Épsilon

Volume 1 | Issue 25 Article 2

2015-12-01

Evaluación de métodos convencionales y de aprendizaje de máquina para completar series de precipitación

María Alejandra Caicedo Londoño
Universidad de La Salle, Bogotá, macaicedo@unisalle.edu.co

Juan Carlos Chacón Hurtado
Unesco-IHE University, DELFT, Holanda, j.chaconhurtado@unesco-ihe.org

Evaluación de métodos convencionales y de aprendizaje de máquina para completar series de precipitación

María Alejandra Caicedo Londoño[1]
Juan Carlos Chacón Hurtado[2]

RESUMEN
En este trabajo se discuten los resultados obtenidos tras evaluar ocho métodos basados en el aprendizaje de máquina, y tres métodos correspondientes a los que históricamente se han empleado para completar datos faltantes en series de tiempo. Los datos utilizados para los análisis corresponden a totales mensuales de precipitación, recolectados por el Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia (Ideam) en cuatro estaciones meteorológicas localizadas en la cuenca del río Baché, en el municipio de Palermo (Huila, Colombia). Para llevar a cabo la evaluación de los métodos, se reprodujeron los datos existentes, a manera de faltantes, y sobre la diferencia de estos se calcularon tres métricas distintas de error: raíz de error medio cuadrático (REMC), eficiencia de Nash-Sutcliffe (NSE) y sesgo. Los resultados muestran que los métodos de aprendizaje de máquina para completar series de tiempo son fiables, ya que resultados similares, y en algunos casos mejores, pueden ser alcanzados sin una precisa implementación, y, consecuentemente, una mayor atención a estos puede llevar a resultados menos inciertos.

Palabras clave: aprendizaje de máquina, datos faltantes, interpolación, precipitación mensual, regresión.

[1] Correo electrónico: macaicedo@unisalle.edu.co
[2] Unesco-IHE University, DELFT, Holanda. Correo electrónico: j.chaconhurtado@unesco-ihe.org.

Fecha de recepción: 8 de noviembre de 2014 • Fecha de aprobación: 14 de marzo de 2015

Cómo citar este artículo: Caicedo Londoño, M. A. y Chacón Hurtado, J. C. (2015). Evaluación de métodos convencionales y de aprendizaje de máquina para completar series de precipitación. *Épsilon,* (25), 11-38.

Assessment of Conventional and Machine Learning Methods for Completing Precipitation Series

ABSTRACT

This paper discusses the results of an assessment of eight machine learning-based methods and three methods that have historically been used to complete missing data in time series. Data used for the analysis correspond to monthly precipitation totals collected by the Colombian Institute of Hydrology, Meteorology and Environmental Studies (Ideam) at four weather stations in the Baché river basin (Palermo municipality, Huila, Colombia). In order to evaluate the methods, the existing data was reproduced as missing data, and three different error metrics were calculated based on the difference between them: Root Mean Square Error (RMSE), Nash-Sutcliffe Efficiency (NSE) and bias. Results show that machine learning methods for completing time series are reliable, seeing as similar (and in some cases, better) results can be achieved without an accurate implementation and, consequently, a greater attention to them can lead to less uncertain results.

Keywords: machine learning, missing data, interpolation, monthly precipitation, regression.

Avaliação de métodos convencionais e de aprendizagem de máquina para completar séries de precipitação

RESUMO

Neste trabalho se discutem os resultados obtidos após a avaliação de oito métodos baseados na aprendizagem de máquina, e três métodos correspondentes aos que historicamente têm sido empregados para completar dados faltantes em séries de tempo. Os dados utilizados para as análises correspondem a totais mensais de precipitação, coletados pelo Instituto de Hidrologia, Meteorologia e Estudos Ambientais da Colômbia (Ideam) em quatro estações meteorológicas localizadas na bacia do rio Baché, no município de Palermo, Huila, Colômbia. Para levar a cabo a avaliação dos métodos, se reproduziram os dados existentes, a maneira de faltantes, e em base a diferença destes se calcularam três métricas diferentes de erro: raiz de erro médio quadrático (REMC), eficiência de Nash-Sutcliffe (NSE) e sesgo. Os resultados mostram que os métodos de aprendizagem de máquina para completar séries de tempo são confiáveis, já que resultados similares, e em alguns casos melhores, podem ser alcançados sem uma precisa implementação, e, consequentemente, uma maior atenção a estes pode levar a resultados menos incertos.

Palavras chave: aprendizagem de máquina, dados faltantes, interpolação, precipitação mensal, regressão.

Introducción

Estudios detallados que permitan estimar con fiabilidad la precipitación son herramientas indispensables para cuantificar y caracterizar la disponibilidad del recurso hídrico y su disposición final entre los diferentes usos (doméstico, agricultura, ganadería, industria, entre otros). Sin embargo, para obtener resultados coherentes y con un alto grado de correspondencia a la realidad es indispensable partir de registros periódicos que den continuidad a las observaciones de los fenómenos que se estudian.

Son múltiples las situaciones que pueden afectar la calidad de los datos o registros de precipitación. Algunas de ellas pueden ser inherentes al ausentismo del operador y a posibles fallos en las estaciones meteorológicas, cuyos registros pueden verse alterados por el cambio de lugar o movimiento del instrumento, el cambio del espacio físico del entorno donde se encuentra la estación, o simplemente no registrar el dato en el momento adecuado. Adicionalmente, el alto costo que trae consigo la instalación de una estación pluviométrica limita el número de las estaciones que se pueden instalar. Todas las anteriores son tan solo algunas situaciones que se pueden mencionar y que llevarían a la inconsistencia o ausencia de datos.

Este documento inicia con una introducción corta sobre los métodos utilizados para completar datos faltantes en series de precipitación. Después, se hace una introducción al caso de estudio y el diseño experimental, que incluye el análisis descriptivo de los datos. Luego se presenta el análisis y la discusión de resultados, para, finalmente, concluir y presentar las recomendaciones que se deben tener presentes en los estudios posteriores, necesarios para llevar a cabo una investigación completa de este tema.

Métodos convencionales para la reconstrucción de series de precipitación

US Weather Bureau

Es un método propuesto por el Departamento de Hidrología del Servicio Nacional de Meteorología de los Estados Unidos, US Weather Bureau (Paulhus y Kohler, 1952). Se basa fundamentalmente en dos métodos de interpolación para completar

series de precipitación diaria. Si los datos de la precipitación media anual en cualquiera de las estaciones índice difiere de la estación de estudio (estación con la serie incompleta) en más de un 10 %, se aplicará una triple interpolación entre la precipitación anual media en cada estación índice, la precipitación anual media en la estación de estudio y la precipitación mensual en cada estación referenciada, para al final obtener la precipitación diaria faltante en la estación de estudio.

Por otro lado, si los datos de precipitación anual media en cada estación índice se encuentran dentro del 10 % de la correspondiente a la estación de estudio, se considera aceptable completar la serie con el ajuste obtenido mediante un promedio aritmético simple. Este método se extiende también a totales mensuales de precipitación, en los cuales solamente se deberá modificar la escala temporal de los análisis.

Promedios

El método consiste en tomar como referencia la serie completa de precipitación anual media de una estación índice. Considerando la precipitación anual media en la estación de estudio, se realiza un promedio aritmético de los registros de precipitación anual que cuenten con datos, tanto en la estación índice como en la estación de estudio. En el caso que se presenten más de una estación índice, se procede de igual manera con cada una de ellas, para luego obtener un nuevo promedio de los registros anuales faltantes de precipitación en la estación de estudio.

Método de la recta de regresión

Es un método que sirve para estimar datos faltantes de una serie temporal de precipitación cuando se cuenta con varias estaciones índices. Su metodología inicia realizando una regresión lineal entre los registros de la estación de estudio y cada una de las estaciones índices. Estadísticamente, el máximo absoluto del coeficiente de correlación (r) corresponde al de la estación índice más adecuada. Este parámetro se obtiene ajustando la serie de datos a una recta de regresión que minimice la distancia media cuadrática.

Un coeficiente de regresión igual a cero (r = 0) significa una correlación nula, lo cual implica que no existe ningún grado de asociación entre las dos series de datos. Un coeficiente de regresión igual a la unidad (r = 1) indica una correlación directa

óptima, mientras que un coeficiente de regresión negativo (r = -1) significa una regresión inversa óptima.

Métodos basados en aprendizaje de máquina para completar series de precipitación

Las técnicas de regresión basadas en el aprendizaje de máquina son herramientas que permiten, a través de determinadas reglas, inferir una serie de datos a partir de otras. El subgrupo de estas reglas es usado en la reconstrucción de series de tiempo que corresponden a aprendizaje supervisado, en el cual la intención es minimizar una métrica de error entre los datos simulados y los datos medidos durante un periodo de entrenamiento. Las técnicas empleadas en este estudio son:

- Procesos de Gauss
- Regresión lineal
- Perceptrón multicapa (redes neuronales)
- Regresión simple lineal
- Máquinas de soporte de vectores para regresión
- Tablas de decisión
- Árboles de decisión M5P
- Promedio largo plazo (Zero-R)

Procesos de Gauss

Los procesos de Gauss (Rasmussen y Williams, 2006) son una forma no paramétrica de regresión, basada en el principio de que las variables que van a ser predichas y las predictoras provienen de una distribución de Gauss multivariada. Aunque este es un método menos paramétrico (Edben, 2008) que los ajustes de regresión de mínimos cuadrados, la definición de una función de covarianza es requerida para aplicar las reglas de regresión. Como consecuencia, dentro de los beneficios de esta aproximación, se encuentra que la predicción de la variable es función únicamente de los datos y por tanto no está precondicionada por la dinámica del proceso. Dentro de los procesos de Gauss más conocidos se encuentran Kriging (Journel y Huijbregts, 1978) y Filtro Kalman (Kalman, 1960).

