Generación de datos faltantes de lluvia anual usando programación genética considerando sitios cercanos

Margarita Preciado Jiménez, Maritza Liliana Arganis Juárez, Ma. De los Ángeles Suarez Medina
Instituto Mexicano de Tecnología del Agua, Morelos, México, preciado@tlaloc.imta.mx, msuarez@tlaloc.imta.mx
Instituto de Ingeniería, Ciudad de México, México, MArganisJ@iingen.unam.mx

Resumen
En este trabajo se realizó la interpolación y extrapolación de datos de precipitación con modelos obtenidos con ayuda de un algoritmo aleatorio de programación genética para aumentar la longitud del registro de series de precipitaciones diarias máximas anuales de estaciones cercanas entre sí, de manera que la media y otros estadísticos del registro se logren conservar aproximadamente.

Introducción

Las técnicas de rellenado de datos para análisis hidrológicos de estudios de ingeniería son herramientas muy importantes, ya que estos permiten contar con bases de datos homogéneas y que pueden ser utilizadas de manera confiable; se han llevado a cabo diversos estudios en esta materia, utilizando técnicas tradicionales de regresiones lineales múltiples y métodos de optimización los cuales se basan en algoritmos aleatorios para la obtención de los parámetros de los modelos (Arganis et al., 2009, Campos-Aranda, 2015). En las estaciones climatológicas es muy frecuente contar con registros incompletos generando a su vez bases de datos porosas, por lo que a su vez se obtienen resultados dudosos y poco consistentes durante la utilización de los mismos. Para llevar a cabo un estudio hidrológico que permita conocer de los elementos meteorológicos, su variabilidad, frecuencia y probabilidad de valores críticos y relacionarlos con la exigencias del proyecto se requiere que los registros históricos sean continuos y con coherencia, de esta manera se minimicen los riesgo de resultados erróneos con el fin de contar con datos precisos de lluvia escurrimiento para evaluar y pronosticar, permitiendo así, llevar a cabo a nivel de detalle minimizando los errores y no sesgar los resultados (Massetti, 2013).

En la literatura se proponen diferentes métodos estadísticos para el llenado de datos faltantes, cabe mencionar la guía de prácticas climatológicas de la Organización Meteorológica Mundial (WMO, 1983), el cual propone los siguientes métodos estadísticos para el relleno de datos faltantes como son regresión simple, múltiple razón, q y razón-normal q (RN). ASCE (1996), propone el método de ponderación de distancia inversa IDW, también conocido como el método U.S National Weather Servic por su implementación en estudios hidrológicos y geográficos al igual que estudios realizados por Hubbard (1994); Sokol y Stekl (1994); Palomino y Martin (1995); Teegavarapu y Chandramouli (2005). Aparicio (2011) y Campos (1998) indican que este método puede emplearse cuando se basa en registros simultáneos de tres estaciones que se encuentren lo más cerca posible a la estación en estudio. Una ventaja importante de la IDW es que se utiliza en cualquier paso de tiempo (Teegavarapu y Chandramouli, 2005). Existen otros métodos estadísticos entre ellos el análisis de componentes principales y el análisis de grupo (Huth y Nemesova, 1995), el método de Kriging (Saborowski y Stock, 1994) y la interpolación óptima (Bussieres y Hogg,1989). Wagner et al. (2012) sugieren que este no reporta mejorías con respeto a el inverso de la distancia (IDW) y que para su empleo requieren una cantidad suficiente de datos para producir un semivariograma fiable.

Campos (1998) propone el método empírico racional deductivo cuando no se disponen estaciones cercanas y se tiene más de 10 años de registros de la estación de interés, empleado por Puertas et al. (2011) y Guevara (2003) para estimar los datos faltantes de precipitación. Jiménez et al. (2004) utilizaron el método regresión lineal con las siguientes condiciones: distancia menor de 25 km, altitud de ± 30 m y con el mismo tipo de clima, y cuando no se cumplían utilizó el método Racional Deductivo. La finalidad de este trabajo fue determinar la confiablidad de los métodos de relleno: U.S National Weather Servicie, deductivo racional, regresión múltiple y simple y utilizar el mejor para rellenar los datos faltantes de las series: precipitación, temperatura máxima, temperatura mínima de las estaciones dentro de la zona de estudio. En este trabajo se presenta la programación genética como una herramienta con la cual se puede contar para el llenado y generación de fatos faltantes.

