

Pronóstico de la Precipitación Máxima en 24 Horas en la Zona Cafetera

Erasmó A. Rodríguez S.*, José A. Torres R.** , Pedro J. León Díaz**

RESUMEN

Se presenta un estudio para determinar los parámetros estadísticos y las características de los residuos en las series de precipitación máxima en 24 horas con un nivel de agregación temporal mensual, para estaciones localizadas sobre la cuenca del río Cauca, en los departamentos de Caldas, Quindío, Risaralda y la parte norte del departamento del Valle del Cauca. Las series de tiempo se ajustan a modelos de predicción ARIMA¹ cuyos resultados, basados en los criterios de menor error y pruebas estadísticas, conducen a obtener un residuo de tipo ruido blanco. Para validar el modelo de predicción, en las estaciones que cuentan con registros posteriores al período concurrente seleccionado, se realizan pruebas al residuo para verificar que corresponden a variables de tipo normal independiente. Finalmente, se hace una correlación cruzada entre registros de estaciones vecinas, o que se ajusten a un mismo modelo ARIMA con el fin de identificar la existencia de posibles relaciones estadísticas a escala regional.

ABSTRACT

The study presented here has been conducted with the aim of determine the statistical properties of maximum 24-hr rainfall monthly series. The investigation has considered rainfall stations located in the Provinces of Caldas, Quindío, Risaralda and the northern part of Valle del Cauca. Forecasting models like ARIMA have been adjusted and validated using the monthly series of 24-hr rainfall. Besides, analyses of cross correlation have been made in order to study the possible regional statistical relationships.

INTRODUCCIÓN

La importancia de conocer y poder pronosticar la precipitación máxima en 24 horas radica en la relación de la precipitación con eventos como movimientos de remoción en masa, inundaciones, avalanchas, etc., que pueden permitir adoptar medidas preventivas y de mitigación

para dichos eventos y también es de vital importancia para el cálculo de caudales de diseño en proyectos de obras civiles.

El objetivo del estudio es encontrar la variación de los errores de pronóstico y la variación de los parámetros estadísticos del residuo entre las series ajustada y registrada en un área con suficiente información.

La técnica empleada corresponde a un análisis estocástico de las series de tiempo en la zona cafetera colombiana, región que cuenta con buena cantidad y calidad de información.

La metodología empleada en el estudio es independizar las componentes determinísticas y estocásticas de las series de tiempo con el ánimo de ajustarlas al mejor modelo de predicción, considerando aquel que arroje los menores residuos entre el ajuste y los datos registrados, y satisfaga el mayor número de pruebas estadísticas de ajuste.

Para validar el modelo de predicción se realiza una comparación de los datos de predicción con los datos registrados para el mismo período, para las estaciones que cuentan con esta información. El modelo es aceptado si el residuo se comporta como ruido blanco.

El último paso es efectuar una correlación cruzada entre estaciones vecinas y series ajustadas a modelos ARIMA similares con el objetivo de conocer si existe algún tipo de correlación entre dichas estaciones.

Para llevar a cabo el estudio se cuenta con herramientas computacionales de versión académica tales como los programas STATLETS², STATS³ y STATGRAPHICS⁴, que son de gran ayuda, pero limitan el alcance del estudio pues permiten el manejo de series con un número máximo de 100 registros.

* Profesor Asistente Departamento de Ingeniería Civil, Unidad de Hidráulica.

**Estudiantes Maestría en Recursos Hidráulicos, Facultad de Ingeniería, Universidad Nacional de Colombia.

¹ ARIMA: Modelo autorregresivo integrado de promedio móvil.

² STATLETS es marca registrada de NWP Associates, Inc. 1997.

³ STATS es marca registrada de Decision Analyst Inc & Comunicometría S.C.

