Investigaciones Geográficas • Instituto de Geografía • UNAM
eISSN: 2448-7279 • DOI: dx.doi.org/10.14350/rig.59837 • ARTÍCULOS
Núm. 99 • Agosto • 2019 • e59837
www.investigacionesgeograficas.unam.mx

Evaluación de desempeño de métodos de relleno de datos pluviométricos en dos zonas morfoestructurales del Centro Sur de Chile

Performance Evaluation of Rainfall Data Fill-in Methods in Two Morphostructural Areas of South-Central Chile

Jenny Sofía Gómez Guerrero* y Mauricio Iván Aguayo Arias**

Recibido: 16/01/2019. Aprobado: 02/05/2019. Publicado en línea: 20/06/2019

Resumen. Una de las principales preocupaciones de los científicos al trabajar con datos temporales es la calidad de la información. Los datos meteorológicos, que son entradas de modelos y predicciones hidroclimáticas, generalmente carecen de series completas. El uso de técnicas de relleno frecuentemente ignora las características orográficas del área de estudio y la precisión del método, produciendo alteraciones en los resultados con importantes consecuencias. El objetivo de este trabajo es evaluar los métodos de relleno de datos pluviométricos razón normal y modelo de regresión lineal (LRM, por sus siglas en inglés), por medio de un análisis del error de estimación aplicado a un registro de 32 años de precipitaciones en dos unidades morfoestructurales distintas localizadas en la región del Biobío, Centro Sur de Chile: la planicie costera y el valle central. Los resultados evidenciaron que el método de Razón Normal presenta menor variabilidad en los errores de estimación y una mejor aproximación a los datos reales para ambas zonas.

Palabras clave: relleno de series pluviométricas, evaluación de desempeño, zonas morfoestructurales, razón normal, regresión lineal.

Abstract. The quality of the information in meteorological data time series has always been a concern for the scientific community. The scarcity of information requires the use of data fill-in techniques and methods that frequently ignore the orographic features of the study area, as well as the method accuracy, leading to inaccurate results with important consequences.

In this context, this paper seeks to evaluate two methods for filling rainfall data, namely Normal Ratio and Linear Regression Model (LRM), applied to two morphostructural zones in the south central region of Chile, through an error analysis of a 32-year series of precipitation data.

Both methods were compared considering 65 of 112 stations across the region, located on the coastal plain and central valley. Subsequently, two time-consistent base stations were defined, one for each area; pluviometric and proximity criteria, as well as the amount of information available, were applied to choose five neighboring stations.

After calculating the correlation between stations, using a probability analysis by quartiles and the Shapiro-Wilk test the normality of the LRM models was confirmed, as well as the homogeneity of the adjusted predictions and residuals.

* Facultad de Ciencias Ambientales, Universidad de Concepción (Chile). Víctor Lamas 1290, Concepción, Chile. Email: jennygomez@udec.cl, js.academi@gmail.com. Autor de correspondencia.
** Facultad de Ciencias Ambientales. Universidad de Concepción (Chile). Víctor Lamas 1290, Concepción, Chile. Email: maaguayo@udec.cl

The Normal Ratio method evaluated rainfall estimates by weighting mean annual rainfall in the neighboring stations, where each weighting factor corresponds to the ratio between the precipitation figure recorded in the auxiliary station and the mean annual rainfall of the respective station.

The performance of each method was assessed using the following estimators: Mean Error, Coefficient of Determination (CoD), Mean Squared Error (MSE), Root-Mean-Square Error (RMSE), Sum of Squared Residuals (SSR), Mean Relative Error (MRE), and Mean Absolute Percentage Error (MAPE).

The statistical analysis reveals a greater range of temporal variation in precipitation in the Central Valley relative to the Coastal Zone, except for one station, and a positive relationship between altitude and a broader pluviometric range. LRM shows greater data dispersion at station Chiguayante; moreover, according to the CoD, this is the station with the lowest prediction potential.

In most of the cases analyzed, we found an inverse relationship between the sum of squared residuals (SSR) and the number of annual precipitation data available in each station.

The estimators SSR, MSE, and RMSE penalize large residuals, revealing that for the 32-year series studied, The Normal Ratio yields better performance and lower prediction error in the target stations in both morphostructural areas, with Dichato as the station with the lowest mean error and Mayulermo as the station with the lowest mean relative error, for both methods in the sample selected.

As Dichato was the station with the greatest Euclidean distance from the base, the distance is discarded as a major predictive factor, contrary to our findings regarding data dispersion.

The analysis of residuals (SSR, MSE, RMSE) indicated that the Linear Regression Model is influenced by outliers. However, these values were considered, since eliminating the extreme values, as is usually done in regression analysis, may result in losing relevant information about maximum and minimum precipitation that is useful in the analysis of extreme climatic events such as drought. The efficiency of both methods for predicting actual values was evaluated through the estimators SSR and CoD, showing that in the present analysis, the Normal Ratio involves a higher CoD and a lower residual variability. Although regression remains a widely used and recommended method, the Normal Ratio should be reconsidered for the prediction of missing data in precipitation series in areas of south central Chile with records available for neighboring stations that could support the equation for the data required.

The quadratic estimators MSE and RMSE allow inferring that those stations showing a lower mean error, where the predictive methods analyzed were most successful, were the stations where precipitation showed a more stable behavior around the mean.

The dimensionless estimators MRE and MAPE confirmed the advantage of the Normal Ratio and determined that the best mean performance of the prediction was related to data dispersion rather than to the Euclidean distance between stations and the base station.

The two methods evaluated offer a simple way to estimate meteorological data when the information available is insufficient; however, the Normal Ratio demonstrated a better performance relative to LRM for estimating missing precipitation data, regardless of the geomorphological area selected.

Keywords: Fill of rainfall data, performance evaluation, morphostructural zones, Linear Regression, Normal Ratio.