Localización de zona de estudio

La laguna de Catemaco es el tercer cuerpo acuífero en extensión de México; sobre su ribera occidental se encuentra asentada la ciudad de Catemaco, perteneciente al estado de Veracruz, este cuerpo de agua se forma a partir de escurrimientos por lluvias torrenciales, propias del clima de bosque tropical que lo rodea, así como por decenas de arroyos, y varios ríos, incluyendo río Cuetzalapan.

Método

Programación genética (PG)

La programación genética (PG) es una herramienta de optimización del cómputo evolutivo que permite construir modelos matemáticos en los que una variable dependiente se puede relacionar con una o más variables independientes, también con este tipo de algoritmos se pueden generar programas de cómputo. Su aplicación es similar a la de un algoritmo genético simple debido a que consiste en generar una población inicial de n individuos, donde cada individuo representa un modelo matemático que se probará en una función objetivo que se desee optimizar. Los individuos están formados por operadores matemáticos (pueden ser exclusivamente aritméticos de suma, resta, multiplicación, división o se pueden añadir funciones trascendentes del tipo seno, coseno, exponencial o cualquier otro tipo de operador como puede ser el operador valor absoluto), también en los individuos pueden aparecer términos constantes. Los mejores individuos se seleccionan un cierto número de veces y posteriormente se realizan las operaciones de intercambio o cruza (llamado cruce también), que consiste en considerar partes o nodos formados por operadores matemáticos y constantes de un individuo e intercambiarlos con los de otro individuo lo cual genera nuevos modelos matemáticos teniéndose al final el mismo tamaño de población que la inicial y estos nuevos modelos pasan a la siguiente generación (iteración) y se vuelven a probar en la función objetivo; el proceso se repite hasta completar el número de generaciones (iteraciones) consideradas para terminar el proceso y el modelo matemático que reporta el mejor resultado para la función objetivo en la última generación representa la solución del problema.

Datos de entrada al modelo de programación genética

En el problema analizado se consideran como variables independientes la precipitación medida en dos estaciones climatológicas cercanas a una cuenca y como variable dependiente otra estación climatológica de la misma cuenca en la que se observa un periodo común de observación entre los tres sitios. Con el modelo matemático obtenido y, al contar con más años de registro en las dos estaciones cercanas, se completa el registro de la tercera estación (extrapolando en años anteriores o posteriores) con la finalidad de

aumentar la longitud del periodo común de los tres registros. Cuando se presenta un año en el que no hay registro en dos estaciones para obtener la tercera el rellenado de ese año se realiza con la media histórica en esa estación, y cada vez que se tengan dos estaciones con dato se selecciona el periodo común de esas estaciones para obtener un nuevo modelo de PG con el que se completa el dato de la tercera estación de manera que cada vez se va teniendo un periodo común más amplio; hasta lograr series de igual longitud que se han rellenado con modelos de interpolación en su mayoría y en puntos aislados con la media de la serie histórica. En cada nuevo registro se hace una reestimación de los estadísticos de cada estación y se calculan las diferencias respecto a los estadísticos históricos para observar su variación.

Análisis estadístico de series máximas anuales

El análisis estadístico de un registro de datos máximos anuales se realiza con los siguientes pasos: 1) se ordenan los datos de mayor a menor 2) Se les asigna un periodo de retorno con una función empírica (la ecuación de Weibull es comúnmente utilizada) 3) Se determinan los parámetros de distintas funciones de distribución usando una técnica que pueden ser las de momentos o la de máxima verosimilitud o algún método de optimización 4) Se calculan los valores que reporta la función para el periodo de retorno de cada dato y se obtiene el error estándar de ajuste. 5) Se selecciona la función que tenga el menor error estándar de ajuste 6) La función seleccionada se utiliza para obtener eventos de diseño para distintos periodos de retorno (comúnmente las obras en hidrología, dependiendo de la importancia y tamaño de las obras se diseñan para periodos de retorno de 2,5,10,50,100,1000,5000 y 10000 años).