Regresión lineal

Se utiliza para establecer los valores de los coeficientes de las variables predictoras que permitan minimizar la diferencia entre los datos simulados y los de entrenamiento. En principio la eficiencia del desempeño se mide respecto al error cuadrático medio (Helsel y Hirsch, 1992). Sin embargo, a la hora de implementarse dicha condición puede variar. Este es sin duda uno de los más populares métodos de regresión dada su simplicidad en formulación e interpretación de resultados.

Perceptrón multicapa

Este es un sistema de regresión de tipo redes neuronales prealimentadas (Svozil, Kvasnicka y Pospichal, 1997), en el cual los parámetros predictores son transformados para ajustar las variables simuladas. En esta configuración, cada perceptrón actúa como una neurona a la cual se le aplica una transformación lineal para cada una de las variables de entrada. Posteriormente, los resultados obtenidos de dichas neuronas se modifican de nuevo a través de una transformación logística en el espacio de las variables predictorias. Estas transformaciones en principio deben minimizar la diferencia entre la serie de datos simulada y los registros utilizados para esta.

En este método se pueden identificar parámetros e hiperparámetros (MacKay, 1999), los cuales definen el comportamiento del sistema. Los primeros corresponden a los coeficientes de las transformaciones lineales y logísticas. Los últimos son parámetros que se deben establecer previamente, en los que el diseñador del sistema debe determinar el número de neuronas en la red. En principio, el incremento en el número de neuronas siempre llevará a mejores ajustes en las series de entrenamiento; sin embargo, el número de estas se debe limitar para evitar problemas de sobreparametrización.

Máquinas de soporte de vectores para regresión

La idea fundamental de esta aproximación es la de transformar el espacio de predictores de tal manera que se pueda identificar un hiperplano en el que se minimice la distancia entre los predictores y las variables simuladas (Cristianini y Shawe-Taylor, 2000). Para lograr este propósito es necesario establecer un Kernel (reglas de transformación), que será utilizado para incrementar la dimensionalidad

del problema, de tal modo que las formas más complejas puedan ser generalizadas desde las transformaciones más simples.

La selección del Kernel en estos casos es, sin duda, uno de los puntos críticos en la parametrización de este método. El efecto de este recae principalmente en la eficiencia y en la convergencia de la solución, y no necesariamente en los resultados de esta para procesos débilmente lineales.

Reglas M5

De manera similar a los árboles de decisión, las M5 construyen una serie de reglas en las cuales el resultado final pertenece a un árbol de decisión M5 (Wang y Witten, 1996).

Árboles de decisión M5P

De la misma forma que las tablas de decisión, los árboles de decisión usan regresión lineal a los resultados de la clasificación de una manera análoga a las reglas M5 (Frank, Wang, Inglis, Holmes y Witten, 1997). En principio, esta alternativa se puede entender como regresiones lineales en paquetes de datos, lo que permitiría obtener mejores ajustes en cuanto al error cuadrático medio durante el entrenamiento, con respecto a métodos como regresión lineal y tablas de decisión. No obstante, los resultados son necesariamente superiores durante el proceso de validación. En este método es importante limitar el número de nodos finales con el fin de evitar problemas de sobreparametrización.

0-R

Es usualmente utilizado como método de referencia, dado que trabaja con la media de la variable predicha como elemento de regresión. En principio, si los resultados del modelo de predicción son peores que los del modelo 0-R, se da por supuesto que no aportan algún tipo de información.

Tablas (árboles) de decisión

Las tablas de decisión son un conjunto de reglas que permiten determinar la pertenencia de una variable a cierto grupo (Mitchell, 1997). En aplicaciones de regre-

sión, el conjunto de variables predictoras clasifica a la variable predicha en grupos, para la cual la variable predicha cuenta con un valor asignado. Este paradigma es fundamentalmente aplicado a variables discretas; sin embargo, su generalización puede llevar a resultados fáciles de interpretar en intervalos de tiempo de simulación notablemente bajos.

Descripción del caso de estudio

El uso potencial del suelo, las características topográficas de la región y la abundancia del recurso hídrico en las zonas planas tiende idealmente hacia el aprovechamiento del suelo en sistemas forestales o silviculturales; mientras que el agua proveniente de las zonas montañosas con gran importancia por su efecto captador y regulador es utilizada en un mayor porcentaje para el consumo humano, pues abastece acueductos que alimentan el municipio de Palermo y poblaciones cercanas. En menor proporción, pero no menos importante, se cuenta con actividades como piscicultura y riegos de cultivos de arroz, que se refieren a la economía característica del municipio (Consejo Municipal de Palermo Huila, 2013).

Claramente es posible hacer una conexión entre el uso del suelo de la región de estudio y la necesidad de contar con datos de precipitación altamente correlacionados entre todas las estaciones pluviométricas instaladas en la zona, pues son estos datos la fuente primaria de información en los estudios para el pronóstico del clima, lo cual influencia no solo su economía, sino también las grandes áreas propensas a inundación localizadas en el valle, la franja delimitada entre los ríos Baché y Magdalena.

Localización e información general de la zona de estudio

La cuenca hidrográfica del río Baché, con una orientación suroeste-noroeste y una corriente principal de una longitud aproximada de 115 km, cuenta con una extensión de 1041,93 km^2. Se localiza sobre la vertiente oriental de la cordillera Central en jurisdicción de los municipios de Santa María, Palermo, Teruel, Aipe y Neiva, departamento del Huila (figura 1). Las elevaciones mínimas y máximas corresponden, respectivamente, a los 384 m s. n. m. en su desembocadura sobre el río Magdalena, y 3400 m s. n. m. en las estribaciones del nevado del Huila (Carmona, 2002)

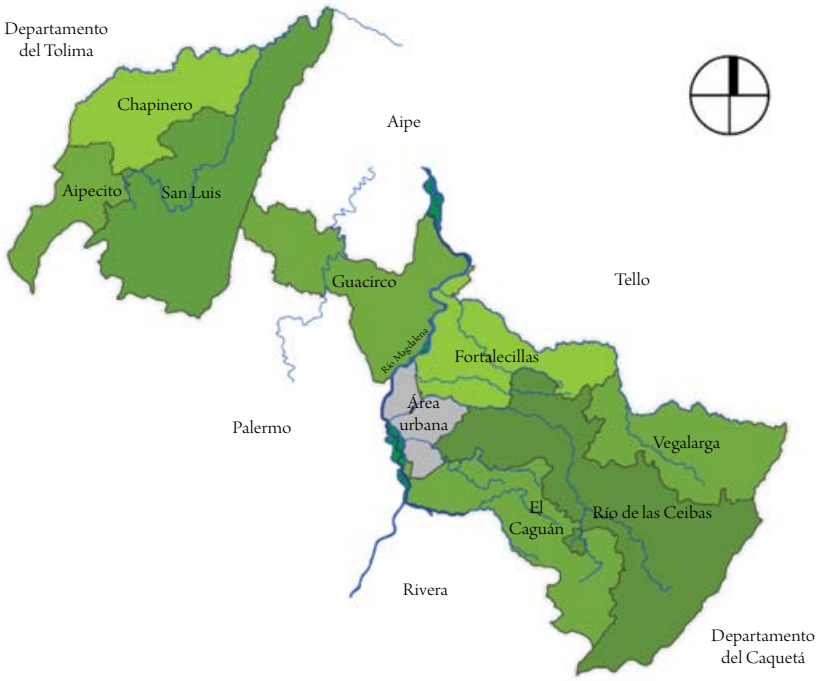

Figura 1. **Localización geográfica de la cuenca del río Baché**

Fuente: Wikipedia (2015).

La fisiografía del municipio de Palermo presenta paisajes como montañas, piedemonte, lomeríos y valles, con climas que corresponden de igual forma: frío y húmedo, medio y húmedo, cálido seco y muy seco. La temperatura media del municipio presenta unas temperaturas que oscilan entre los 15 °C en las zonas de las cordilleras y 27 °C para las zonas bajas, las cuales corresponden a los valles de los ríos Magdalena y Baché (Consejo Municipal de Palermo Huila, 2013).

Estaciones pluviométricas

Los registros de precipitación utilizados en este estudio fueron suministrados por la regional Huila-Caquetá del Ideam, dentro del proyecto del Plan Departamental de la Cuenca del Río Baché. En total se trabajaron con 1191 registros sin contar con los datos perdidos (figura 2). La estación Rionegro se localiza a una elevación de 560 m s. n. m., fue instalada en 1979. La estación Totumo se localiza a una

elevación de 700 m s. n. m., fue instalada en 1971. La estación Volcán se localiza a una elevación de 1105 m s. n. m., fue instalada en 1976. La estación Paraguay se localiza a una elevación de 1300 m s. n. m., fue instalada en 1986. Por esta razón se seleccionó el periodo de análisis comprendido entre 1986 y 2012, y en los meses de enero a diciembre.