INTRODUCCION

El estudio de eventos climatológicos extremos, el diseño de obras hidráulicas y la modelización hidrológica y climatológica, entre otras aplicaciones, requieren como datos de entrada series temporales de precipitación. Es frecuente encontrar series de precipitación incompletas lo que dificulta su utilización en modelos e índices para la caracterización hidrológica o climatológica de un determinado lugar.

El objetivo principal del relleno de datos es estimar la precipitación de los registros faltantes con el menor error posible respecto de su ocurrencia real. Dentro de las técnicas para relleno de datos pluviométricos existe una amplia variedad; sin embargo, la elección del método debería depender de las características geomorfológicas del área de estudio. Para la estimación de precipitaciones faltantes generalmente se usan métodos de ponderación tradicionales como el Inverse Distance Weighting Method (IDWM) (ASCE,1996 citado en Toro, Arteaga, Vázquez e Ibáñez, 2015), el cual consiste en el cálculo de la precipitación de un punto desconocido con base en el promedio ponderado de los valores conocidos dentro del vecindario de este punto, usando como ponderación el inverso de la distancia al punto conocido (Lu y Wong, 2008) y métodos basados en datos (Teegavarapu, 2009; Teegavarapu y Chandramouli, 2005). Algunos de los métodos tradicionales basados en la distancia han sido cuestionados pues no tienen en cuenta la posibilidad de que exista auto correlación espacial negativa entre observaciones ubicadas en distintas regiones topográficas (Teegavarapu y Chandramouli, 2005).

Por otra parte, algunos de los métodos basados en datos como el promedio aritmético (Pizarro, Ramirez y Flores, 2003), conocido como "Station-Average Method", es práctico y simple, pero su precisión se ve afectada cuando la diferencia en los promedios anuales de las estaciones regionales difiere en más del 10% de la captura anual en la estación de interés (McCuen, 1998). Esto es especialmente difícil de encontrar en zonas montañosas donde existen efectos orográficos de la elevación en las mediciones; por tanto, estos métodos no son adecuados para regiones montañosas (Tung, 1983). También se encuentran otros métodos, como la regresión lineal, ampliamente utilizado en Chile (Pizarro, Ausensi, Aravena, Sangüesa, León y Balocchi, 2009) y en Sudamérica (Luna y Lavado, 2015), desde su recomendación en la Guía para la Elaboración del Balance Hídrico de América del Sur (UNESCO-ROSTLAC, 1982) y los métodos basados en análisis de series de tiempo.

Chile es un país caracterizado por tener diversidad de regímenes climáticos debido a sus límites geográficos naturales, su extensión, condiciones orográficas y cercanía al mar. Estas condiciones favorecen una amplia gama de climas, como el desierto, estepa, mediterráneo, templado lluvioso, oceánico, tundra y polar. Además, se observan diferencias en el régimen de lluvias, dependiendo de la ubicación del área de estudio con respecto a la cordillera costera y andina (Sarricolea, Herrera y Araya, 2013). La zona centro-sur, por ejemplo, posee un relieve con cordones montañosos y valles que dividen longitudinalmente el territorio, lo que modifica los patrones de precipitación e influye en la presencia de variabilidad climática entre zonas relativamente cercanas.

Actualmente, la información meteorológica disponible en las bases de datos de acceso público del país está distribuida desigualmente en el espacio y se encuentra una gran cantidad de datos faltantes que hace difícil el análisis de series temporales.

Al ser la precipitación la variable más importante que alimenta los procesos hidrológicos subsecuentes (Fekete, Vörösmarty, Roads y Willmott, 2004), cualquier error sistemático o aleatorio en su medición, así como en la estimación de sus datos, tiene gran relevancia en los resultados posteriores de los modelos y balances hidrológicos. La estimación de datos faltantes es, por tanto, una de las tareas más importantes requeridas en muchos estudios de modelización hidrológica y climatológica (Lu y Wong, 2008; Teegavarapu y Chandramouli, 2005) El objetivo de este trabajo es realizar un análisis comparativo de dos métodos de relleno de datos meteorológicos: la regresión lineal y el método de razón normal, utilizando series anuales de al menos 30 años de precipitación, en dos zonas morfoestructuralmente distintas de la Región del Biobío[1] y Ñuble del centro-sur de Chile: la Planicie Litoral o "Zona Costera", y la depresión intermedia o" Valle Central", por medio de un análisis de desempeño y la comparación y análisis de siete estimadores de error.

METODOLOGÍA

Área de estudio

La Región del Biobío es una región que se extiende desde los 36° 00' a los 38° 30' S. Ocupa 37 mil km^2 y es la segunda más poblada de Chile (Figura 1). Esta región se encuentra ubicada en la zona centro-sur del país y posee un clima mediterráneo con estación húmeda prolongada. Sin embargo, existen diferencias climáticas a nivel local originadas principalmente por cambios en la latitud, características físicas del terreno y distancia al mar. Las precipitaciones anuales, por ejemplo, presentan en el litoral variaciones entre 700 y 1200 mm, en la zona intermedia entre 950 y 1500 mm y en la zona andina y precordillerana, sobre los 1400 mm (DMC, 2001). Los meses más lluviosos se registran entre mayo y agosto. Longitudinalmente, el centro-sur de Chile presenta cuatro unidades morfoestructurales de relieve conocidas como "macroformas", organizadas de oeste a este: Planicie Litoral, Cordillera de la Costa, Depresión intermedia o Valle Central y Cordillera de Los Andes.

La orografía del centro-sur de Chile altera los patrones de precipitación de manera significativa,

[1] En adelante, Región del Biobío hará referencia a las regiones Biobío y Ñuble juntas, según la división política de Chile vigente hasta el año 2018 (Figura 1).