Aplicación y resultados

La metodología se aplicó a datos de precipitación diaria máxima anual de tres estaciones climatológicas localizadas en la cuenca de la presa Canseco, en el municipio de Catemaco, Veracruz, a saber: 30204 Catemaco (CFE), 30033 Coyame, 30294 Sontecomapan. Los datos se obtuvieron de las normales climatológicas reportadas por la CONAGUA y de datos reportados por la Comisión Federal de Electricidad (CFE). Los registros originales aparecen en la Tabla 1 y en las Figuras 2 a 4; sus estadísticos aparecen en la Tabla 2.

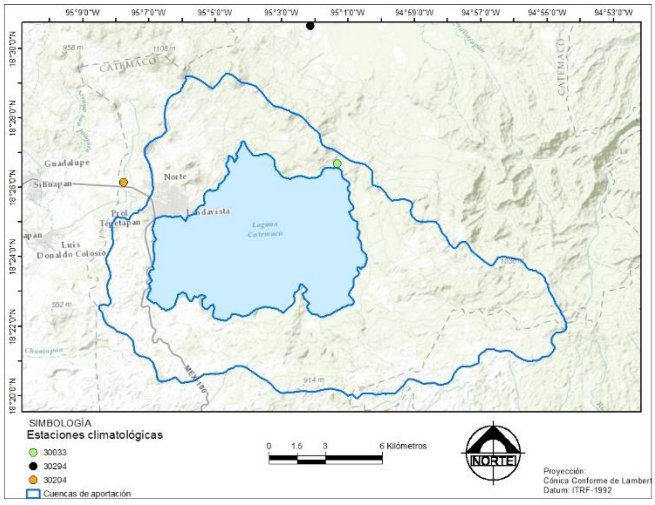

Figura 2. Sitio de estudio. Estaciones climatológicas seleccionadas en la cuenca de la Presa Canseco, Ver.

Tabla 1. Precipitación diaria máxima anual. Estaciones climatológicas. Cuenca Presa Canseco, Ver. Registros originales

Año	Catemaco CFE 30204 hp, mm	Coyame 30033 hp, mm	Sontecomapan 30294 hp, mm	Año	Catemaco CFE 30204 hp, mm	Coyame 30033 hp, mm	Sontecomapan 30294 hp, mm
1952	175.5	222.5		1984	135.5	268.4	229.4
1953	224	241.8		1985	59	123.8	
1954	100	143		1986	130	333.9	310.9
1955	283	300		1987	137.5	382.4	253.9
1956	93.5	249.6		1988	180	379.6	181.9
1957	112	320.5		1989	191.5	265.7	274.5
1958	107	300.5		1990	204.5	257.8	183
1959	151.3	265		1991	216.5	415.8	382.3
1960	107	303		1992	203.5	183.7	308
1961	116	276		1993	280	214.8	205.8
1962	97.5	320		1994	40	184.3	140
1963	85	148		1995	130.5	200.7	153.9
1964	98	244.3		1996	234.5	230.3	193.2
1965	133.4	266.5		1997	51	366.5	305.4
1966	198.4	226.3		1998		214.8	158.6
1967	300.5	320		1999		251.4	247.5
1968	244.5	328		2000		191.6	364.6
1969	175	296.5		2001		301.8	337.2
1970	137.2	221		2002		225.9	144.3
1971	218	270		2003		208.6	
1972	200	336		2004		341.9	309.6
1973	297.7	336		2005		359.9	151.2
1974	164.5	209		2006		224.9	326.2
1975	126.4	166.3		2007		187.6	186.2
1976	186	268.5	191.5	2008		419.6	214.7
1977	168.3	336	255.5	2009		268.7	196.8
1978	171	220.4	237.3	2010		191.8	142.8
1979	140	157	149.4	2011		156.8	203.4
1980	173.9	229	203.5	2012		282.9	247.6
1981	140	263.8	281.1	2013			203.2
1982	189	266.8	178.9	2014			
1983	83	157.5	201.6	2015	107.2	212.6	238.6