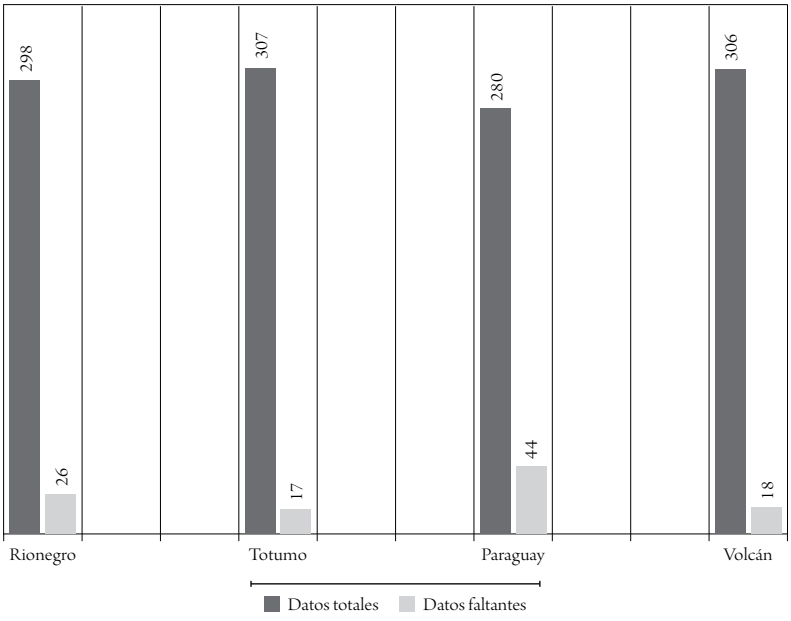

Figura 2. **Número de datos totales y faltantes en cada estación pluviométrica, periodo de análisis 1986-2012**

Fuente: elaboración propia.

La precipitación media mensual para todo el periodo de estudio (figura 3) evidencia que de junio a septiembre se presentan los valores más bajos de precipitación, mientras que de octubre a diciembre, los más altos. Esto permite hacerse una idea de la variación estacional de la precipitación en la zona. En la tabla 1 se presentan los parámetros estadísticos más representativos para los datos de precipitación mensual en cada una de las estaciones.

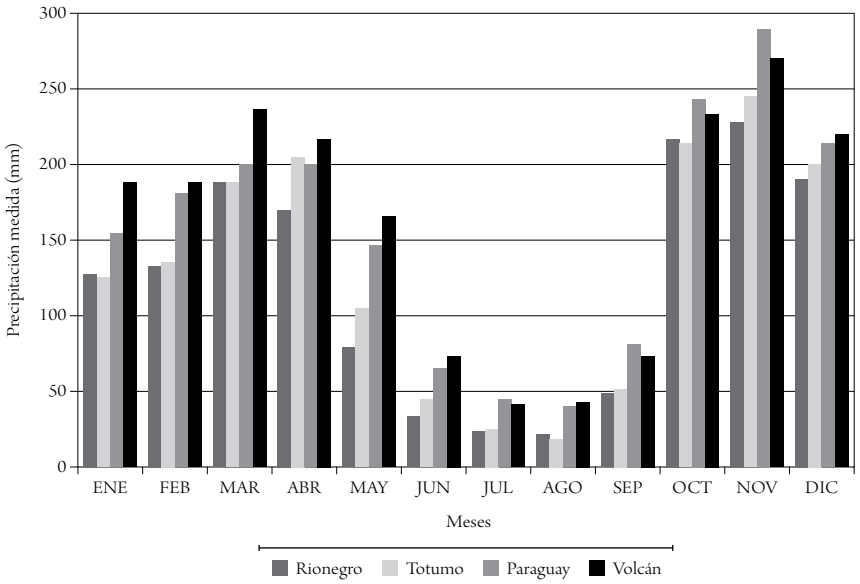

Figura 3. **Variación de la precipitación media mensual en cada estación, 1986-2012**

Fuente: elaboración propia.

Diseño experimental

En este documento se realizó la evaluación de ocho métodos de aprendizaje de máquina y tres métodos correspondientes convencionales. Para llevar a cabo la evaluación de los métodos mencionados, se reprodujeron los datos existentes a manera de faltantes, y sobre la diferencia de estos, se calcularon tres métricas de error distintas: la raíz del error cuadrático medio (REMC), la eficiencia de Nash-Sutcliffe (NSE) y el sesgo. Las dos primeras corresponden a medidas de dispersión del error, mientras que la última representa la desviación estándar entre los valores medios.

Los modelos de aprendizaje de máquina y de regresión lineal se parametrizaron (entrenados) con el 60 % de los datos, mientras el 40 % restante fueron utilizados para la validación. Al dividir la serie de esta manera, se está garantizando la posibilidad de realizar pruebas independientes con datos que no han sido utilizados para el entrenamiento de los modelos, simulando las condiciones habituales de uso.

Épsilon, nº 25 • julio-diciembre 2015 • pp. 11-38 • ISSN 1692-1259-ISSN-E 2389-8828

Tabla 1. **Parámetros estadísticos de los datos de precipitación mensual, 1986-2012**

Estación	Elevación	Datos perdidos (%)	Parámetros estadísticos	Meses											
				1	2	3	4	5	6	7	8	9	10	11	12
Rionegro	560	8,0	Pm* (mm)	127,40	132,50	188,50	170,10	79,57	34,37	23,54	21,09	49,12	216,80	228,00	189,50
			Sp*	75,72	103,90	95,04	126,90	70,83	45,29	35,86	20,16	44,46	108,40	125,80	91,82
			Cv*	0,59	0,78	0,50	0,75	0,89	1,32	1,52	0,96	0,90	0,50	0,55	0,48
Totumo	700	5,2	Pm* (mm)	126,30	136,50	187,80	204,30	105,70	45,05	25,06	18,00	52,08	214,70	245,10	201,30
			Sp*	83,90	111,10	94,00	125,60	68,00	51,50	50,30	28,50	57,80	109,70	122,30	94,80
			Cv*	0,66	0,81	0,50	0,61	0,64	1,14	2,01	1,58	1,11	0,51	0,50	0,47
Paraguay	1300	13,6	Pm* (mm)	155,00	180,90	201,10	199,10	146,50	65,26	44,12	40,26	81,39	243,00	289,40	214,50
			Sp*	115,50	167,00	143,90	127,60	128,20	78,90	90,20	33,20	67,50	126,90	164,70	106,50
			Cv*	0,75	0,92	0,72	0,64	0,88	1,21	2,05	0,82	0,83	0,52	0,57	0,50
Volcán	1105	5,6	Pm* (mm)	187,80	188,70	236,60	216,90	166,40	73,70	41,92	42,97	74,28	234,30	270,00	220,00
			Sp*	80,30	136,90	117,00	138,40	88,00	78,60	56,10	46,00	115,10	110,00	105,60	95,40
			Cv*	0,43	0,73	0,49	0,64	0,53	1,07	1,34	1,07	1,55	0,47	0,39	0,43

Pm*: precipitación media mensual; Sp*: desviación estándar de la precipitación mensual; Cv*: coeficiente de variación de la precipitación mensual.

Fuente: elaboración propia.

Para el desarrollo del estudio no se consideró una investigación exhaustiva de los hiperparámetros en los modelos de aprendizaje de máquina. Esto llevaría a un estudio detallado de cada uno de los métodos, lo que se considera fuera del alcance del presente documento. No obstante, sí se establece una línea base que permita determinar la viabilidad y posibles limitaciones en el uso de estas técnicas y su valor frente a los métodos convencionales.

Análisis de datos

Como se describió anteriormente, las series de datos utilizadas cuentan con registros tomados desde 1986 hasta 2012. En la figura 4 se puede apreciar cómo las series de datos no son homogéneas debido a una significativa falta de datos en 2008. Los motivos por las cuales se presenta la incongruencia en este año se desconocen dado que dichos registros no fueron registrados como datos perdidos en la información suministrada por la entidad, y es posible suponer que es producto de una falla humana al momento de realizar la copia de datos. Por lo tanto, la información será descartada en el presente estudio, considerando que no es fundamental el uso de series de tiempo continuas en un análisis de este tipo.

Como medida inicial se realizó un análisis de correlación (figura 5) entre las distintas estaciones. Este análisis permite identificar similitud estadística entre las variables medidas y por tanto servirá para establecer cuáles son las estaciones índice que suministran una mayor cantidad de información en la estimación de los datos perdidos.

De manera análoga, el análisis de autocorrelación (figura 6) permite identificar patrones en la línea de tiempo de las series de precipitación, lo que muestra oscilaciones que pueden corresponder a ciclos naturales de los fenómenos de precipitación. Dentro de estos ciclos se encuentran oscilaciones anuales típicas del ciclo hidrológico anual, así como oscilaciones de más larga frecuencia tales como los fenómenos del niño y la niña. Este análisis permite establecer cuál es la similitud estadística de periodos anteriores en la simulación de datos perdidos. Los resultados también muestran que los ciclos anuales son similares; sugieren que incluir información acerca de los años anteriores en el mismo mes puede contribuir con la estimación de los datos perdidos.