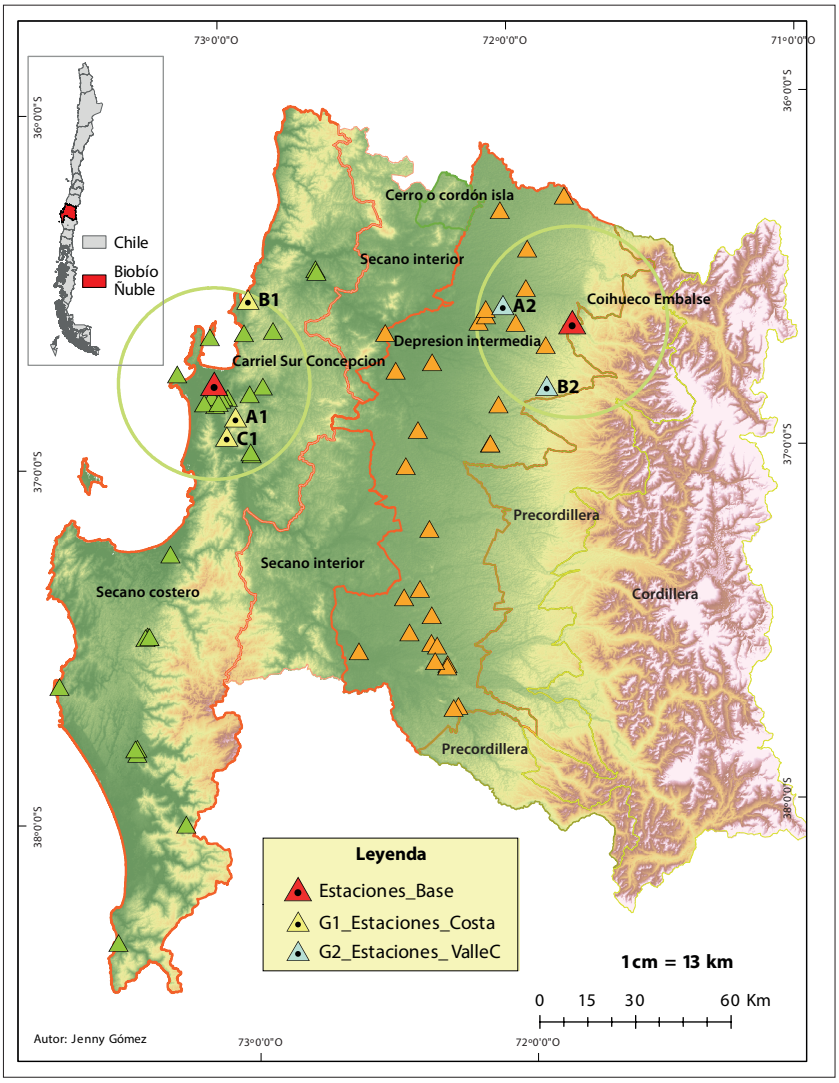

Figura 1. Selección de estaciones pluviométricas en Regiones del Biobío y Ñuble, en las Zonas morfoestructurales Costera y Valle Central. Fuente: elaboración propia con datos DGA y ODEP.

aumentando en la zona de precordillera de los Andes, la cantidad de lluvia registrada bajo su influencia, incluso hasta en el doble con respecto a la zona costera (Sarricolea, Herrera y Araya, 2013), razón por la cual se ha decidido evaluar los métodos de relleno de datos meteorológicos en dos subgrupos (Figura 1): 1) la planicie litoral (en adelante, Zona Costera-Región con estaciones color verde) y 2) la depresión intermedia, en adelante, Zona del Valle Central-Región con estaciones en color naranja. Esto con el fin que la aplicación de los métodos se efectúe en zonas relativamente homogéneas y pluviométricamente comparables, así como también permita hacer un análisis independiente en ambas zonas de interés.

Fuente de datos

Los datos pluviométricos fueron seleccionados teniendo en cuenta criterios de homogeneidad en las precipitaciones basados en dos zonas ambientales de comparación diferenciadas por su morfología: la Zona Costera (G1) y el Valle Central (G2) (Figura 1). Estos datos corresponden a precipitaciones acumuladas mensuales de estaciones pluviométricas de la Dirección General de Aguas (DGA) y la Dirección Meteorológica de Chile (DMC),

recopilados por el Centro de Ciencia del Clima y la Resiliencia (CR2), (CR2, 2016).

Usando dichas zonas como límites ambientales, se calcula una nueva base de datos de precipitación acumulada anual a partir de las estaciones pluviométricas en la planicie costera y valle central. Para este estudio, se contó con un universo de 112 estaciones en la Región del Biobío. El periodo de análisis se determinó por los años hidrológicos de precipitación acumulada anual comprendidos entre el 1º de abril de un año y el 31 de marzo del año siguiente, entre 1983 al 2015, para un total de 32 años empleados. En un primer filtro se descartaron las estaciones ubicadas fuera de las dos regiones morfoestructurales anteriormente definidas y aquellas que no se encontraran activas durante el periodo de estudio, y se obtuvo un total de 65 estaciones. Además, se usó un criterio de proximidad geográfica y una adaptación del método de análisis pluviométrico de dos o más estaciones, propuesto por Martínez de Azagra y Navarro (2007), para establecer la selección de estaciones en cada uno de los subgrupos, con centro en una estación base. Dicha estación se procura que tenga en lo posible un registro completo y que mantenga consistencia en los datos (Benitez, 1998). Las estaciones base seleccionadas para G1 y G2 se muestran en la Figura 1. Cada uno de los grupos corresponde a estaciones localizadas en un radio de 30 km de distancia de la estación base correspondiente. El listado de estaciones que cumplen estos criterios se presenta en los cuadros 1 y 2.