Tabla 2. Estadístico originales. Estaciones climatológicas

Estadístico	Catemaco CFE 30204	Coyame 30033	Sontecomapan 30294
media	159.53	258.99	229.55
desvest	63.39	69.17	65.96
coef asim	0.41	0.28	0.61
cv	0.40	0.27	0.29

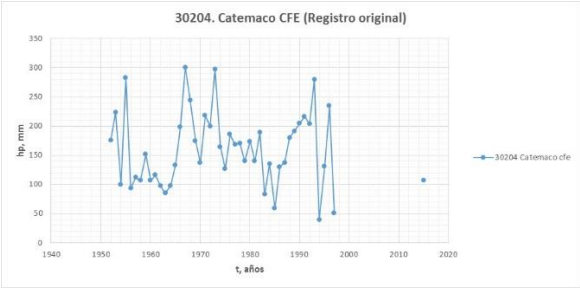

Figura 3. Registros originales de precipitación diaria máxima anual. 30204 Catemaco CFE

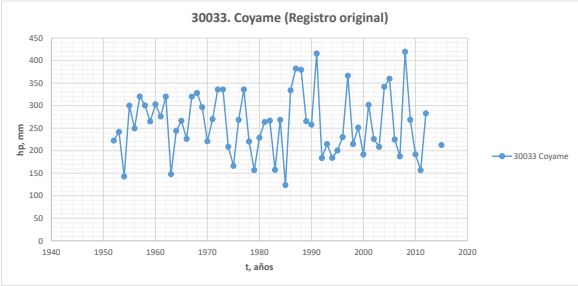

Figura 4. Registros originales de precipitación diaria máxima anual. 30033 Coyame

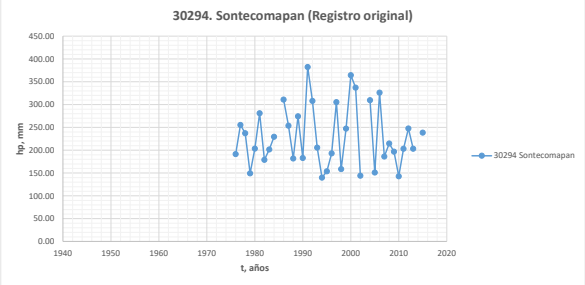

Figura 5. Registros originales de precipitación diaria máxima anual. 30294 Sontecomapan

De la Tabla 1 se identificó el periodo común de 1976 a 1984 en las tres estaciones y se puso como variable dependiente a la estación que le faltaba el dato de 1985, que fue Sontecomapan, y como variables independientes se consideraron los datos de Catemaco CFE y de Coyame, aunque los coeficientes de variación son más parecidos entre Sontecomoapan y Coyame. Con esos datos se construyó el archivo para alimentar al algoritmo de programación genética (PG) que se encuentra codificado en Matlab (Mathworks, 2005),programa original del Instituto en Investigaciones en Matemáticas Aplicadas y en Sistemas. Para la programación genética se consideraron operadores de

suma, resta y multiplicación, con el fin de tener formas sencillas de los modelos. La ecuación PG1 obtenida para este primer intervalo de datos es:

$$hpsonte=-0.891951*hpcatemaco+2.297248*hpcoyame-0.00332149*hpcoyame^2+1.231278 \quad (1)$$

dónde: *hpsonte* precipitación diaria máxima anual en sontecomapan, en mm, *hpcatemaco* precipitación diaria máxima anual en Catemaco CFE, en mm, *hpcoyame* precipitación diaria máxima anual en Coyame, en mm

El error medio cuadrático obtenido con esa ecuación fue de 566.91. Al dibujar los datos medidos contra los calculados se obtuvo un coeficiente de determinación R^2= 0.6137, un poco bajo, se notó una ligera dispersión de los datos respecto a la función identidad (Figura 5).