Épsilon, Nº 25 • julio-diciembre 2015 • pp. 11-38 • ISSN 1692-1259-ISSN-E 2389-8828

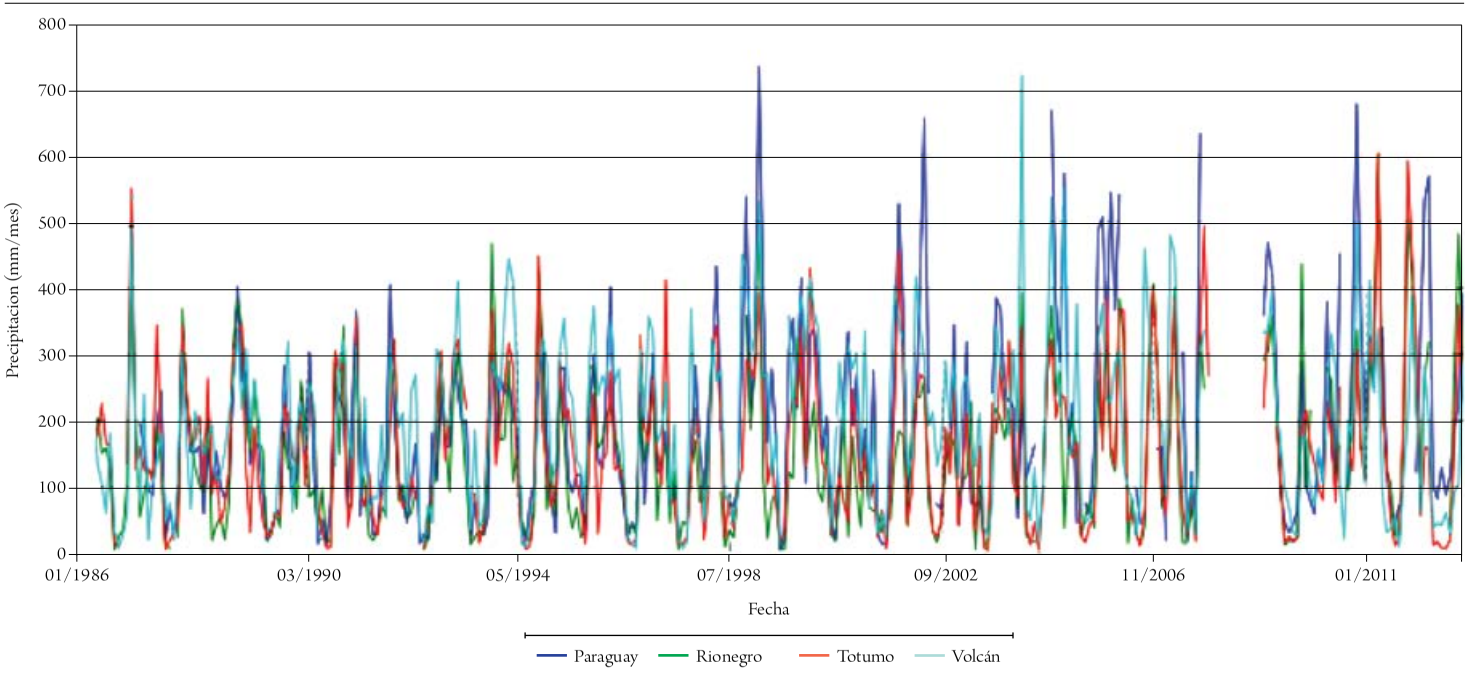

Figura 4. **Series de datos para estaciones en cuenca del río Bache**

Fuente: datos adaptados del Ideam.

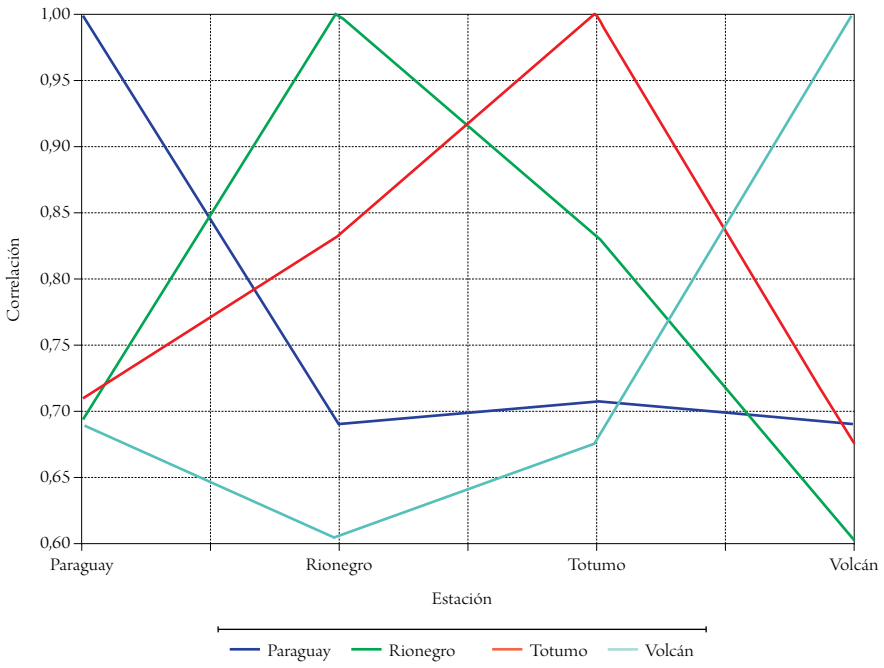

Figura 5. **Coeficiente de correlación entre estaciones**

Fuente:

Resultados

Teniendo en cuenta los resultados del análisis de correlación y autocorrelación, se estableció que las medidas más significativas que contribuyen a la predicción de los valores perdidos son las mediciones del mismo periodo en las otras estaciones, al igual que los registros de la misma estación pero en un periodo de tiempo comprendido entre 12 y 24 meses previo del dato perdido. Es importante establecer que según los resultados de autocorrelación de las series, se presenta un comportamiento similar casi indefinidamente; sin embargo, se decidió truncar a dos periodos con el fin de reducir el número de variables predictoras en los modelos de aprendizaje de máquina. Esto repercute en una mejor relación variables-observaciones y, por lo tanto, se podría esperar un mejor desempeño de estas variables.

Los resultados para todos los métodos de entrenamiento se presentan en las figuras 7, 8 y 9. En estos se puede apreciar que los resultados son bastante consistentes entre todos los métodos, beneficiando particularmente a los basados en aprendi-

Épsilon, Nº 25 • julio-diciembre 2015 • pp. 11-38 • ISSN 1692-1259-ISSN-E 2389-8828

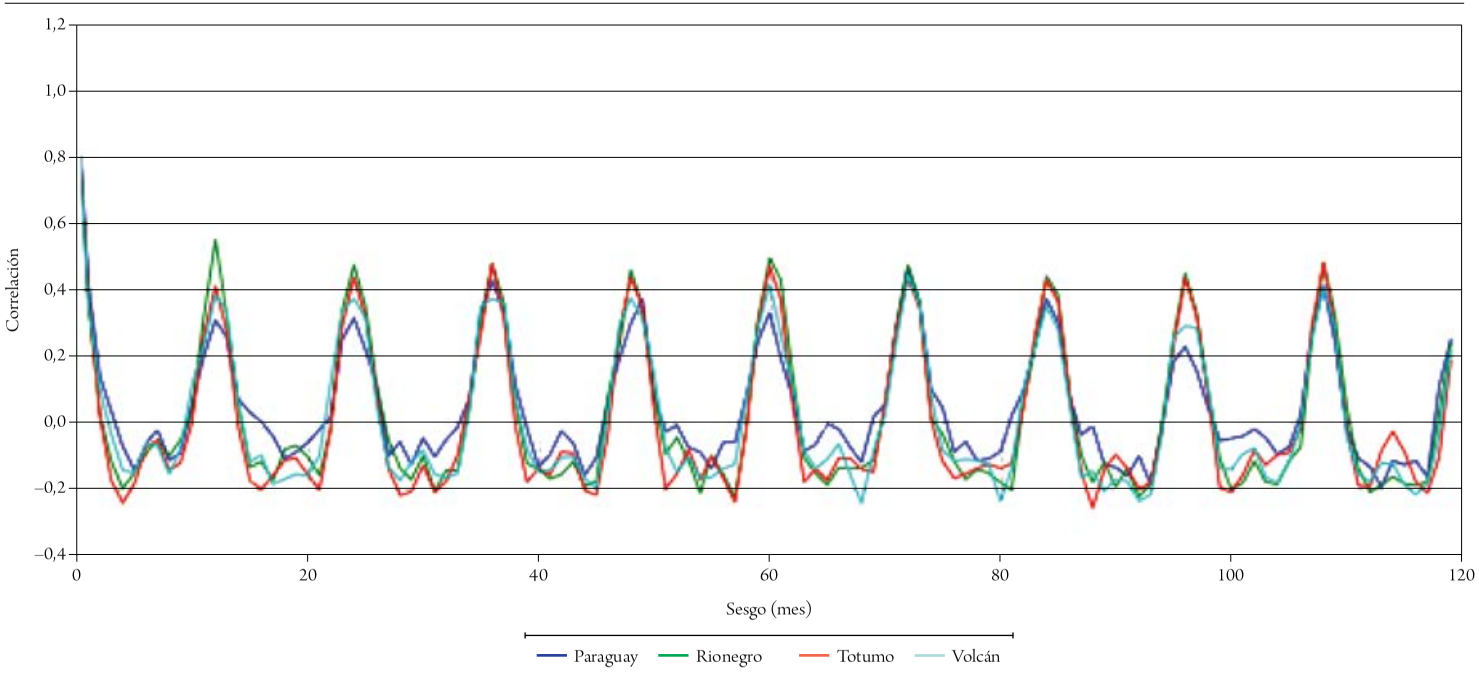

Figura 6. **Autocorrelación para cada una de las estaciones**

Fuente: elaboración propia.

zaje de máquina (a excepción de los procesos de Gauss), en comparación con los métodos convencionales para cada una de las métricas simuladas. Es importante considerar que el método 0-R se usa como predicción del valor medio de la serie a largo plazo, y por tanto se espera que alguna variable predictora presente un mejor desempeño que este valor. Adicional a esto, se puede observar que los resultados entre estaciones son considerablemente homogéneos.

