i=1}^{i=n} a_i(\tau).\vec{e}_i$$

2.2 Necesidad de una depuración progresiva.

Los vectores $\vec{e}_i$ (también denominados patrones) son calculados a partir de la nube de puntos-dato disponibles. En la misma puede existir un pequeño grupo de valores disparatados, que incidan en la determinación de tales patrones, afectando sensiblemente los mismos.

En Silveira *et al.* (1991), se muestra que en una población con r=4000 eventos, incluso tan sólo dos valores disparatados podían alterar significativamente los patrones. Este hecho obliga a realizar un proceso de depuración recursivo, en el que, en primera instancia, se buscan solamente los casos más gruesos. Como se verá luego, progresivamente se puede ajustar el criterio, para proceder a la detección de problemas más sutiles.

3. Aplicación a un caso concreto: Cuenca del río Tacuarembó

3.1 Características generales de la zona en estudio

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^2, 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. El período en estudio comprende casi 15 años, del 1/1/75 al 2/12/89, con lo que se supera largamente el número mínimo de muestras recomendado por Hawkins, 1974.

3.2 Métodos utilizados para la detección de errores

a) Datos marginales en la distribución univariada

Consiste en determinar rangos "razonables" para los valores registrados en cada estación, y señalar los casos en que no se verifican. En el banco de datos disponible, era usual que se confundieran registros tomados en mm, con registros tomados en décimas. Ello pudo ser corregido si los mismos eran superiores a 100 mm/día, pero era impracticable hacerlo en otros casos.

En el caso de la lluvia, este método puede detectar eventos claramente anómalos por exceso, pero es incapaz de identificar un cero como erróneo, dado que el 80% de los datos, lo es.

b) Comparación de series de Thiessen

Se calcula la precipitación media en el área mediante el método de Thiessen (ver, por ejemplo, Jácome Sarmento *et al.*, 1990), utilizando varios subconjuntos de estaciones tomadas de entre las n disponibles. Las series temporales de lluvias medias en el área son comparadas, y cuando difieren "demasiado", se analiza ese día en particular.

Los resultados obtenidos (no presentados aquí) permiten decir que este criterio es más potente que el anterior, detectando errores en aproximadamente el 30% de los días señalados (Silveira *et al.*, 1991). Además, no es requerido que ellos sean, en sí mismos, marginales.

c) Datos marginales en la distribución multivariada

En el caso estudiado, típicamente dos de cada tres días tenía alguna ausencia. Por ello, para cada τ, deben distinguirse dos situaciones:

c.1) Se dispone de registros en las n estaciones.

c.2) Falta algún registro.

En el primer caso, es posible calcular directamente las n coordenadas $a_i(\tau)$. Si para algún i, $a_i(\tau)$ no está dentro del i-ésimo rango especificado, los n registros de lluvia utilizados en su determinación son revisados. Estos rangos se determinan a partir de la distribución de a_i para todo el período.

En el segundo caso, puede aplicarse el criterio basado en la proximidad, o puede utilizarse algún procedimiento para estimar el(los) dato(s) faltante(s), de forma de reducir el problema al caso anterior. En López *et al.* (1994) se compara el desempeño de cuatro métodos, resultando más eficiente el de penalización de coeficientes principales, allí descrito.

Una vez determinados los datos ausentes, habría que tratarlos como en el c.1), controlando cada a_i. Tanto aquí como en el trabajo de Silveira *et al.*(1991) se aplicó un criterio más laxo; el día era revisado si alguno de los registros estimados era negativo o mayor que 100 mm/día. Por mayores detalles, se remite a López *et al.* (1994).

En las figuras 3 y 4 se presentan la distribuciones típicas de los a correspondientes a los patrones más y menos importantes. Arbitrariamente se decidió utilizar rangos simétricos $(-\alpha_i, \alpha_i)$. Para cada a_i, α_i se determina de forma de dejar fuera menos del 4% de los eventos. Si la distribución es cuasi-simétrica (patrones 2, 3, ..., 17, figs. 3 y 4) ello implica rechazar aproximadamente el 2% en cada extremo de la distribución. Para distribuciones fuertemente asimétricas (patrón 1, fig. 3) se descarta sólo de un lado de la distribución.