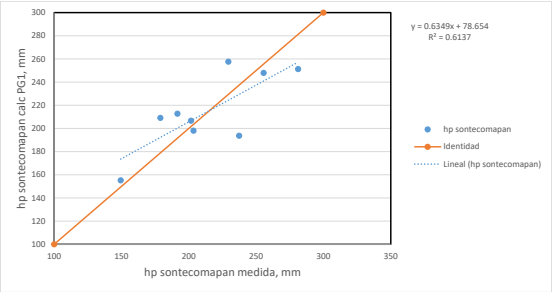

Figura 6. Comparación de datos medidos vs calculados con PG1 vs la identidad. Sontecomapan 1976 a 1984

Con la ecuación (1) denominada en las figuras como PG1 se obtuvo el dato de 1985 de la estación Sontecomapan y con ello se logró obtener un periodo común de registro en las tres estaciones desde 1976 a 1997, para este periodo de registro se obtuvo un nuevo modelo llamado PG2 :

$$hpsonte=0.457559*hpcoyame+108.725142 \quad (2)$$

donde : *hpsonte* precipitación diaria máxima anual en sontecomapan, en mm, *hpcatemaco* precipitación diaria máxima anual en Catemaco CFE, en mm, *hpcoyame* precipitación diaria máxima anual en Coyame, en mm

El error medio cuadrático obtenido con este modelo fue de 2425.32, con un coeficiente de determinación R^2=0.3491 (Figura 6).

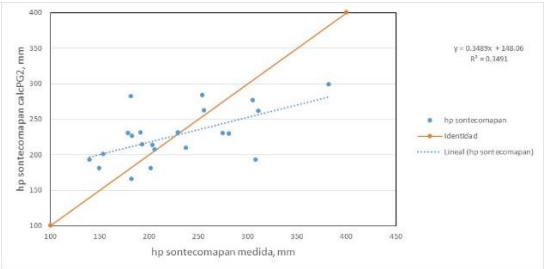

Figura 7. Comparación de datos medidos vs calculados con PG1 vs la identidad. Sontecomapan 1976 a 1997

El modelo PG2 (ecuación 2) se utilizó para rellenar los datos de los años 1952 a 1975 de Sontecomapan, con lo que se obtuvieron registros de igual periodo común de 1952 a 1997 para las tres estaciones, y, con el fin de rellenar datos faltantes ahora en la estación Catemaco CFE, se consideraron como variables independientes a las estaciones Coyame y Sontecomapan y como dependiente a Catemaco, generándose el modelo llamado PG3 (ecuación 3):

$$hpcatemaco=-0.924087*(-0.00054*((-69.486828-hpsonte)-(hpsonte*(2*hpcoyame-2*hpsonte)+2*hpcoyame^2))-hpsonte) \quad (3)$$

donde:
hpcatemaco precipitación diaria máxima anual en Catemaco CFE, en mm
hpcoyame precipitación diaria máxima anual en Coyame, en mm
hpsonte precipitación diaria máxima anual en sontecomapan, en mm

El error medio cuadrático obtenido con el modelo PG3 (ecuación 3) fue de 3434.36 y el coeficiente de determinación entre los datos medidos y calculados fue R^2=0.1343 (Figura 7).

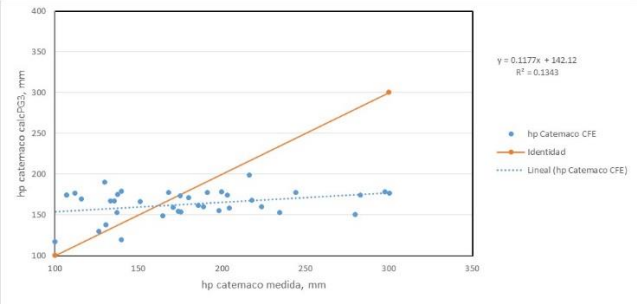

Figura 8. Comparación de datos medidos vs calculados con PG1 vs la identidad. Catemaco CFE. 1952 a 1997

Con el modelo PG3 (ec 3) se rellenaron los datos para la estación Catemaco de 1998 al 2002, teniéndose hasta ese momento completados los registros de 1952 al 2002.