Durante el periodo de validación se pudo establecer que los resultados son notablemente diferentes (figuras 10, 11 y 12). Para el periodo simulado se pueden apreciar significativas discrepancias entre las estaciones, si se agrupa a Paraguay y Volcán en un grupo, mientras que Rionegro y Totumo, en otro. Para el primero se evidencia que el desempeño en todos los métodos es notablemente insuficiente en comparación con el segundo, con lo cual es evidente que existen discrepancias en los datos que el análisis de correlación por sí solo no puede detectar. Tales discrepancias se reflejan por igual en ambos grupos sin diferenciar el método utilizado. Este hecho permite concluir que puede ser la naturaleza de los datos la que infiere en cambios en sus propiedades estadísticas durante los dos periodos de análisis.

Otra posible explicación para estos resultados se evidencia en el sobreajuste de los métodos, dado que imposibilitan la generalización de resultados a partir de los datos de entrenamiento. Esto ocurre en muchos casos por una mala elección de los hiperparámetros de los modelos de aprendizaje, series de tiempo muy cortas, o motores de optimización que no convergen en mínimos globales. La causa fundamental de estos problemas no forma parte del alcance de este estudio; sin embargo, se reconoce como una fuente de error que debe ser considerada en estudios más detallados de problemas de este tipo.

Los resultados de las regresiones fueron posteriormente validados a través del análisis de cuantiles. Se pudo observar que las tablas de decisión y la regresión 0-R se vieron especialmente afectadas debido a que las primeras solamente proveen resultados en clases discretas de valores continuos (figuras 13 y 14). Sin embargo, la tendencia general de la regresión indica que la relación entre cuantiles es lineal y es posible ajustarla por medio de una parametrización distinta de las clases de salida en el modelo. La regresión 0-R, como se ha establecido anteriormente, utiliza el promedio a largo plazo como la mejor estimación del dato perdido; en este sentido, el análisis de cuantiles revela una tendencia totalmente vertical, que se puede traducir en una inhabilidad del modelo para reproducir la variabilidad temporal del proceso.

Épsilon, nº 25 • julio-diciembre 2015 • pp. 11-38 • ISSN 1692-1259-ISSN-E 2389-8828

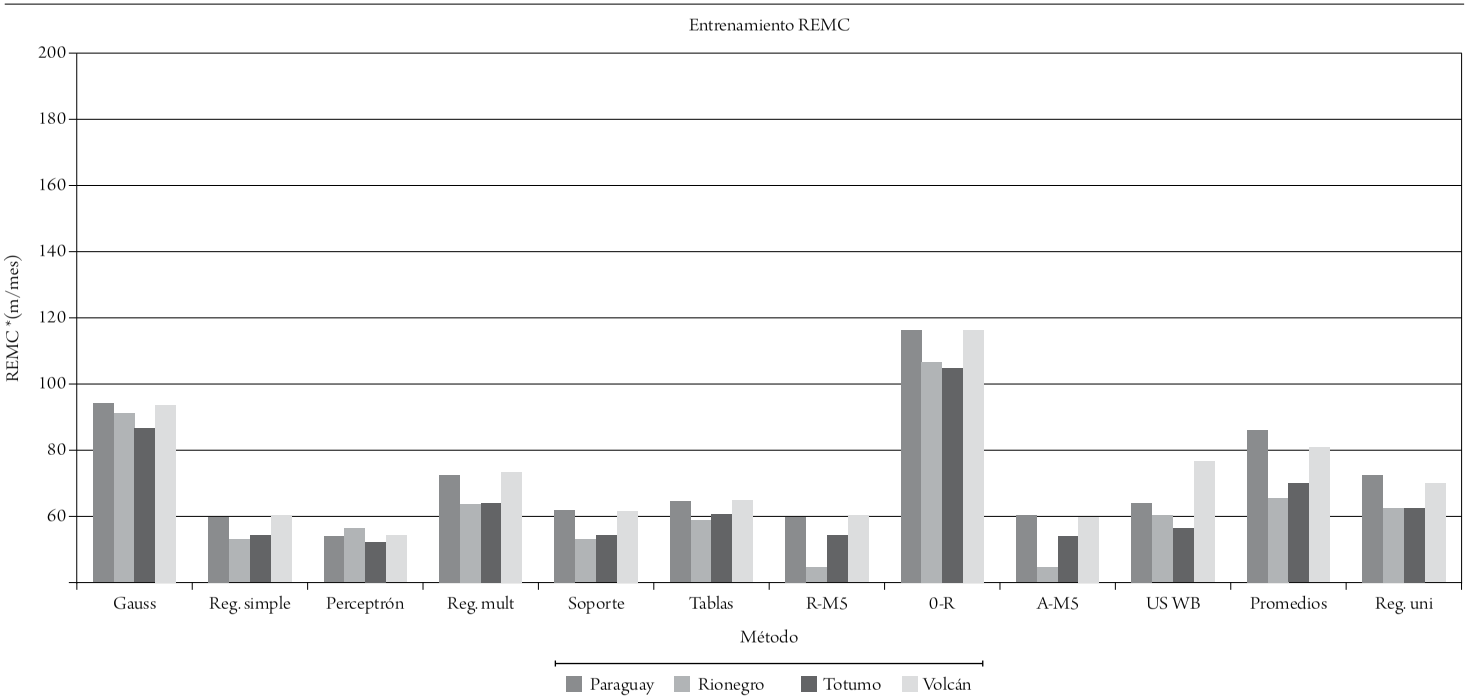

Figura 7. **Error medio cuadrático en entrenamiento**

Fuente: elaboración própia.

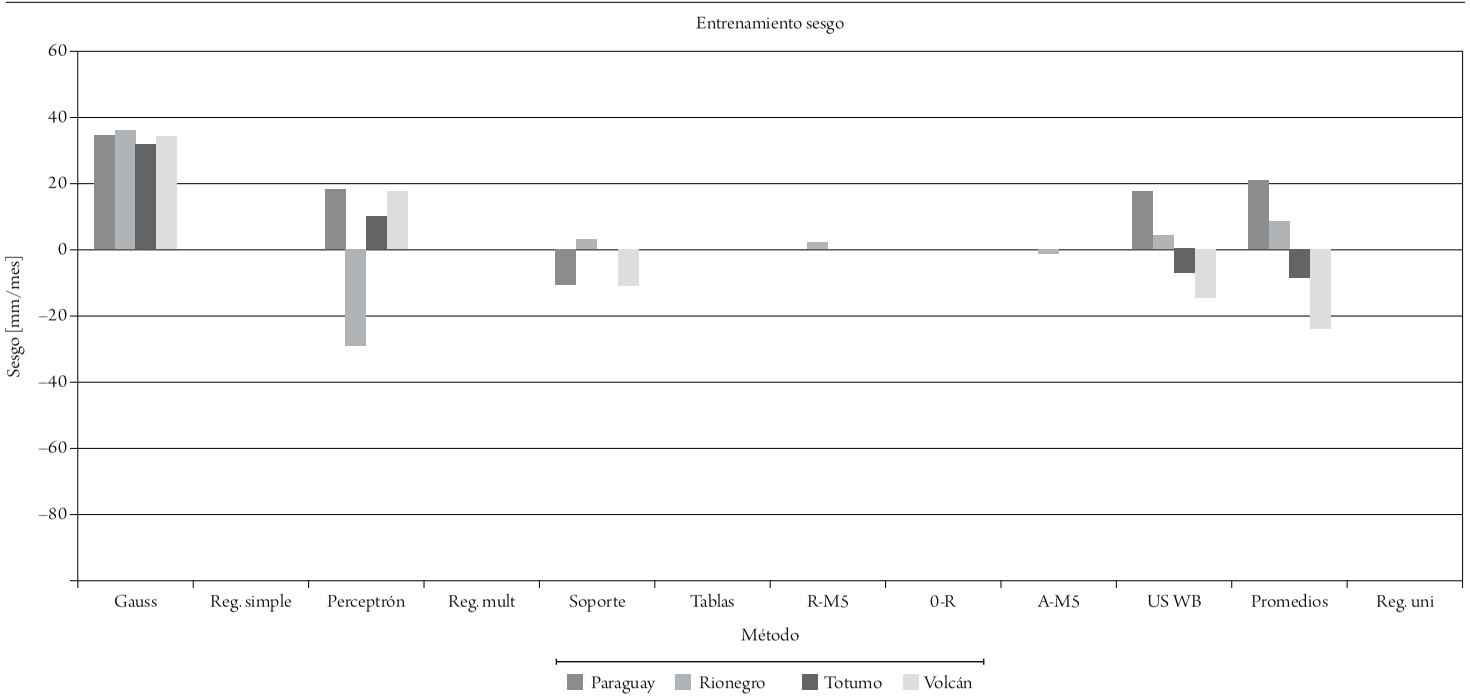

Figura 8. **Eficiencia de Nash-Sutcliffe en entrenamiento**

Fuente: elaboración propia.