4. Resultados obtenidos

4.1 Sobre datos simulados

Con objeto de evaluar la habilidad del método, se seleccionó un grupo de n = 13 estaciones cuidadosamente depuradas, que tenían promedialmente más del 95% de los registros en el período de r=5450 días (aproximadamente 15 años).

Se generó un conjunto de 2832 ternas aleatorias Estación-fecha-dato, para sustituir los valores medidos, lo que representó algo más del 4% de la población.

Los falsos datos de lluvia se generaron siguiendo una distribución réplica de la de los datos reales. En la fig. 5 se representa la misma sin el valor correspondiente a 0 mm/día.

Se aplicó la metodología sugerida para la detección de errores, en busca de los datos sembrados en la base.

En las tablas I y II se presenta la evolución del total de datos erróneos detectados, en cada paso del proceso sugerido. Entre la primer y segunda columna, la diferencia consiste en que se recalculan los márgenes α_i. Los eventos detectados en la primer columna son ignorados a los efectos del cálculo de α_i, pero son seguramente detectados en la segunda pasada. Tampoco se recalcula la matriz de covarianza, ni los vectores propios, ni por tanto los a_i.

Otra estrategia es detectar, depurar y recalcular todo. Los resultados se presentan en la primer columna. Para la siguiente depuración, se eliminan los datos marginales detectados y se procede a recalcular la matriz de covarianza, y los vectores propios.

Se indican en negrita los resultados correspondientes a días con alguna ausencia. En la Tabla I, se expresan los resultados obtenidos en relación a los datos que son teóricamente consultados en planilla; en la Tabla II, se presenta respecto al total de datos sembrados que aún quedan en la base.

	Primer pasada $\frac{\text{total detectados}}{\text{total revisados}}.10^3$	*Segunda pasada* $\frac{\text{total detectados}}{\text{total revisados}}.10^3$
Primer depuración	$\frac{360+\mathbf{211}}{7644+\mathbf{6318}}.10^3 = 41$	$\frac{2065+\mathbf{215}}{54067+\mathbf{6435}}.10^3 = 38$
Segunda depuración	$\frac{151+\mathbf{35}}{5798+\mathbf{5863}}.10^3 = 16$	$\frac{1784+\mathbf{40}}{50924+\mathbf{5837}}.10^3 = 32$
Tercer depuración	$\frac{68+\mathbf{36}}{4966+\mathbf{7514}}.10^3 = 8$	$\frac{276+\mathbf{39}}{9555+\mathbf{7397}}.10^3 = 19$

Tabla I: Evolución de la depuración, en relación a los datos a revisar. Los términos de la tabla corresponden al siguiente esquema $(A+B)/(C+D).10^3$, siendo A: datos anómalos detectados en días sin ausencias; B: datos anómalos detectados en días con ausencias (**en negrita**); C: datos revisados en días sin ausencias y D: Datos revisados en días con ausencias (**en negrita**).

	Primer pasada $\frac{\text{total detectados}}{\text{total sembrados}}$	*Segunda pasada* $\frac{\text{total detectados}}{\text{total sembrados}}$
Primer depuración	$\frac{360+\mathbf{211}}{2832} = \frac{571}{2832} = 0.20$	$\frac{2065+\mathbf{215}}{2832} = \frac{2280}{2832} = 0.81$
Segunda depuración	$\frac{151+\mathbf{35}}{2832-571} = \frac{186}{2261} = 0.08$	$\frac{1784+\mathbf{40}}{2261} = \frac{1824}{2261} = 0.81$
Tercer depuración	$\frac{68+\mathbf{36}}{2261-186} = \frac{104}{2075} = 0.05$	$\frac{276+\mathbf{39}}{2075} = \frac{315}{2075} = 0.15$

Tabla II: Evolución de la depuración, en relación al remanente de errores sembrados. Los términos de la tabla corresponden al siguiente esquema $(A+B)/C$, siendo A: datos anómalos detectados en días sin ausencias; B: datos anómalos detectados en días con ausencias (**en negrita**) y C: datos erróneos aún en la base de datos.

Los resultados muestran que es más conveniente ajustar los márgenes, que recalcular los patrones. Así, para dos pasadas, quedan en evidencia el 81% de los datos sembrados, que pertenecen al 49% de los días revisados.