Al faltar el dato del año 2003 para Sontecomapan, se optó por rellenar con la media ese valor debido a que no se contaba con datos simultáneos de las otras dos estaciones (sólo se tenía el dato de la estación Coyame), al rellenar este año faltante, se utilizó el periodo común de 1952 al 2003 para obtener un nuevo modelo matemático para estimar la precipitación máxima anual de Catemaco CFE al que se denominó PG4 (ecuación 4):

$$hpcatemaco=0.550922*hpsonte+36.04453958 \quad (4)$$

El error medio cuadrático obtenido con el modelo PG4 (ecuación 4) fue de 3890.00 y el coeficiente de determinación obtenido entre los datos medidos y calculados bajó a R^2=0.049.

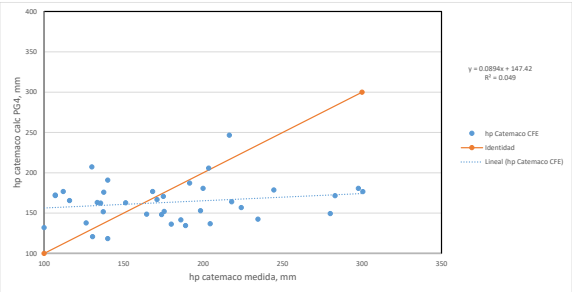

Figura 9. Comparación de datos medidos vs calculados con PG1 vs la identidad. Catemaco CFE. 1952 al 2003

Con el modelo PG4 (ecuación 4) se rellenaron los datos de la estación Catemaco CFE desde el 2003 al 2012 usando los datos de las otras dos estaciones. Posteriormente se rellenó con la media el dato del 2013 de Coyame porque no había datos medidos simultáneamente en Catemaco y en Sontecomapan, con esos datos se volvió a usar el modelo PG4 para rellenar el dato de Catemaco CFE en el 2013. Después se rellenó el dato del 2014 en Coyame y en Sontecomapan y finalmente con esos datos la ecuación PG4 se obtuvo el dato del 2014 en la estación Catemaco CFE. Lo anterior permitió obtener los registros de 1952 al 2015 para las tres estaciones consideradas. Los registros ampliados aparecen en forma gráfica en las Figuras 8 a 10.