Épsilon, Nº 25 • julio-diciembre 2015 • pp. 11-38 • ISSN 1692-1259-ISSN-E 2389-8828

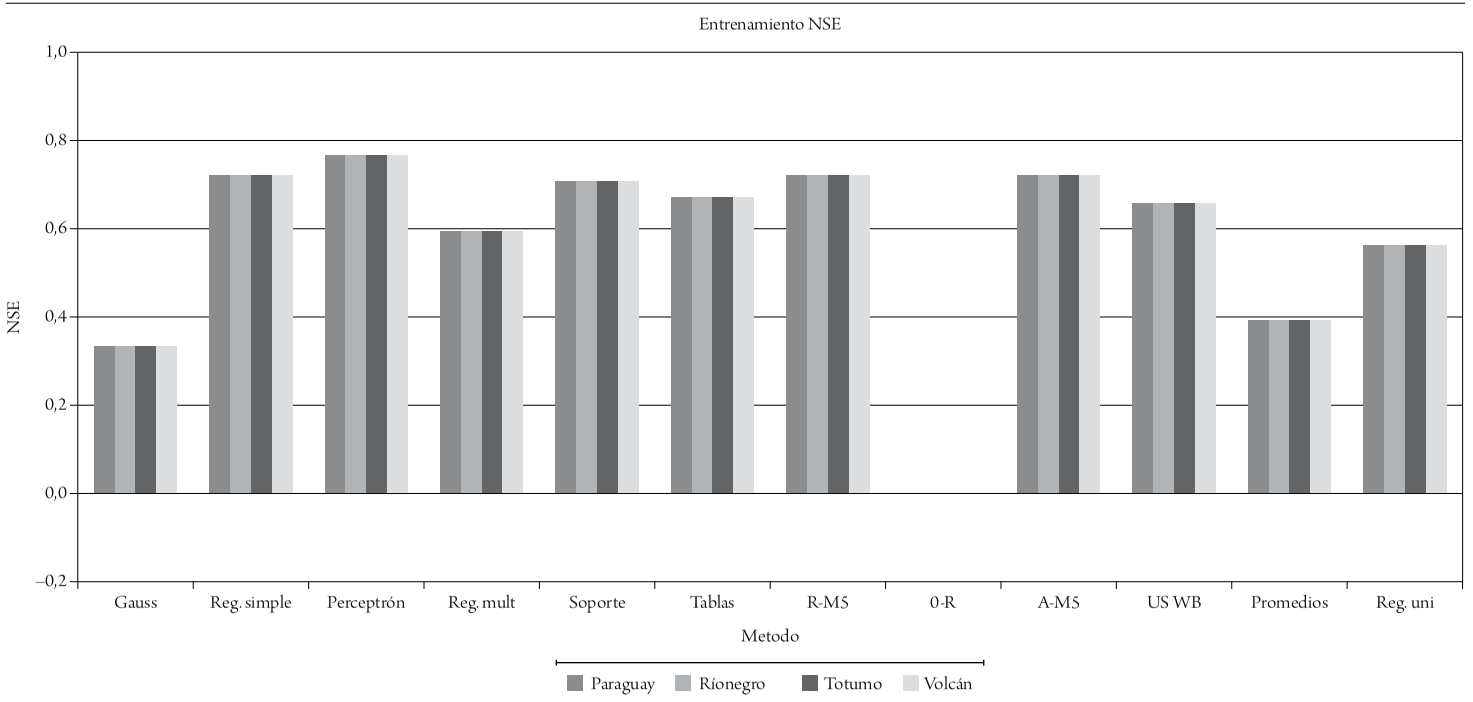

Figura 9. **Sesgo en entrenamiento**

Fuente: elaboración propia.

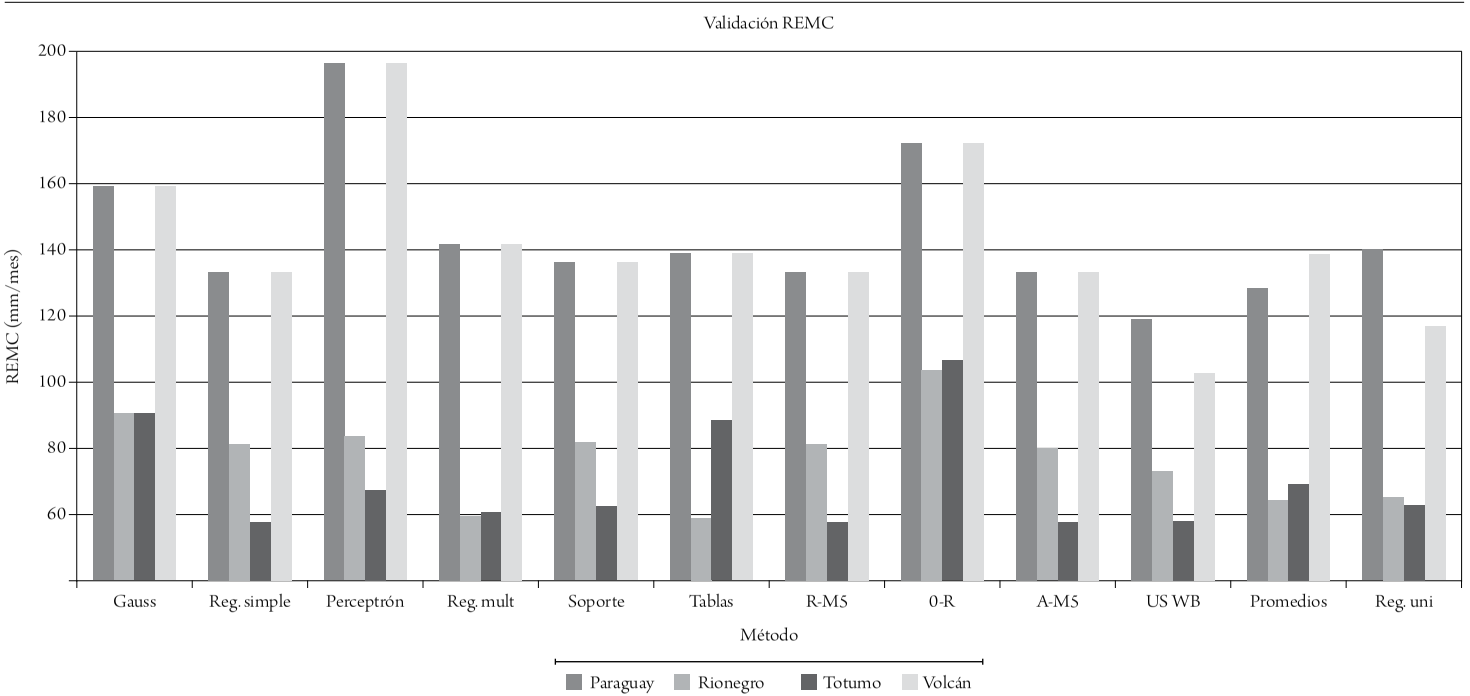

Figura 10. **Raíz del error medio cuadrático en validación**

Fuente: elaboración propia.

Épsilon, nº 25 • julio-diciembre 2015 • pp. 11-38 • ISSN 1692-1259-ISSN-E 2389-8828