⁴ STATGRAPHICS es marca registrada de Statistical Graphics Inc., 1985-1993

I. METODOLOGÍA

A. SELECCIÓN DE LA INFORMACIÓN Y PERÍODO CONCURRENTE

Para el desarrollo del estudio se cuenta con registros de 23 estaciones climatológicas, ubicadas en la zona cafetera que tienen un área de cubrimiento aproximada de 4.300 km² y un nivel de agregación temporal mensual. Véase figura 1.

Para la selección de la información de las diferentes estaciones se han descartado aquellas estaciones que presenten registros incompletos para períodos de más de 1 año, y aquellas que no cuentan con registros dentro del período concurrente seleccionado.



Figura 1. Localización de estaciones

Debido a que una de las herramientas disponibles para el tratamiento de los datos es el programa STATLETS, cuya versión académica permite el manejo de máximo 100 registros, se ha utilizado un período de 8 años, que corresponde al lapso entre 1981 y 1988.

Cuadro 1. Información de estaciones meteorológicas

ESTACION	CODIGO	LATITUD grad-min	LONGITUD grad-min	COTA msnm
VIRGINIA-LA ALERTA	4614018	4-54	75-53	900
APTO MATECAÑA	4613504	4-49	75-44	1342
SAN ISIDRO	4613018	4-48	75-51	1220
LA BOHEMIA	4612510	4-52	75-54	1020
APTO EL EDEN	4612506	4-27	75-46	1204
LA CAMELIA	4612012	4-19	75-49	1245
EL ALAMBRADO	4612018	4-24	75-49	1100
CUMBARCO	4612513	4-12	75-49	1740
LA ITALIA	4610079	4-02	75-46	2740
HDA LA LUCERNA	4610007	4-13	76-09	942
GALICIA	4610041	4-11	76-03	1174
ZARZAL	4610078	4-24	76-04	916
EL ALCAZAR	4610040	4-19	76-02	963
COROZAL	4612013	4-24	75-55	1158
SAN PEDRO	4610074	4-32	76-02	909
OBANDO	4610030	4-36	75-59	960
PALMA SOLA	4610083	4-42	75-58	910
ALCALA	4612015	4-40	75-48	1250
SUB LA ROSA	4613012	4-50	75-42	1440

B. ANÁLISIS DE LAS SERIES DE TIEMPO

El análisis de series de tiempo consiste en realizar una revisión del comportamiento estadístico de las series, para determinar la periodicidad y/o dependencia de las misma, si existe o no, un componente aleatorio en ellas y determinar la importancia de cada componente. Para dicho análisis se emplea la prueba de bondad de ajuste Chi cuadrado, el análisis de autocorrelación y el análisis espectral.

1. PRUEBA DE ESTACIONARIDAD

La prueba de estacionaridad se aplica a todas las series de tiempo de las estaciones seleccionadas, con el objetivo de determinar si la serie contiene un componente aleatorio o no. Con base en los resultados del programa STATS se han determinado los valores de la prueba Chi cuadrado y la probabilidad de que los valores esperados sean diferentes a los valores registrados.

Se pudo determinar que las series no son estacionarias debido a que los valores de probabilidad obtenidos están por fuera de los límites de confianza del 95 %, por lo tanto las series tienen un componente aleatorio.

2. ANÁLISIS DE AUTOCORRELACIÓN

El análisis de correlación se realiza para una serie circular discreta con la finalidad de determinar dependencia entre los registros.

A manera de ejemplo se presenta el correlograma para la estación Virginia – La Alerta 4614018 (figura 2), en donde se observa un comportamiento no completamente aleatorio con una cierta periodicidad, por lo cual es necesario realizar otros tipos de análisis, tal como se describen a continuación.