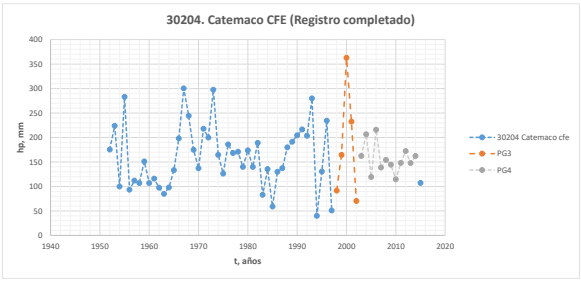

Figura 10. Registro completado Estación 30204 Catemaco CFE

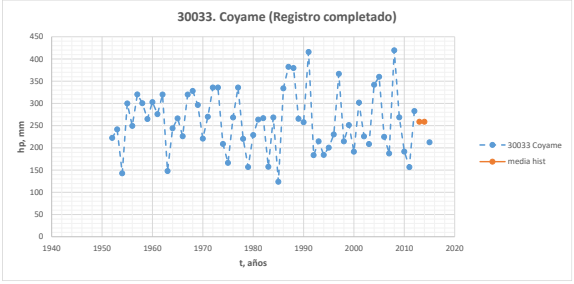

Figura 11. Registro completado Estación 30033 Coyame

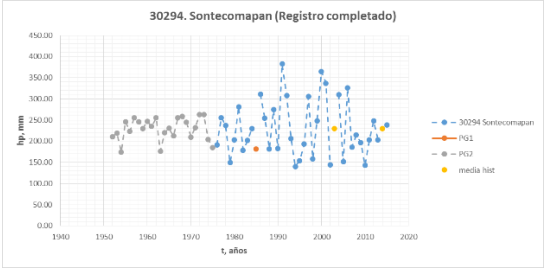

Figura 12. Registro completado Estación 30294 Sontecomapan

En Catemaco y Sontecomapan se observó que los datos completados fueron de menor magnitud en general que los que se tenían registrados históricamente. Los registros ampliados con este procedimiento de rellenado se analizaron estadísticamente para obtener precipitaciones de diseño (de 24 horas) para distintos periodos de retorno. Los resultados se compararon con los ajustes realizados con los registros originales para identificar la posibilidad de sobredimensionamiento o subdimensionamiento de la avenida de diseño al aplicar este procedimiento (Figura 11). En los ajustes se observaron diferencias hasta de 11, 72 y 182 mm para periodos de retorno de 10,000 años en la estación Catemaco CFE, Coyame y Sontecomapan con los modelos de PG y rellenado con la media, indicarían sobredimensionamiento. Para periodos de retorno intermedios (100 años) las diferencias son de 4.5, 34 y 83 mm en los eventos de diseño (sobredimensionamiento) y sólo para los periodos de retorno de 2 y 5 años se tendrían subdimensionamiento en Catemaco CFE, y para 2 años en Coyame y Sontecomapan.

La comparación de los estadísticos históricos respecto a los registros completados (Tablas 2 y 3) indicó escasa variación en la media y en el coeficiente de variación de los registros completados respecto a los originales; en la estación Sontecomapan con rellenado se obtuvieron las diferencias más altas en el mismo atribuidos a las diferencias en la desviación estándar.

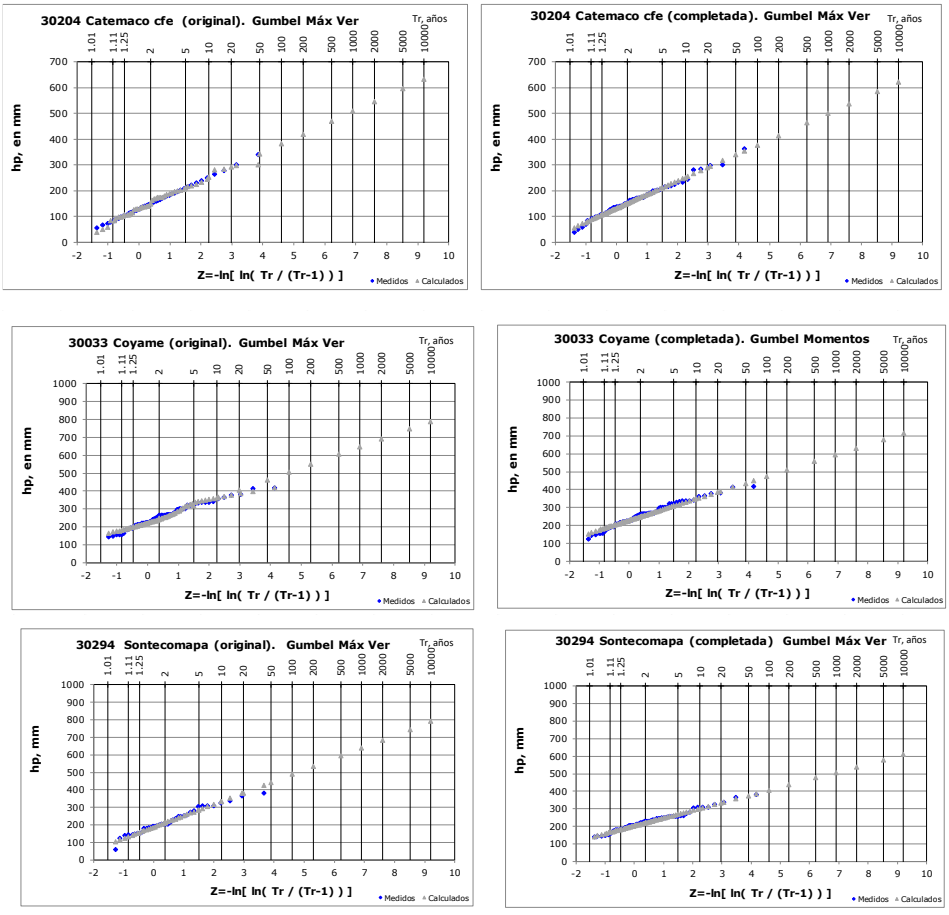

Figura 13. Comparación de ajustes de las tres estaciones analizadas (registro original y ampliado)

Tabla 3. Comparación de estadísticos con el rellenado de datos.