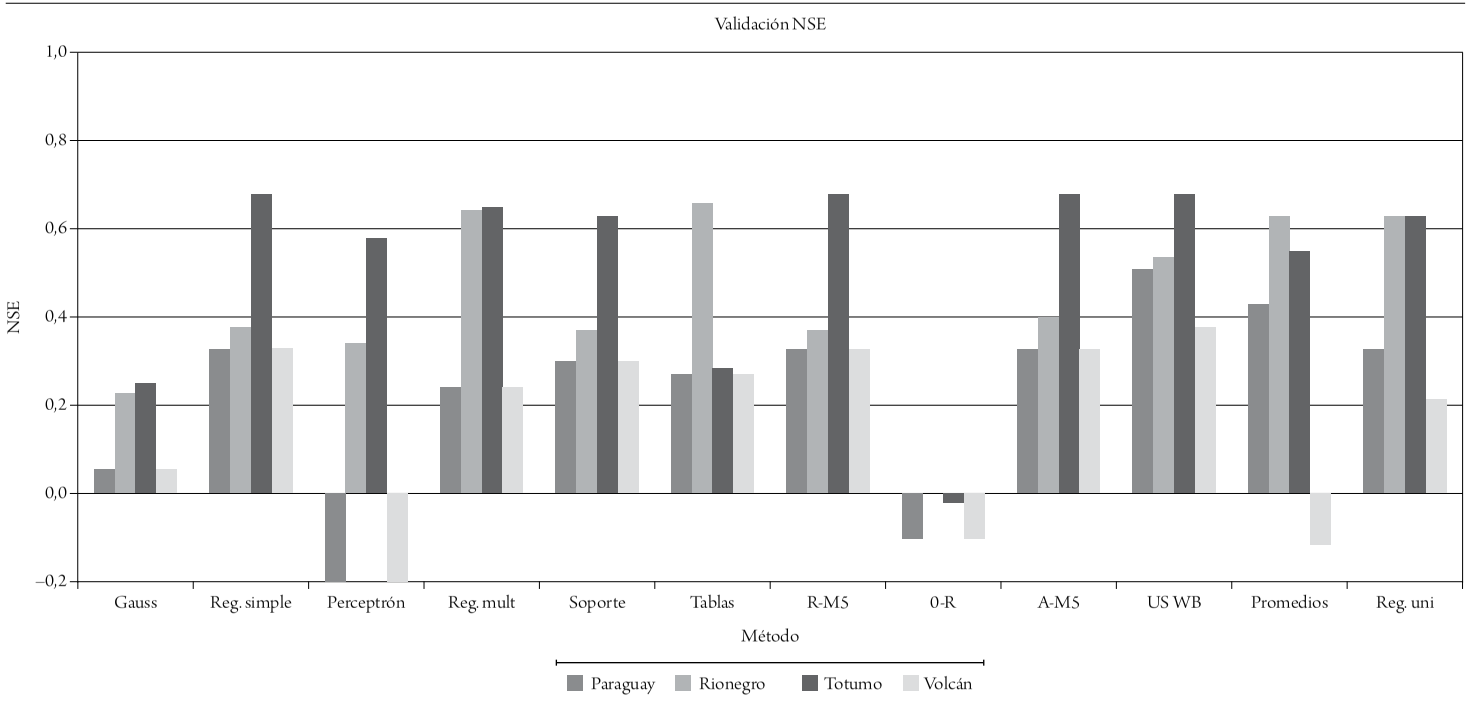

Figura 11. **Eficiencia de Nash-Sutcliffe en validación**

Fuente: elaboración propia.

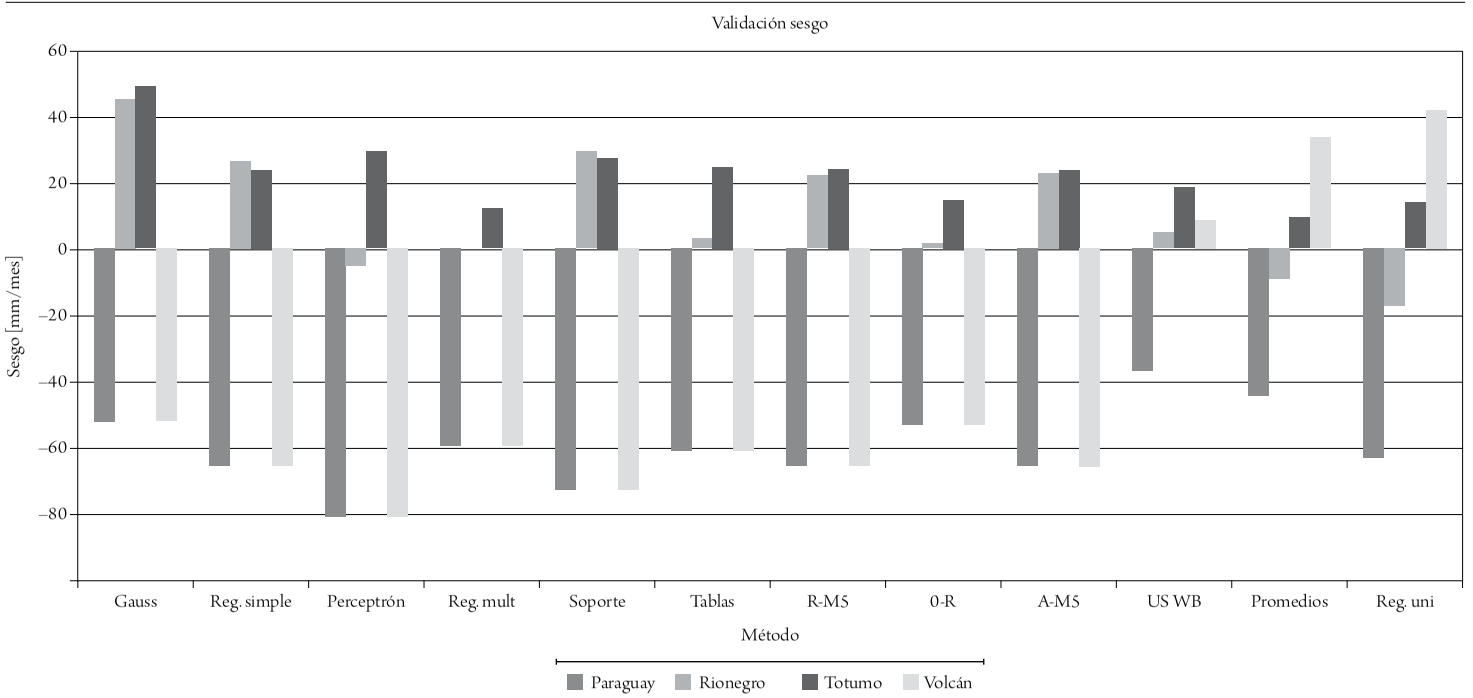

Figura 12. **Sesgo en validación**

Fuente: elaboración propia.

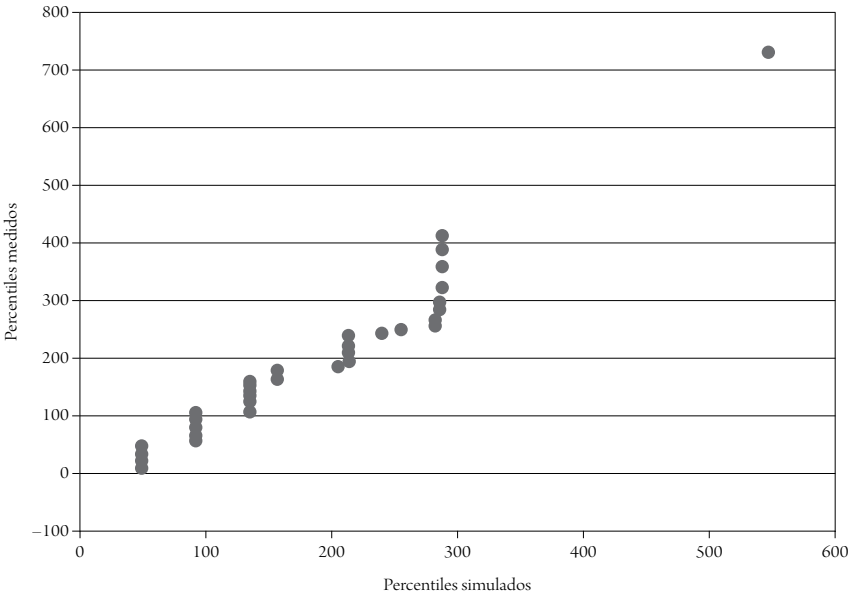

Figura 13. **Probabilidad de percentiles simulados y medidos usando tablas de decisión (Paraguay)**

Fuente: elaboración propia.

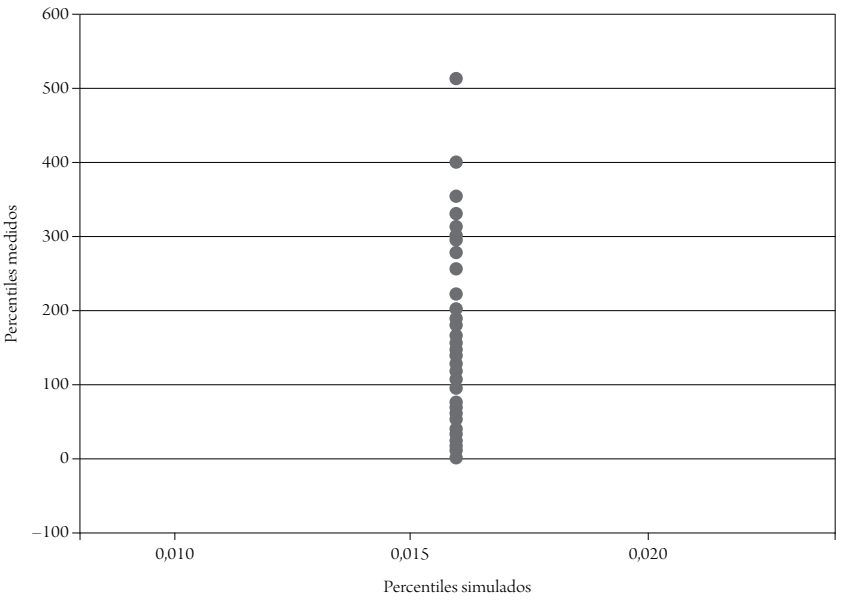

Figura 14. **Probabilidad de percentiles simulados y medidos usando 0-R (Paraguay)**

Fuente: elaboración propia.

épsilon, nº 25 • julio-diciembre 2015 • pp. 11-38 • ISSN 1692-1259-ISSN-E 2389-8828

Por otra parte, los resultados de los procesos de Gauss, especialmente al completar la serie de datos de la estación Volcán (figura 15), presentaron significativas desviaciones en el análisis de cuantiles. Estas desviaciones en los procesos de Gauss reflejan características de datos que no corresponden a los normalmente distribuidos, siendo esta una de las suposiciones del método. La corrección de esta deficiencia se puede obtener transformando los datos previamente al proceso de entrenamiento de los modelos de aprendizaje.