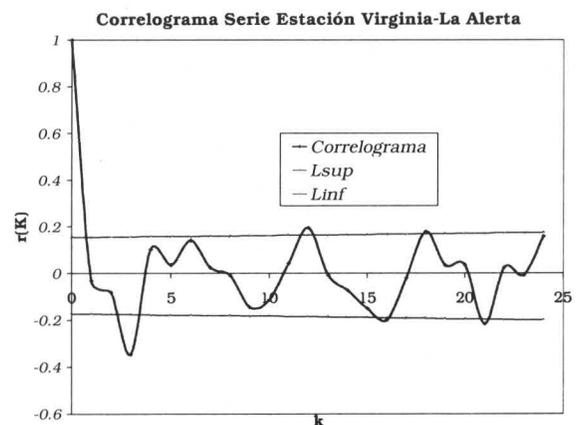


Figura 2. Función de autocorrelación estación Virginia – La Alerta

3. ANÁLISIS ESPECTRAL

El análisis espectral discreto se realiza para verificar la periodicidad e identificar las frecuencias que representan en gran parte la varianza de la serie analizada. Para la estación Virginia

– La Alerta 4614018 se observa que la frecuencia 1/6 correspondiente a una periodicidad de 6 meses; es la más relevante (ver Figura 3), según este análisis se puede afirmar que por el régimen bimodal de las lluvias, se presenta un período semestral húmedo.

También se realiza el periodograma acumulado para verificar la importancia de cada uno de los componentes de la serie, encontrándose que el componente aleatorio es importante y que la serie es completamente aleatoria con un 95 % de confianza.

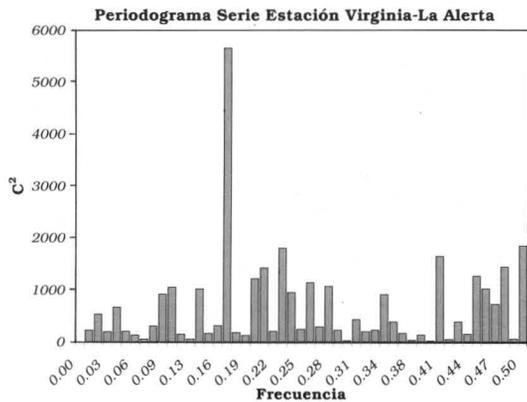


Figura 3. Periodograma estación Virginia – La Alerta

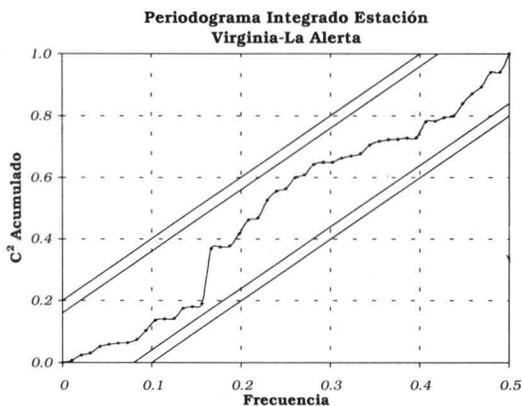


Figura 4. Periodograma integrado estación Virginia – La Alerta

4. ESTACIONALIDAD

Se define estacionalidad como el movimiento periódico de una serie, más o menos regular en donde se conoce a priori la frecuencia, tal como el ciclo anual.

El objetivo de este análisis, es confirmar la ciclicidad. Existen dos modelos típicos de estacionalidad: un modelo multiplicativo de la forma:

$$Y_t = T_t * C_t * S_t * R_t \tag{1}$$

y un modelo aditivo de la forma

$$Y = T_t + C_t + S_t + R_t \tag{2}$$

Donde Y es la serie de tiempo, T es la tendencia que siguen los datos a largo plazo, C es la ciclicidad que siguen los datos alrededor de la tendencia, S es la estacionalidad o periodicidad regular y R corresponde al ruido blanco o fluctuaciones impredecibles en los datos.