El universo de estaciones de análisis se redujo a un total de 28 estaciones y dos estaciones base. Como último requerimiento se definió que las estaciones tuviesen al menos un 90% de los datos, no obstante, la carencia de información encontrada en las estaciones de ambas zonas requirió ampliar este criterio hasta aproximadamente un 30%, teniendo en cuenta que, a mayor porcentaje de información faltante, mayor es el error de estimación acumulativo de la serie meteorológica. Se obtienen,

Cuadro 1. Estaciones vecinas (19) a la estación base 1* (Carriel Sur CCP) en Zona Costera, y porcentajes de información faltante en el periodo 1983-2015.[1] Promedio de 32 años de precipitación.[2]

N°	Fuente	Nombre de la estación	Inicio de la obs.[1]	Fin de la obs.[1]	N° de obs.	Promedio[2] (mm)	%Datos F
1	DMC	Carriel Sur CCP*	abr-83	mar-15	384	86.5	0%
2	DMC	Punta Hualpén Faro	abr-83	nov-93	128	53.83	66.67%
3	DGA	Concepción DGA	ene-96	mar-15	231	96.61	39.84%
4	DGA	Rio Biobío en Desembocadura	abr-01	mar-15	141	73.62	63.28%
5	DGA	Cerro Verde	dic-89	dic-95	68	103.64	82.29%
6	DMC	Cerro Caracol	abr-83	mar-15	292	87.93	23.96%
7	DMC	Bellavista Univ. de Concepción	abr-83	dic-92	36	116.31	90.63%
8	DGA	Concepción Edif. Mop	jul-00	abr-06	70	114.25	81.77%
9	DGA	Estero Nonguen Frente UBB	may-09	mar-15	66	91.80	82.81%
10	DMC	ChiguayanteEssbio	abr-83	dic-09	312	104.60	18.75%
11	DMC	NonguenEssbio	abr-83	mar-15	365	109.23	4.95%
12	DGA	Andalien	nov-00	mar-15	168	94.22	56.25%
13	DGA	Las Pataguas	ene-93	mar-15	241	82.72	37.24%
14	DGA	Rafael	ene-93	mar-15	257	111.58	33.07%
15	DGA	Estero Bellavista en Tome	jun-09	mar-15	58	76.25	84.90%
16	DMC	Isla Quiriquina Faro	abr-83	oct-84	19	93.44	95.05%

Cuadro 1. Continúa.

N°	Fuente	Nombre de la estación	Inicio de la obs.[1]	Fin de la obs.[1]	N° de obs.	Promedio[2] (mm)	%Datos F
17	DGA	Dichato	abr-83	mar-15	374	80.48	2.60%
18	DMC	Hualqui Sendos	abr-83	dic-91	103	121.98	73.18%
19	DGA	Estero Hualqui en Desemb.	may-09	mar-15	48	78.99	87.50%
20	DMC	Punta Tumbes Faro	Inact-Dic78	N/A	N/A	N/A	100.00%

Fuente: elaboración propia. Cálculos con base en datos DGA y DMC tomados de Base de Datos CR2 (2016).

Cuadro 2. Estaciones vecinas (9) a la estación base 2* (Coihueco Embalse) en Valle Central y porcentaje de información faltante en el periodo 1983-2015.[1]

N°	Fuente	Nombre de la estación	Inicio de la obs.[1]	Fin de la obs.[1]	N° de datos	Promedio (mm)	%Faltante
1	DGA	Coihueco Embalse*	Apr-83	mar-15	384	118.3	0.00%
2	DMC	Bernardo O'Higgins Chillan AD.	Apr-83	mar-15	357	88.76	7.03%
3	DGA	Mayulermo	feb-92	mar-15	275	123.57	28.39%
4	DGA	Chillan Viejo	Apr-83	mar-15	383	84.99	0.26%
5	DMC	San Carlos Sendos	Apr-83	jul-03	241	90.11	37.24%
6	DMC	Pinto Municipalidad	mar-99	Dec-08	106	122.71	72.40%
7	DMC	Instituto Profesional Adv. Chillan	Jan-86	Dec-95	91	102.99	76.30%
8	DGA	Canal de la Luz en Chillan	sep-08	mar-15	71	57.09	81.51%
9	DMC	Santa Rosa de Cato	Inact-Dec80	N/A	N/A	N/A	100%
10	DGA	Chillan Sendos	Inact-Dec82	N/A	N/A	N/A	100%

Fuente: elaboración propia. Cálculos en base a datos DGA y DMC tomados de Base de Datos CR2 (2016)

finalmente, los grupos con las estaciones DGA y DMC (Figura 1):

G1:
- Carriel Sur Concepción (Base 1)
- ChiguayanteEssbio
- NonguénEssbio
- Dichato

G2:
- Coihueco Embalse (Base 2)
- Bernardo O'Higgins Chillan AD.
- Mayulermo

Esta selección de estaciones puede ser corroborada por la alta correlación entre los datos de las estaciones vecinas y la estación base (Cuadro 3).

Una descripción general de las estaciones seleccionadas se muestra en el Cuadro 4.

Método de regresión lineal (LRM)

Este método, que ha sido descrito ampliamente y con diversidad de aplicaciones (Kazmier, 1998; Luna y Lavado, 2015; McCuen, 1998; Mendenhall, 1990; Rojo, 2007), permite obtener una ecuación que describe el comportamiento de dos variables diferentes en función de los datos obtenidos. Esta ecuación establece una relación bivariada que es útil para inferir datos desconocidos en la variable de interés. Para su aplicación se deben verificar o

Cuadro 3. Coeficientes de correlación entre las estaciones vecinas y la estación base, para cada zona. Etiquetas conforme a la Figura 1.

Estación base	Estación vecina	Coef. de correlación
G1. Carriel Sur Concepción	A1. NonguenEssbio	0.82
	B1. Dichato	0.84
	C1. ChiguayanteEssbio	0.73
G2. Coihueco Embalse	A2.Bernardo O'Higgins Chillan AD.	0.89
	B2. Mayulermo	0.94

Fuente: elaboración propia.