Estadístico	PG1	PG2	PG3		PG4	media en 2013 y 2014	media en2014	PG4
	con 1985	con más años*	con más años**	media en 2003	añadiendo2003al2012	completado***	completado***	completado***
Estación	sontecomapan	sontecomapan	catemaco	sontecomapan	catemaco	coyame	sontecomapan	catemaco
clave	30294	30294	30204	30294	30204	30033	30294	30204
media	228.31	228.58	162.99	228.60	162.99	258.99	228.61	161.05
desvest	61.41	53.53	69.40	53.09	69.40	68.07	52.67	63.51
coef asim	0.66	0.68	0.62	0.68	0.62	0.28	0.68	0.72
cv	0.2690	0.2342	0.4258	0.2323	0.4258	0.2628	0.2304	0.3944

Diferencias vs históricos

Estadístico	PG1	PG2	PG3		PG4	media en 2013 y 2014	media en2014	PG4
	con 1985	con más años	con más años	media en 2003	añadiendo2003al2012	completado	ompletado media en 201	completado
Estación	sontecomapan	sontecomapan	catemaco	sontecomapan	catemaco	coyame	sontecomapan	catemaco
clave	30294	30294	30204	30294	30204	30033	30294	30204
media	1.25	0.97	-3.46	0.96	-3.46	0.00	0.94	-1.52
desvest	4.56	12.44	-6.00	12.87	-6.00	1.11	13.29	-0.12
coef asim	-0.04	-0.06	-0.22	-0.07	-0.22	0.00	-0.07	-0.32
cv	0.0184	0.0532	-0.0284	0.0551	-0.0284	0.0043	0.0570	0.0030

*se añadió de 1952 a 1975
** Se añadió de 1998 al 2002
*** Registro de 1952 al 2015

Conclusiones

Con programación genética se determinaron modelos matemáticos de interpolación y de extrapolación de datos de precipitaciones diarias máximas anuales que permitió el rellenado de datos de estaciones cercanas entre sí. El método es de relativa simplicidad y los estadísticos de los registros ampliados presentan una ligera disminución en su magnitud respecto a los originales, lo anterior también se reflejó en las extrapolaciones para periodos de retorno grandes.

Bibliografía

Arganis-Juárez., ML, Val-Segura, R., Prats-Rodriguez, J., Rodríguez-Vázquez, K., Domínguez-Mora, R. and Dolz-Ripoll, J. (2009) Genetic Programming and standardization in modeling water temperature. Advances in Civil Engineering. Hindawi Publishing Corporation. Volume 2009

Campos-Aranda, DF. Estimación simultánea de datos hidrológicos anuales faltantes en múltiples sitios. Ingeniería Investigación y Tecnología FI-UNAM, volumen XVI (número 2), abril-junio 2015: 295-306

Estimación del error

Se empleó la raíz cuadrada del cuadrado medio del error (RCCME) (Rivas y Carmona, 2010; Teegavarapu y Chandramouli, 2005), el error medio absoluto (MEA) (Teegavarapu y Chandramouli, 2005 y Kashani y Dinpashoh, 2012), el error relativo (RE), coeficiente determinación (R2) y Índice de concordancia de Willmott (d) (Rivas y Carmona, 2010 y Willmott, 1981). El modelo perfecto es cuando R2=1 y MAE= RCCME= RE= 0 y el mejor modelo debe tender a los límites anteriores, siendo excelente cuando d ≥0.95, RE ≤ 0.20 y R2> 0.8 (Caíet al., 2007 y Pereira, 2004).

Resultados y discusión

Para la variable precipitación se obtuvieron valores para la RCCME de: 16.57, 19.16