El enfoque general para obtener una descomposición sigue varios pasos:

- Suavizar los datos por medio de un modelo de promedio móvil de longitud de la estacionalidad o periodicidad conocida. Por ejemplo, un promedio móvil de longitud 12 se aplicaría a datos mensuales. Esto estima la tendencia cíclica.
- Dividir los datos por el promedio móvil (si se usa el método multiplicativo) o restarlo de los datos (si se usa el método aditivo). Esto estima la estacionalidad.
- Promediar los resultados para cada estación separadamente y ajustar los datos para que un mes promedio equivalga a 100 (si se usa el método multiplicativo) o 0 (si se usa el método aditivo). Así se obtienen los índices estacionales.
- Ajustar los datos para la tendencia estimada, el ciclo y estacionalidad, restando el componente residual.
- Dividir los valores originales de datos por el índice estacional apropiado usando el método multiplicativo, o restando el índice usando el método aditivo.

Como se observa en el periodograma en la figura 3 existe una tendencia cíclica de la serie que puede verificar la existencia de estacionalidad en la serie.

C. AJUSTE DEL MODELO DE PREDICCIÓN

1. TIPOS DE MODELOS DE PREDICCIÓN

Existen diferentes modelos de predicción como son los modelos de corrida aleatoria, media constante, tendencia lineal, tendencia cuadrática, tendencia exponencial, promedio móvil y ARIMA entre otros. A continuación se realiza una breve descripción de los métodos utilizados en este estudio.

- **Corrida aleatoria:** este modelo asume que en la serie de tiempo es igualmente probable para corridas por arriba o abajo desde su posición actual al tiempo t. El mejor pronóstico para el t+k es el valor de la serie al tiempo t, es decir,

$$F_{t+k} = Y_t \tag{3}$$

- **Media constante:** Este modelo asume que los datos varían aleatoriamente alrededor una media fija. El mejor pronóstico para el tiempo t+k es el promedio de todos los valores previos, es decir,

$$F_{t+k} = \text{promedio de } (Y_1, Y_2, \dots, Y_t) \quad (4)$$

• **Tendencia lineal:** este modelo supone que los datos varían aleatoriamente alrededor de una línea de pendiente b Con intercepto a : El mejor pronóstico por el tiempo $t+k$ es una extrapolación de los datos por mínimos cuadrados, es decir,

$$F_{t+k} = a + b*(t+k) \quad (5)$$

• **Tendencia Cuadrática:** El modelo asume que los datos varían aleatoriamente alrededor una regresión cuadrática. El mejor pronóstico para el tiempo $t+k$ es una extrapolación de los datos por mínimos cuadrados, es decir,

$$F_{t+k} = a + b*(t+k) + c*(t+k)^2 \quad (6)$$

• **Tendencia Exponencial:** El modelo asume que los datos varían alrededor de una regresión exponencial. El mejor pronóstico para el tiempo $t+k$ es una extrapolación de los datos a través de la siguiente curva.

$$F_{t+k} = \exp(a + b*(t+k)) \quad (7)$$

• **Promedio móvil simple:** El modelo pronostica datos futuros a partir de un promedio móvil de ventana m .

$$F_{t+k} = (Y_{t-m} + Y_{t-m-1} + \dots + Y_t) / m \quad (8)$$

• **Suavizado exponencial simple:** este modelo pronostica valores futuros tomando un promedio ponderado de todos los datos, donde se da más peso a las observaciones recientes que a observaciones antiguas. El pronóstico para el tiempo $t+k$ está dado por:

$$F_{t+k} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \alpha(1-\alpha)^3 Y_{t-3} + \dots \quad (9)$$

Donde α es una constante suavizadora que debe estar entre 0 y 1. Cuando el valor de α es cercano a 0, el mayor peso se da a observaciones antiguas. Cuando α tiende a 1, el modelo tiende a la media. Cuando el modelo tiende a 1, el modelo aproxima una corrida aleatoria.

• **Modelos ARIMA:** son modelos autorregresivos integrados de promedio móvil que expresan los datos para el tiempo t como una combinación lineal de datos previos.

Estos modelos son, en teoría, la clase más general de modelos para pronosticar una serie de tiempo y se pueden identificar como (p,d,q) .