Cuadro 4. Ubicación, elevación y precipitación media anual (PMA) para las estaciones vecinas G1 y G2.

Estación	Fuente	Latitud (S)	Longitud (W)	Elevación (msnm)	PMA (mm)
Carriel Sur	DMC	-36.7792	-73.0622	12	86.5
ChiguayanteEssbio	DMC	-36.9294	-73.0267	39	104.60
Dichato	DGA	-36.5456	-72.9311	11	80.48
NonguénEssbio	DMC	-36.8769	-72.9931	126	109.23
Coihueco Emb.	DGA	-36.6408	-71.7989	314	118.30
Bernardo O'Higgins	DMC	-36.5872	-72.04	151	88.76
Mayulermo	DGA	-36.8189	-71.8944	371	123.57

Fuente: elaboración propia con base en Datos DGA y DMC.

asumir ciertos supuestos sobre los datos como lo son la normalidad, la homogeneidad de varianzas y la independencia. La ecuación que se describe en la regresión simple corresponde a la relación lineal:

$$y = \beta_0 + \beta_1 x + u$$
$$E(y|x) = \beta_0 + \beta_1 x \,; u = 0 \quad (1)$$

En la ecuación *y* es función lineal de *x,* y *u* es el término independiente asociado a la perturbación de todos los factores aleatorios exógenos distintos a *x,* que inciden en la variable dependiente. *E* denota el valor esperado de *y* para un valor determinado de *x*, β_0 es el intercepto y β_1 corresponde al parámetro estimado de la variable *x*.

La ecuación lineal que arroja el modelo permite estimar precipitaciones en las estaciones con datos faltantes a partir de una estación de referencia.

En la aplicación del modelo LRM se debe comprobar la consistencia de los modelos por medio de la verificación de algunos supuestos. En nuestro caso se tomó como supuesto la no colinealidad de las estimaciones y se comprobó la normalidad de los modelos. Para esto se aplicaron los métodos Q-Q plot que permite el análisis de probabilidad por cuartiles y el método Shapiro Wilk, igualmente se realizaron test para comprobar la homogeneidad de residuos.

Método de la razón normal

Este método consiste en un enfoque de estimación basada en datos que calcula la precipitación como una ponderación de la precipitación anual promedio en un cierto número de estaciones vecinas (Aparicio, 2004; Carrera-Villacrés, Guevara-García, Tamayo-Bacacela, Balarezo-Aguilar, Narváez-Rivera y Morocho-López, 2016; Linsley,

Kohler, y Paulhus, 1958; Monsalve, 2009). Se diferencia del modelo lineal en que no requiere que los residuos cumplan con supuestos de normalidad u homogeneidad.

Se expresa mediante la siguiente ecuación (2):

$$P_x = \frac{1}{n}\left[\left(\frac{N_x}{N_1}\right)P_1 + \left(\frac{N_x}{N_2}\right)P_2 + \ldots + \left(\frac{N_x}{N_n}\right)P_n\right] \quad (2)$$

Donde *Px* corresponde a la altura de la precipitación faltante en la estación en estudio. Los pesos que se usan en las ponderaciones corresponden a proporciones entre la captura de precipitación del medidor en la estación auxiliar *i* (Pi) y la precipitación media anual (Ni) correspondiente.

Error de estimación y análisis de desempeño

Un estimador es un estadístico que corresponde a una función de una variable muestral que asigna a un parámetro el valor (estimación) que toma la función en una muestra específica (Ruiz-Maya y Martín Pliego, 1999). Análogamente, para este ejercicio, las estimaciones corresponden a los valores experimentales de la precipitación. Para evaluar el desempeño de estos métodos de relleno de datos meteorológicos se determinó el mejor estimador de la variable meteorológica faltante usando un análisis de los residuos de las estimaciones teóricas.

El desempeño de los métodos estudiados en la predicción de los valores de precipitación en las estaciones vecinas (variable dependiente) se puede estimar en primera instancia con base en la relación entre el valor observado (y) y el valor ajustado o estimado ($\hat{y}$). Este se conoce como *error absoluto* (3) y corresponde al valor absoluto de la diferencia entre valor observado y el valor estimado.

$$e = Y - \hat{Y} \quad (3)$$

Debido a que es justamente dicha información (valor observado) de la cual el investigador no dispone, las comparaciones sobre la precisión y ajuste de los métodos de relleno de datos se realizaron en base a los valores no faltantes. En la sumatoria de los errores para el cálculo del error medio se considera el valor absoluto para evitar la anulación debida al signo. Un estimativo más exacto que no tiene en cuenta el signo del error corresponde a la media de la suma de las desviaciones al cuadrado de los residuos o error cuadrático medio (MSE).

El error cuadrático es un método para evaluar una técnica de elaboración de pronósticos donde cada residuo se eleva al cuadrado, por lo que este enfoque sanciona errores grandes, pero puede conducir a valores de error exagerados.

$$MSE = \frac{\Sigma(Y_t - \hat{Y}_t)^2}{n} \quad (4)$$

$$\text{RMSE} = \frac{1}{n}\sqrt{\sum_1^n(\hat{y}_i - y_i)} \quad (5)$$

Para esta evaluación se compararon el error promedio, la suma de las desviaciones al cuadrado (MSE) (4), la raíz del error cuadrático medio (RMSE) (5), la suma de los cuadrados de los residuos (SSR) (7), el coeficiente de determinación (8) y el error relativo medio (MRE) (9).