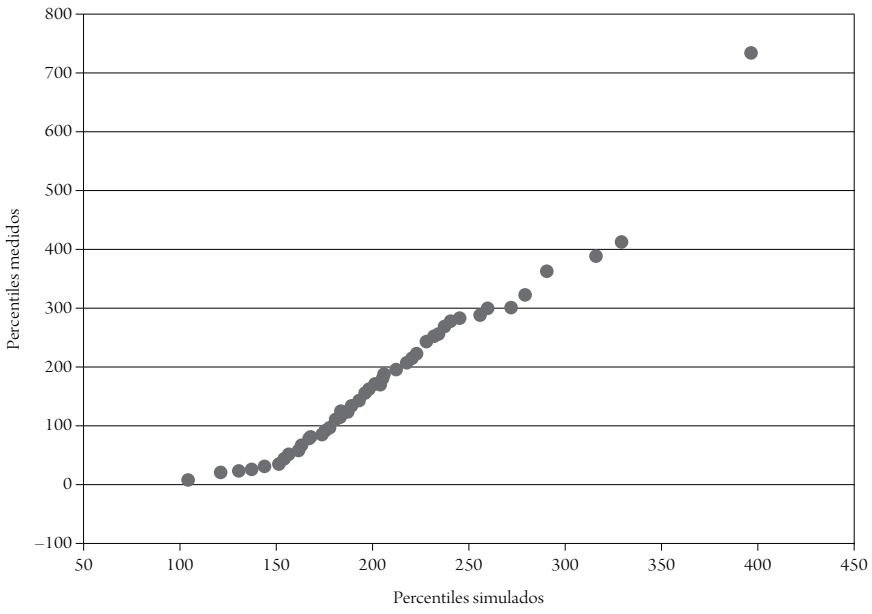

Figura 15. **Probabilidad de percentiles simulados y medidos usando procesos de Gauss (estación Volcán)**

Fuente: elaboración propia.

Los resultados de completar la serie de datos utilizando modelos de árboles de regresión M5 (figura 16) también presentaron consistentes subestimaciones en años considerados como húmedos. Estas reglas al parecer son heredadas de los intervalos de clases dictados a través de los árboles de decisión tal como se puede observar en la figura 13. Estos vacíos indican la sobregeneralización del modelo en estas áreas, lo cual lo hace especialmente susceptible a ruido de series externas. En este sentido, es necesario reevaluar las reglas para la asignación de las ramas finales de los árboles de decisión, antes de aplicar las regresiones en los nodos finales.

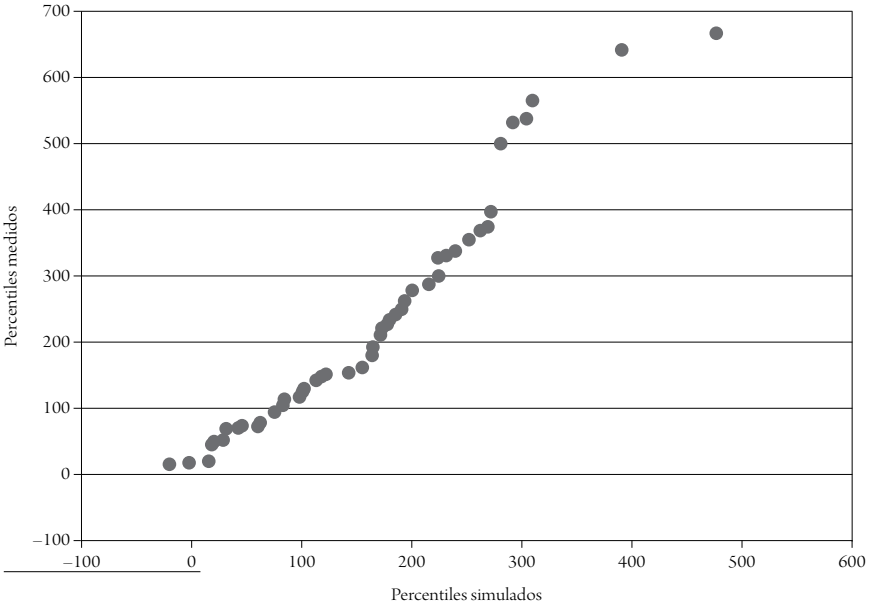

Figura 16. **Probabilidad de percentiles simulados y medidos usando AM5 (estación Volcán)**

Fuente: elaboración propia.

Como se estableció, los resultados obtenidos del perceptrón están lejos de ser aceptables, debido a su pobre desempeño. Este comportamiento se debe en especial a una inadecuada selección de los hiperparámetros, los cuales no pueden ser parcialmente reajustados durante el proceso de entrenamiento del método. Como consecuencia de ello, de un análisis de cuantiles de un modelo conceptualmente deficiente, difícilmente se obtendrán resultados satisfactorios. En este sentido, se sugiere que el método sea ajustado y validado antes de proceder con los análisis posteriores.

Finalmente, es posible afirmar que los métodos tradicionales presentan un comportamiento estadístico adecuado para completar series de datos. Respecto a los cuantiles, los simulados por lo general son mayores a los medidos, y existe un ligero sesgo entre las funciones de distribución de probabilidad.

Épsilon, Nº 25 • julio-diciembre 2015 • pp. 11-38 • ISSN 1692-1259-ISSN-E 2389-8828

Conclusiones y recomendaciones

Los métodos de aprendizaje de máquina son una buena una alternativa a los métodos convencionales para completar series de datos perdidos. Esto se puede evidenciar en los resultados obtenidos en comparación con métodos convencionales, considerando el hecho de que los detalles en el uso de estos métodos están lejos de ser refinados, y por tanto una más delicada parametrización, al igual que una adecuada selección de los hiperparámetros, podrá repercutir en mejores resultados.

Por otra parte, el entrenamiento de los algoritmos se debe llevar a cabo no solamente sobre la serie de datos (tal como se realizó en este estudio), sino también sobre los pliegues de esta (*folds*). Ello permitirá reducir los efectos de la sobreparametrización en los modelos, especialmente evidente en el uso del perceptrón multicapa. Con el uso de esta aproximación, se espera reducir las discrepancias obtenidas entre el entrenamiento y la validación.

Con respecto a la selección de los periodos de validación y entrenamiento, es posible establecer que se requiere del uso de diferentes técnicas que permitan entrenar las series de manera más regular. Con este propósito, se sugiere utilizar combinaciones aleatorias de las series de tiempo, de tal modo que sea posible obtener una distribución más uniforme de los periodos de entrenamiento y posterior validación, para lograr así series más homogéneas.

Con los resultados obtenidos se llega también a concluir que es necesario realizar una evaluación puntual de los datos, sobre todo los referentes a las estaciones Paraguay y Volcán. Esto se debe a que los resultados de las regresiones se encuentran muy distantes de los obtenidos en los otros métodos, sin importar el método aplicado, a pesar de que las correlaciones entre las distintas estaciones son fundamentalmente similares. Como consecuencia, se sugiere el uso de métricas no lineales de correlación (como información mutua o entropía) para verificar las suposiciones de ajustes entre las series de datos.

En general, los resultados muestran consistencia estadística en el análisis de los cuantiles, a excepción de los mostrados en el análisis de resultados. Con esto es posible establecer la viabilidad de completar series de datos a través de cualquiera de los métodos previamente descritos, dado que no solo el número de datos para completar es bajo en comparación con el número de datos disponibles, sino que

además las funciones de probabilidad de los datos medidos y simulados son considerablemente homogéneas.

Referencias

Consejo Municipal de Palermo Huila. (2013). Acuerdo Municipal 14 de 2013, por medio del cual se adopta la revisión general para la reformulación del plan básico de ordenamiento territorial del municipio de Palermo-Huila. Palermo, Huila, Colombia.

Cristianini, N. y Shawe-Taylor, J. (2000). *An introduction to support vector machines.* Cambridge: Cambridge University Press.

Edben, M. (2008). *Gaussian prosses for regression: a quick introduction.* Oxford: Oxford University.

Frank, E., Wang, Y., Inglis, S., Holmes, G. y Witten, I. (1997). *Using model trees for classification.* Hamilton, New Zealand: Department of Computer Science, University of Waikato.

Helsel, D. y Hirsch, R. (1992). *Statistical methods in water resorces.* Elsevier.

Journel, A. y Huijbregts, C. (1978). *Mining geostatistics.* Londres: Academic Press.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME-Journal of Basic Engineering, 82*, 35-45.

Linsley, K., Paulhus, J. y Kohler M. (1975). *Hydrology for engineers.* Nueva York: McGraw-Hill.

MacKay, D. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation, 11*(5), 1035-1068.

Mitchell, T. (1997). *Machine learning.* Nueva York: McGraw-Hill.

Morán, W. C. (1989). *Hidrología para estudiantes de Ingeniería civil.* Lima: Pontificia Universidad Católica del Perú.

Paulhus, J. L. y Kohler, M. A. (1952). *Interpolation of missing precipitation records.* Washington D. C.: Monthly Weather Review, Hydrologic Services Division, U.S Weather Bureau.

Rasmussen, C. y Williams, C. (2006). *Gaussian processes for machine learning.* Massachusetts: MIT Press.

Svozil, D., Kvasnicka, V. y Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems, 39*(1), 43-62.

Wang, Y. y Witten, H. (1996). *Introduction of model trees for predicting continuous classes.* Hamilton, New Zealand: Department of Computer Science, University of Waikato.

Wikipedia (2015). *Río Baché.* Recuperado de https://es.wikipedia.org/wiki/R%C3%ADo_Bach%C3%A9