Donde:

p = orden del modelo autorregresivo

d = orden de diferenciación

q = orden del modelo promedio móvil

Para datos estacionales el modelo ARIMA se puede expresar como $(p,d,q) \times (P,D,Q)$ donde los valores P,D,Q son los ordenes

autorregresivo, de diferenciación y del promedio móvil respectivamente del componente estacional.

La forma general del modelo ARIMA es:

$$\Phi_p(B) \nabla^d X_t = \Theta_q(B) \varepsilon_t \quad (10)$$

Donde $\Phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)$ es el polinomio autorregresivo de orden p , $\Theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)$ es el polinomio del promedio móvil de orden q , ∇ es el operador diferencial, y ε_t es una secuencia de ruido distribuido normal independiente con media 0 y varianza constante.

2. CRITERIO DE SELECCIÓN

Cada una de las series se ajusta a diferentes modelos de los cuales se selecciona solamente uno basado en los errores de ajuste de la predicción, los cuales son la diferencia entre los datos reales para el tiempo t y el pronóstico. Los errores que presenta el programa STATLETS son los siguientes:

MSE: El promedio o media del cuadrado de los errores.

RMSE: La raíz cuadrada del MSE.

MAE: La media de los valores absolutos de errores de pronóstico.

MAPE: La media de los valores absolutos de los errores, como un porcentaje del valor real.

ME: El promedio o la media de los errores.

MPE: La media del error como un porcentaje de los valores reales. Este es calculado únicamente si todos los valores de datos son mayores que 0.

Los MSE, RMSE, MAE, MAPE, ME y MPE miden la magnitud de los errores de pronóstico. Los mejores modelos obtienen valores menores para estas estadísticas.

Otra forma para determinar el ajuste del modelo se realiza mediante pruebas estadísticas que determinan si el modelo se adapta a los datos originales y si el residuo de los datos corresponde a ruido blanco. Las pruebas que ejecuta el programa STATLETS son:

CORRIDAS: Cuenta el número de corridas arriba y abajo.

RUNM: Cuenta el número de corridas arriba y abajo de la mediana.

AUTO: Ejecuta la prueba de Box-Pierce sobre la correlación del residuo.

MEDIA: Compara la media de la primera mitad de los residuos con la media de la segunda la mitad.

VAR: Compara la varianza de la primera mitad de los residuos con la varianza de la segunda mitad

Si las pruebas son aceptadas se presenta un mensaje OK - Cuando la prueba no es significativa y donde la probabilidad es mayor al 10 %.

La selección del modelo depende de los parámetros observados cuando el modelo pasa todas las pruebas.

3. METODOLOGÍA BOX - JENKINS PARA EL ANÁLISIS DE SERIES DE TIEMPO

La metodología más utilizada en la predicción de series de tiempo es la de Box-Jenkins (1976), que procura ajustar el modelo ARIMA utilizando como criterio de selección el residuo y sus parámetros estadísticos, además involucra la estacionalidad de las series de tiempo. En la figura 5 se muestra un diagrama de flujo de la metodología de Box-Jenkins.

Para la estación Virginia - La Alerta se presenta la predicción de un periodo de tres años ajustado al modelo ARIMA que cumple todas las pruebas estadísticas y con el menor error de ajuste, Véase figura 6.