La suma de los cuadrados de los residuos SSR(7) corresponde a la variabilidad que no es explicada por el modelo, es decir la variabilidad de los residuos, la cual se quiere minimizar en la regresión (Wooldridge, 2013). Numéricamente:

$$SST = SSR + SSE \quad (6)$$

$$SSR = \sum_1^n e_i^2 \quad 7)$$

Donde SST es igual a la suma total de cuadrados (6) y corresponde a la variación total de la variable estimada respecto de la media y SSE es la variabilidad explicada por el modelo de regresión. El coeficiente de determinación (CoD) o R^2 (8) evalúa el grado de ajuste del modelo y se puede expresar como una función de los parámetros anteriores:

$$CoD = R^2 = \frac{SSE}{SST} = \frac{\sum_1^n(\hat{y}_i - \bar{y})}{\sum_1^n(y_i - \bar{y})} \quad (8)$$

$$0 \leq R^2 \leq 1$$

Adicionalmente, se estimó el error relativo

medio (MRE), con el fin de obtener una medida adimensional, que sea independiente de las unidades y del tamaño del error. Y un análogo a este estimador que se expresa en porcentaje: el error porcentual absoluto medio o MAPE.

$$MRE = \frac{1}{n}\sum_{1}^{n}\frac{\hat{y}_i - y_i}{y_i} \quad (9)$$

RESULTADOS

Análisis exploratorio

En un primer análisis exploratorio de los estadígrafos es posible apreciar un mayor rango de variación temporal para las precipitaciones anuales en el Valle Central respecto a la Zona Costera, excepto para la estación Nonguén (1427.9 mm), con amplitudes de 1644.4 mm para la estación Coihueco embalse, 1378.8 mm para Mayulermo y 986.87 mm para Bernardo O'Higgins versus 883.6 mm para Carriel sur, 1254.6 mm para Chiguayante y 668.24 mm para Dichato (Cuadro 5).

Debido a que los dos grupos se encuentran a poca diferencia latitudinal y a que no se descarta la influencia oceánica (oeste-este) en el valle central, podríamos encontrar una clara explicación en la variable altitudinal de estas estaciones, pues la de mayor rango (Coihueco) se encuentra a 314 msnm y la de menor rango (Dichato), se encuentra a 11 msnm Paralelamente, Nonguén, la estación con mayor intervalo de variación dentro de la Zona Costera, es la más alta en su grupo.

Las medias similares de las estaciones centrales Coihueco y Mayulermo, así como sus coeficientes de variación, revelan una gran similitud en el comportamiento de las precipitaciones.

Simulando modelos LRM

A partir del método LRM se obtienen los siguientes cálculos para los estimadores de la regresión.

Cálculo de los estimadores B de la regresión y coeficiente de determinación:

Los coeficientes B_1 del modelo LRM revelan el cambio medio en la precipitación anual en la estación de respuesta por unidad de cambio en la estación predictora (Cuadro 6) y corresponden a la pendiente de la recta de los modelos de regresión (Figura 2).

En la Figura 2 se observa que los puntos cubren equitativamente el rango de las estaciones base. Los valores altos de coeficiente de determinación representan, en general, un buen ajuste de las estimaciones a la variable real. Las rectas mínimo-cuadráticas muestran buen seguimiento de los datos, observándose, no obstante, un mayor rango de dispersión en los datos de Chiguayante. Para esta estación se obtiene el LRM con menor ajuste según el Coeficiente de determinación (CoD) y por tanto, el que tendría menor capacidad predictiva de la precipitación a pesar de tener un coeficiente de regresión cercano a 1.

El grupo 2, de Valle Central, presenta un rango más amplio de variación en las precipitaciones,

Cuadro 5. Estadísticos Básicos. Zona Costera (G1) y Valle Central (G2).

	Estaciones	Promedio	Desv. est.	Máx.	Mín.	C. de variación
G1	Carriel Sur	1037.92	261.79	1524.3	640.7	0.25
	Chiguayante	1194.16	327.67	1885.8	631.2	0.27
	Dichato	947.06	176.24	1263	594.76	0.19
	Nonguén	1310.14	295.48	1946.6	518.7	0.23
G2	Coihueco	1419.58	362.91	2360.1	715.7	0.26
	Bernardo O.	989.56	261.71	1494.12	507.25	0.26
	Mayulermo	1498.23	382.29	2154.1	775.3	0.26

Fuente: elaboración propia.

Cuadro 6. Estimadores de los modelos LMR para cada estación base.

Carriel Sur	Estación	Intercepto (Bo.)	Pendiente (B1)	CoD
lm1	Nonguén	269.593	1.018	0.6798
lm2	Dichato	361.4926	0.5738	0.7147
lm3	Chiguayante	239.0428	0.9101	0.5355
Coihueco	Estación	Intercepto (Bo.)	Pendiente (B1)	CoD
lm4	Bdo. O'Higgins	108.0501	0.6263	0.7495
lm5	Mayulermo	180.7482	0.9277	0.8929

Fuente: elaboración propia.