Si en cambio se deseare recalcular los patrones al detectar los primeros 571 errores, en la segunda depuración se encuentran solamente 186 errores, que es sólo el 21% de los días a revisar.

Tal comportamiento no fue el observado al trabajar con los datos en bruto: allí unos pocos errores afectaban significativamente a los patrones, requiriéndose un par de recálculos hasta estabilizar los mismos.

4.2 Sobre datos reales

En un caso real no puede generarse una tabla como la II. El criterio de "fin de tarea" fue cuando el número de datos erróneos detectados era nulo. Se define como erróneo a estos efectos, todo aquel registro que no coincide con el valor de planilla.

En primera instancia se trabajó para los días completos, sobre un conjunto de 21 estaciones, distinguiéndose dos fases.

En la primera, luego de realizado el cálculo de componentes principales, se procedió a retirar los errores más gruesos. Los mismos fueron identificados porque habiendo afectado levemente al vector de precipitaciones medias, ello se reflejaba en la forma de los primeros patrones (por más detalles, ver Silveira *et al.,* 1991). Ello corresponde a las etapas 1,2 y 3, de la Tabla III.

En una segunda fase del proceso se separaron aquellos días en que el valor absoluto de algún $a_i(\tau_k)$ excedía un margen especificado. Los datos de ese día eran contrastados con la planilla, y corregidas las discrepancias. A continuación se recalculaban los componentes y el proceso se reiniciaba.

Para cada a_i los márgenes se calcularon o bien como el triple de la desviación estándar, o bien no se especificaron. Si bien ese criterio "del triple" es derivado de una cota válida para la distribución normal, el método no requiere ni implica la misma.
En el transcurso de la tarea, se observó que algunos días eran sistemáticamente señalados como sospechosos, estando de acuerdo con la planilla. Se realizó un análisis subjetivo caso a caso, y se procedió a clasificar los datos en consistentes y dudosos. Los primeros se asociaban normalmente a precipitaciones intensas muy concentradas en el espacio; los segundos, muestran típicamente valores muy disímiles en estaciones próximas. Estos fueron retirados temporariamente de la base de datos, de forma de no afectar en el cálculo de los patrones (ver González *et al.*, 1991).
El proceso finaliza cuando los únicos días detectados contienen exclusivamente datos coincidentes con la planilla, y que ya fueron analizados. En la tabla III se muestra la evolución de la depuración durante el proceso. En las primeras tres etapas, se buscaban sólo errores muy gruesos, controlando esencialmente a los coeficientes de los patrones importantes; por ello el pequeño número de días afectados.
Un aspecto a comentar, es el relativo al evaluador del rendimiento. La columna η de las Tablas III y IV muestra un valor que, si bien es independiente del tamaño de la red, parece excesivamente pesimista; en la práctica, resulta más realista el indicado en la columna G (en términos de errores por día revisado), dado que en muchos casos el error era tan obvio, que con consultar sólo uno o dos de los 21 datos, era suficiente.

Etapa	A	B	C	E	F	G	η
1	9	21	6	51			34
2	354	326	154	448		186	87
3	222	267	336	475	395	174	83
4	70	83	206	286	219	126	60
5	72	60	8	132	105	106	51
6	41	2	29	111	18	65	31
7	9	1	12	115	13	19	9
8			1	113	2	1	0
9				109		0	0

Tabla III: Evolución de la depuración de los datos reales, para los días sin ausencias. Se analizan 21 estaciones. Claves: A.- Datos erróneos; B.- El dato no existe en planilla; C.- Datos dudosos; E.- Total de días revisados; F.- Días no tratados anteriormente; G.- (A+B+C)/E*100 Total de errores cada 100 días revisados; η = (A+B+C)/(21*E)*1000 Total de errores cada 1000 datos revisados.