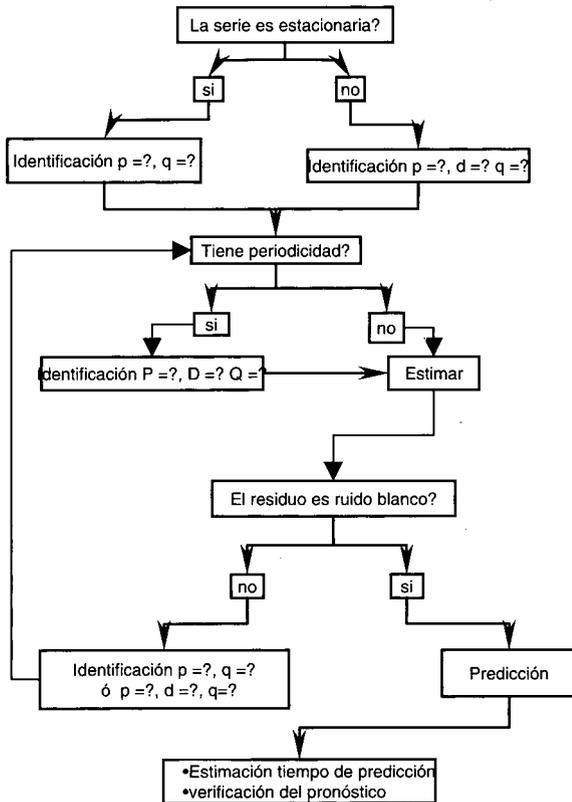


Figura 5 Metodología de Box-Jenkins

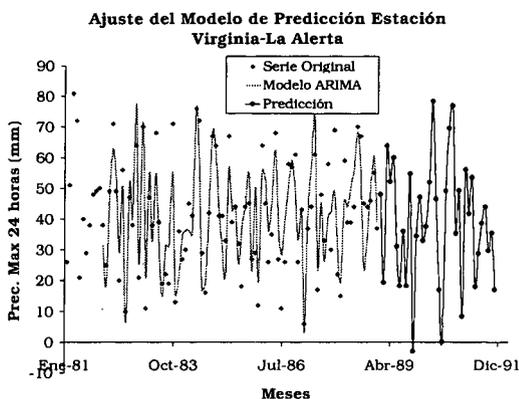


Figura 6. Predicción de registros Estación Virginia – La Alerta.

D. VERIFICACIÓN DE LA PREDICCIÓN

1. ANÁLISIS DEL RESIDUO

Una serie aleatoria de tiempo es una sucesión de números aleatorios sin la estructura dinámica inherente o correlación. Si el modelo de pronóstico ha capturado toda la estructura dinámica en los datos, los residuos deberían formar una sucesión aleatoria de números.

STATLETS ejecuta tres pruebas para verificar si el residuo corresponde a ruido blanco:

- **Las corridas arriba y abajo de la mediana:** Cuenta el número de veces que la serie residual está arriba o abajo de la mediana de muestreo y compara el valor esperado para una sucesión aleatoria de números.
- **Corridas por arriba y abajo:** Cuenta el número de veces que la serie residual de tiempo sube o baja, y la compara con el valor esperado para una sucesión aleatoria de números.
- **Prueba de Box Pierce:** Determina el primer término de la autocorrelación y calcula un estadístico que sigue una distribución Chi cuadrado.

Cada prueba calcula una probabilidad. Una prueba con probabilidad por debajo de 0,05 conduciría a rechazar la hipótesis de que los residuos son aleatorios al 5% el nivel de significancia.

Para comprobar la veracidad de la predicción en la estación Virginia - La Alerta, se ha calculado la diferencia entre los datos reales y la predicción en los años 1989, 1990 y 1991, y se ha realizado el mismo tratamiento para los residuos con el fin de determinar si corresponden a ruido blanco. (Ver figura 7)

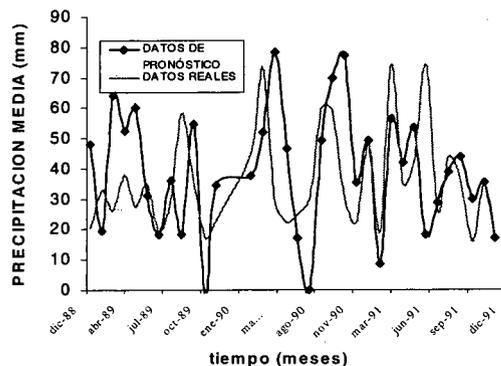


Figura 7. Comparación datos reales - Predicción Estación Virginia La Alerta.