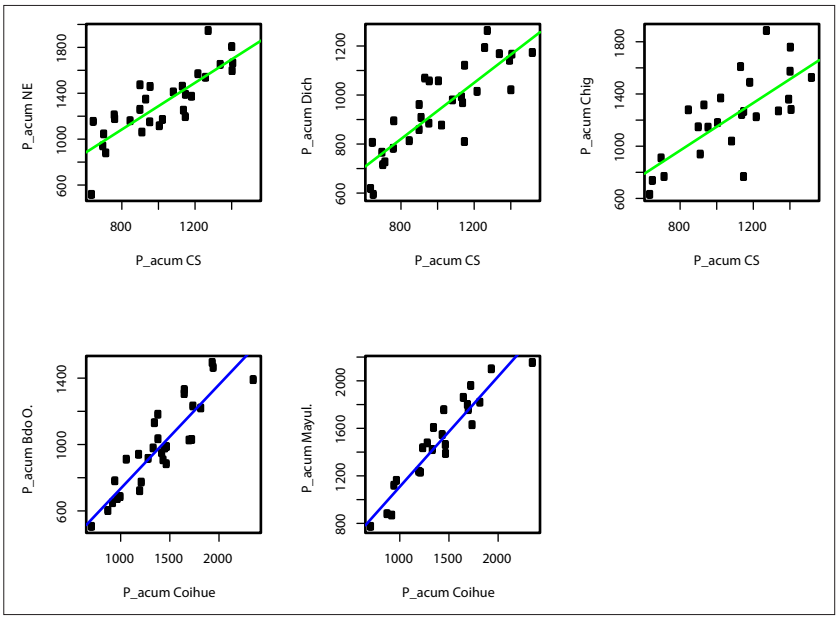

Figura 2. Modelos LRM dependientes de CS (Carriel Sur) en verde. Estaciones NE= NonguénEssbio, Dich= Dichato, Chig= Chiguayante. Modelos LRM dependientes de Coh (Coihueco Embalse) en azul. Estaciones Bdo. O.= Bernardo O'Higgins, Mayul= Mayulermo. Fuente: elaboración propia.

particularmente evidenciado en Coihueco, donde la precipitación máxima supera los 2000 mm y su rango de variación alcanza los 1644 mm. Destaca la relación de dependencia evidenciada entre las variables de la estación Mayulermo y Coihueco Embalse (Figura 2 y Cuadro 6).

Verificación de homogeneidad

Se verificó la homogeneidad de los residuos con respecto a los valores observados en los modelos (Figura 3). Las gráficas a lo largo del eje x permiten verificar la homogeneidad de la variable y no revelan presencia de tendencias que planteen serias inquietudes con respecto a la homogeneidad en las varianzas de las variables, lo cual es indicativo también de independencia entre las predicciones ajustadas y sus residuos.

Verificación de normalidad

Se evaluaron diagramas de probabilidad por cuartiles e histogramas para cada muestra (Figura 4).

Para esto se generó un conjunto de valores aleatorios con distribución normal para probar el ajuste de los modelos. Los resultados permiten inferir que, en general, las distribuciones de los residuos son aproximadamente normales. El único caso que plantea dudas de normalidad es el modelo lm4 (Coihueco-B. O'Higgins). El histograma y el

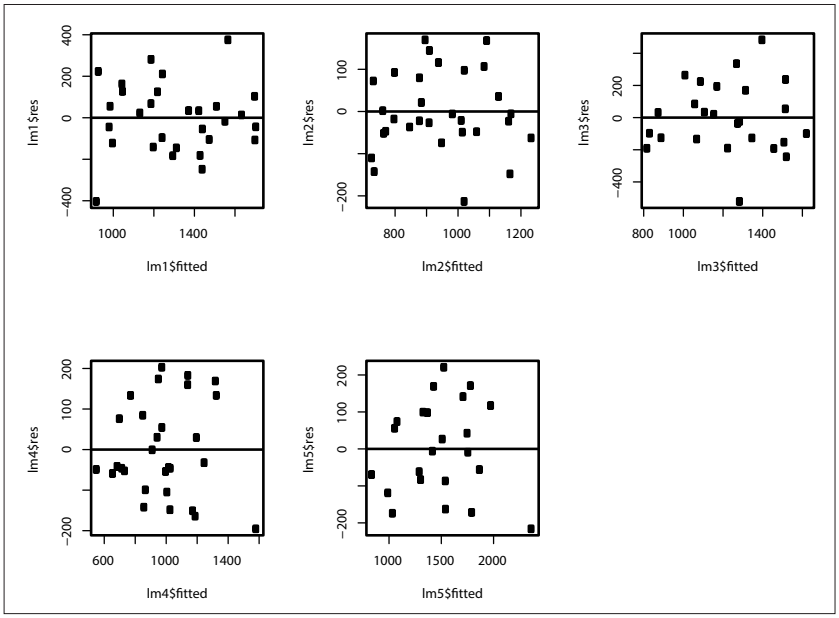

Figura 3. Test de homogeneidad de variables. Modelos lineales 1 al 5 (véase Cuadro 6). Eje *x*: modelos ajustados. Eje *y*: residuos. Fuente: elaboración propia.

gráfico de probabilidad por cuartiles muestra una menor aproximación normal en las colas de las distribuciones (Figura 4). El posterior test numérico Shapiro Wilk permite despejar esta duda.

Test Shapiro Wilk

Para un p–value < 0.05**,** los resultados del test indican que, en general, las distribuciones de las estaciones pueden ser catalogadas como normales; los p-values para cada uno de los modelos lo superan, por tanto, se acepta la hipótesis nula de la normalidad en las distribuciones (Cuadro 7).

RESULTADOS DE LOS ESTIMADORES

Para el manejo de los residuos fueron usados siete estimadores de error: SSR, RMSE, MSE, MRE, MAPE, CoD y error promedio en la regresión y el método de la razón normal.

En la estimación de datos de precipitación en dos zonas morfoestructurales de la región del Biobío en Chile se utilizaron dos estaciones base y cinco estaciones vecinas para comparar la regresión lineal y razón normal.

Se comparó el rendimiento de los métodos de relleno de datos pluviométricos y la bondad de ajuste de las estimaciones obtenidas para las cinco estaciones vecinas en cada grupo (G1, G2), tanto para la regresión como para el método de la razón normal obteniendo los resultados presentados en el Cuadro 8.

El cálculo del error utilizó datos no faltantes; no obstante, en el método de regresión lineal se encontró una relación inversamente proporcional entre el error cuadrático (SSR) y la cantidad de datos de precipitación anual (%Data P anual, Cuadro 8) disponibles en cada estación, por lo cual, para este método, a un mayor número de observaciones se evidencia una reducción en el error acumulado de predicción en la precipitación.