Etapa	A	B	C	D	E	F	G	η
1	344	314	220	945	457		399	210
2	56	27	65	57	495	94	41	22
3	117	118	138	37	558	179	73	39
4	52	69	118	21	536	94	49	26
5	17	36	36	10	586	53	17	9
6	21	12	34	6	560	30	13	7
7	19	20	9	1	659	15	7	4
8					659		0	0

Tabla IV: Evolución de la depuración de datos reales, en aquellos días con alguna ausencia. Se analizan 19 estaciones. Claves: A.- Datos erróneos; B.- El dato no existe en planilla; C.-Datos dudosos; D.- Datos en planilla pero no digitados; E.- Total de días revisados; F.- Días no tratados anteriormente; G.- (A+B+C+D)/E*100 Total de errores cada 100 días revisados; η =(A+B+C+D)/(19*E)*1000 Total de errores cada 1000 datos revisados

Con respecto a los días con ausencias, se manejó el criterio de penalización de coeficientes principales (López *et al.* 1994). En aquellos días en que todos los valores registrados son cero, se completa con ceros los que falten. En otro caso, se penalizaron los 10 coeficientes más débiles con el inverso de la varianza de cada uno. La tabla IV muestra el avance de la tarea, estando los porcentajes referidos al total de datos revisados en cada etapa.

5. Conclusiones

Se ha presentado una metodología de análisis de calidad de datos, basada en el uso del Análisis de Componentes Principales (ACP). Los resultados, en términos de relación esfuerzo-beneficio son muy satisfactorios. En un experimento controlado, se logró identificar un dato erróneo cada dos días revisados, hallándose de esa forma más del 80% del total de los datos sembrados.

El esfuerzo de cálculo requerido es modesto, siendo lo más costoso el cálculo de la matriz de covarianza y los vectores propios de la misma, operación que se realiza un número muy limitado de veces.

Dado que para cada evento se realiza únicamente una transformación lineal, es posible aplicar el método en equipos de muy pequeño porte en tiempo real.

6. Reconocimientos

Carlos López tuvo a su cargo el diseño de la metodología, así como del experimento. Elizabeth González supervisó la depuración del banco de datos, y realizó el relevamiento bibliográfico. Ambos autores analizaron en conjunto los resultados obtenidos. Jorge Goyret implementó parte de los algoritmos, y realizó los cálculos relativos al experimento presentado.

7. Agradecimientos

Se deja constancia que este trabajo formó parte 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 publicar los resultados.

8. Referencias

Barnett, V.; Lewis, T., 1984. "Outliers in statistical data" John Wiley and Sons, 463 pp.

DNM, 1988. "Procedimientos para el control de calidad climatológico" Informe interno de la Dirección Nacional de Meteorología, Nov. 1988, 20 págs.

Fernau, M.E.; Samson, P.J., 1990. "Use of Cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part I: Transport Patterns" Journal of Applied Meteorology, V 29, N 8, 735-750.

Francis, P.E., 1986. "The use of numerical wind and wave models to provide areal and temporal extension to instrument calibration and validation of remotely sensed data" In Proceedings of A workshop on ERS-1 wind and wave calibration, Schliersee, FRG, 2-6 June, 1986 (ESA SP-262, Sept. 1986)

Gnanadesikan, R.; Kettenring, J.R., 1972. "Robust estimates, residuals and outlier detection with multiresponse data" Biometrics, V 28, 81-124.

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.

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.

Hollingsworth, A.; Shaw, D.B.; Lonnberg, P.; Illari, L.; Arpe, K. and Simmons, A.J., 1986. "Monitoring of observation and analysis quality by a data assimilation system" Monthly Weather Review, V 114, N 5, 861-879.

Husain, T., 1989. "Hydrologic uncertainty measure and network design" Water Resources Bulletin, V 25, N 3, 527-534.

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

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, J. F.; Curbelo, R., 1994. "Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias" Estadística, 46, 146, 147, pp. 55-83 También Publicación Técnica del Centro de Cálculo PTCECAL2/92, Centro de Cálculo, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

O'Hagan, A., 1990. "Outliers and credence for location parameter inference" Journal of the American Statistical Association: Theory and Methods, V 85, N 409, 172-176.

Pio, C.A.; Nunes, T.V.; Borrego, C.S.; Martins, J.G., 1989. "Assesment of air pollution sources in an industrial atmosphere using principal components and multilinear regression analysis" The Science of the Total Environment, V 8, 279-292.

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

Sevruk, B., 1982. "Methods of correction for systematic error in point precipitation measurement for operational use" World Meteorological Organization WMO 589, Operational Hydrology Report 21, 89 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., 1992. "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., 1992. "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.
White, D., 1991. "Climate regionalization and rotation of principal components" International Journal of Climatology, V 11, 1-25