E. CORRELACIÓN CRUZADA

El correlograma de la serie se elabora para procesos de variables mutuamente independientes a partir de la siguiente ecuación, que corresponde al correlograma de una serie discreta:

$$r(k) = \frac{\frac{1}{N-k} \sum_{i=1}^{N-k} (x_i)(y_{i+k}) - \frac{1}{(N)^2} \left(\sum_{i=1}^N x_i \right) \left(\sum_{i=1}^N y_i \right)}{\left[\frac{1}{N} \sum_{i=1}^N x_i^2 - \mu x^2 \right] \left[\frac{1}{N} \sum_{i=1}^N y_i^2 - \mu y^2 \right]^{\frac{1}{2}}}$$

Donde N es el número de datos de la muestra, k es el intervalo de tiempo para una serie discreta y μ corresponde a la media de cada serie de tiempo.

En las Figuras 8 y 9, que corresponden a correlogramas cruzados de estaciones vecinas, se observa cierta correlación entre los datos, además de periodicidad común de 6 y 12 meses.

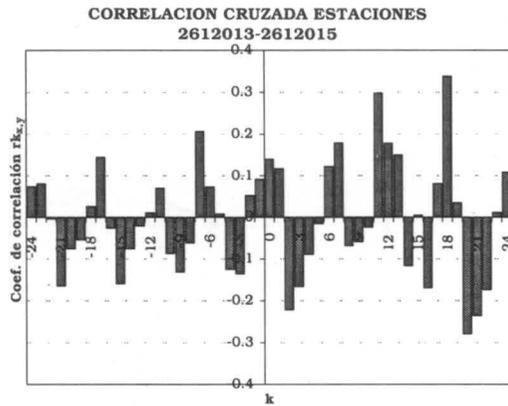


Figura 8 Correlación cruzada entre las estaciones vecinas 2612013-2612015

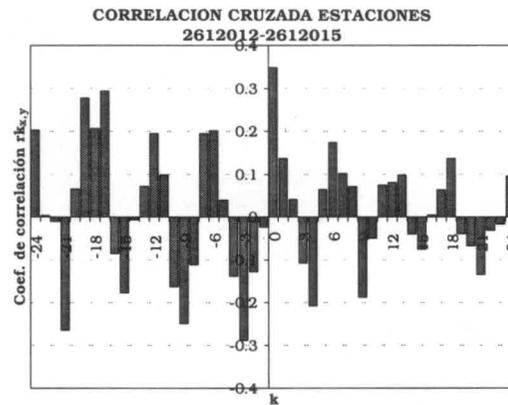


Figura 9. Correlación cruzada entre las estaciones vecinas 2612012-2612015

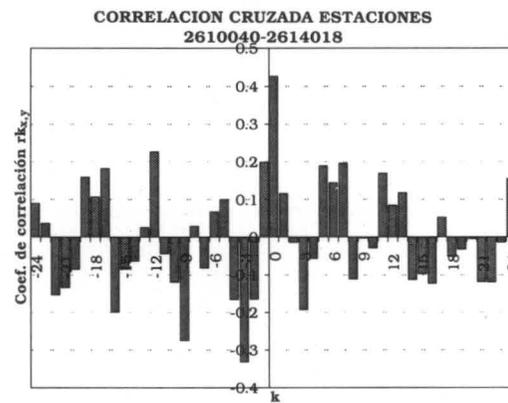


Figura 10. Correlación cruzada entre las estaciones 2610040-2614018 que se ajustan al mismo modelo ARIMA

En la figura 10 se observa correlación pero menor que la observada entre estaciones vecinas, sin embargo, para los intervalos cercanos a cero los valores son significativos.

II. ANÁLISIS DE RESULTADOS Y CONCLUSIONES

En el cuadro 2 se presenta un resumen de los modelos ARIMA que se ajustan mejor a la serie original de los datos para cada una de las estaciones estudiadas, se observa también el error de ajuste (RMSE), y el parámetro estadístico de la desviación estándar del residuo.

Cuadro 2. Resumen modelo ARIMA de mejor ajuste

ESTACION	ERROR DE AJUSTE (RMSE) mm	DESVIACION RESIDUO mm	modelo ARIMA					
			p	d	q	P	D	Q
VIRGINIA-LA ALERTA	9,959	13,756	1	0	1	1	1	6
APTO MATECAÑA	11,638	16,791	1	0	1	1	1	6
SAN ISIDRO	11,395	13,397	1	0	3	2	1	4
LA BOHEMIA	9,765	14,212	1	0	5	0	1	6
APTO EL EDEN	11,630	13,556	0	0	4	2	1	5
LA CAMELIA	13,582	17,404	1	0	1	2	1	5
EL ALAMBRADO	11,410	14,759	0	1	1	2	1	4
CUMBARCO	10,242	13,731	1	0	2	1	1	6
LA ITALIA	14,241	14,979	0	0	5	2	1	4
HDA LA LUCERNA	9,774	14,488	1	0	1	2	1	6
GALICIA	9,673	13,623	1	0	2	1	1	6
ZARZAL	9,817	13,583	2	0	4	1	1	5
EL ALCAZAR	12,297	16,801	1	0	1	1	1	6
COROZAL	10,923	14,619	0	0	1	0	1	6
SAN PEDRO	11,639	12,606	0	0	6	2	1	4
OBANDO	13,303	15,910	1	0	6	1	1	5
PALMA SOLA	11,789	13,746	1	0	4	2	1	4
ALCALA	12,572	15,298	0	0	2	2	1	6
SUB LA ROSA	14,740	20,005	0	1	5	2	1	6
Promedio:	11,6 mm	14,91 mm						

Los errores de ajuste (RMSE) de las estaciones analizadas tienen un promedio de 11,6 mm con valores que oscilan entre 9,67 y 14,74 mm.

Las desviaciones estándar de los residuos del ajuste en las estaciones analizadas tienen un promedio de 14,91 mm y los valores están entre 12,6 y 20,0 mm

- El régimen de lluvias bimodal en la zona de estudio obedece al movimiento del ecuador térmico (Zona de convergencia intertropical), y al movimiento de traslación de la tierra. Dichos fenómenos fijan la estacionalidad de las series analizadas.
- Las series de precipitación máxima mensual en 24 horas para el período concurrente 1981-1988, en las estaciones meteorológicas analizadas, no son estacionarias y tampoco ergódicas.
- Los mejores modelos de ajuste para predicción de precipitación de la zona, son los modelos ARIMA con la estacionalidad encontrada con ayuda del análisis de correlación y análisis espectral.

- Los polinomios de promedio móvil de los modelos ARIMA estacionales se ajustan mejor, en general, a tamaños de ventana altos (verificados hasta 6), posiblemente por ser los tamaños de ventana del período estacional.
- Todos los residuos del ajuste respecto a la serie original se comportan como ruido blanco y en la verificación de la predicción se llega a la misma conclusión, por lo tanto se puede afirmar que los modelos utilizados en dicha predicción son válidos.
- Los mejores coeficientes de correlación cruzada se presentan entre estaciones vecinas y describen una tendencia cíclica regular de período 12 meses.
- El resultado de la correlación cruzada, indica que para esta región es recomendable completar los datos faltantes de los registros con ayuda de estaciones cercanas

BIBLIOGRAFÍA

1. BRAS R., RODRÍGUEZ-ITURBE I. "*Random Functions and Hydrology*". Dover Publications, Inc. E.U. (1993).
2. CHIN, D.A. "A Scale Model of Multivariate Rainfall Time Series". *Journal of Hydrology* (Amsterdam). 168 (1): 1-16. (19__)
3. CHOW, V.T. "*Hidrología Aplicada*". Mc Graw Hill. Bogotá.(1994).
4. FEDDES, R. "*Space And Time Scale Variability And Interdependencies In Hydrological Processes*". Unesco. (1995).