De manera similar ocurre para el método de la razón normal, con excepción de la estación Mayulermo, en la cual un menor número de estaciones conduce a un menor error. En el análisis a detalle de esta situación se ha observado que en dichas estaciones se presentaron algunas precipitaciones anormalmente altas o bajas con respecto a la media anual, por lo cual, en tales situaciones, el método de razón normal podría sobreestimar o subestimar la precipitación.

El estimador que da una idea del grado de aproximación media de los métodos al valor real de la precipitación, independientemente de las

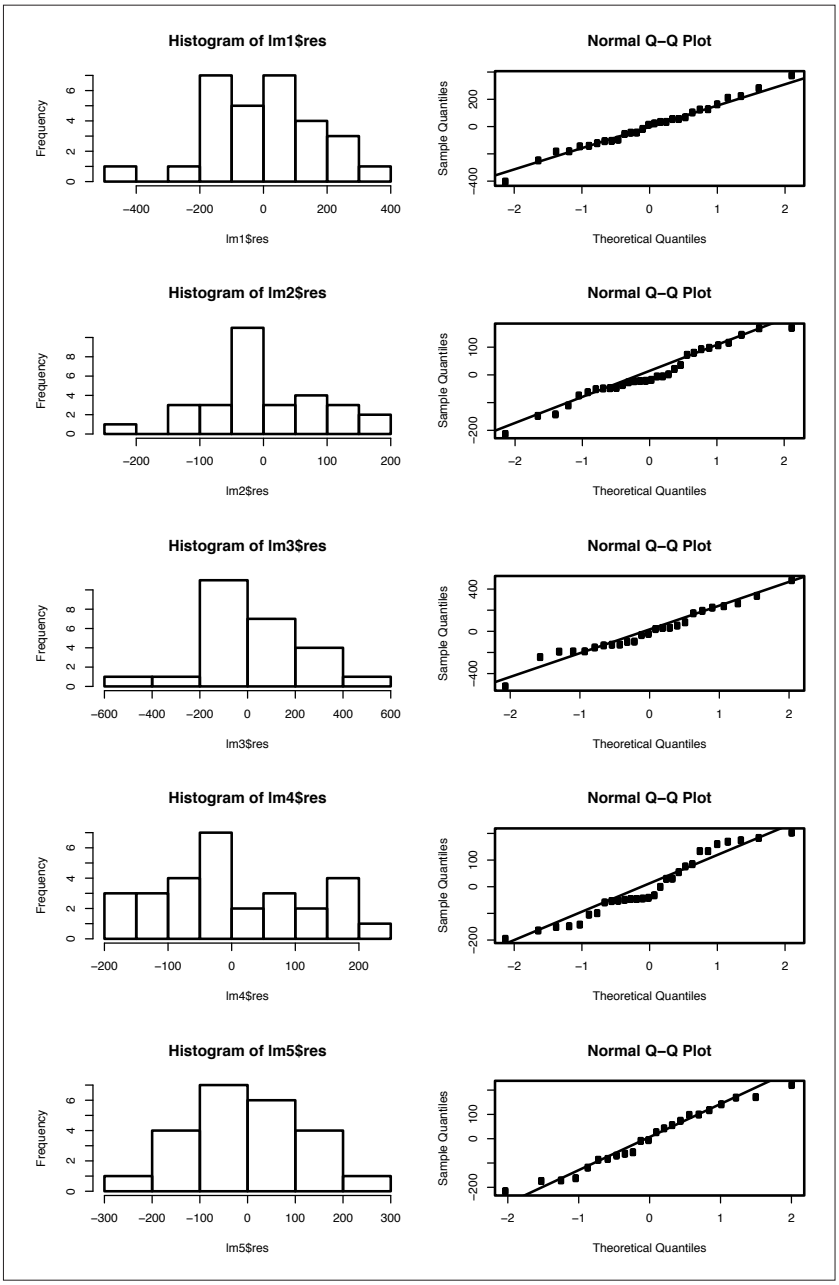

Figura 4. Análisis de Normalidad por método Q-Q Plot para 5 modelos LRM en G1 y G2. lm1 = Nonguén, lm2= Dichato, lm3= Chiguayante, lm4=B. O'Higgins, lm5=Mayulermo. Periodo 1983-2015. Fuente: elaboración propia.

unidades, es el error relativo medio (MRE) y el error porcentual absoluto medio (MAPE), el cual también ratifica un mejor comportamiento de razón normal frente a la regresión para ambos grupos (G1 y G2) y un menor error de estimación entre ambos grupos en la estación Mayulermo del valle central.

Por el contrario, la estación Chiguayante presentó el error más alto en ambos métodos. Los valores faltantes en la estación Chiguayante se encontraban concentrados al final de la serie y fue la estación que presentó un menor coeficiente de correlación respecto a su estación base al igual que un menor coeficiente de determinación. Por tanto,

la distribución de los datos faltantes al interior de las series podría ser un factor determinante para anticipar el desempeño de un método de relleno basado en datos.

DISCUSIÓN

Los resultados permiten inferir que la aproximación por método de razón normal ofrece un menor error para el cálculo de la precipitación independientemente de la zona seleccionada. Los estimadores cuadráticos, como SSR, MSE y RMSE, penalizan residuos grandes, revelando para este periodo de 32 años un mejor rendimiento y menor error de predicción en el método de razón normal en las estaciones objetivo de ambas zonas morfoestructurales, siendo la estación Dichato la estación con menor error promedio y Mayulermo la estación con menor error relativo medio para ambos métodos en la muestra seleccionada. Esto es

Cuadro 7. Test de normalidad Shapiro Wilk para los modelosLm1 a Lm5 (Cuadro 6).

Modelo	W	p-value
Lm1	0.99177	0.9976
Lm2	0.96807	0.4878
Lm3	0.97316	0.7257
Lm4	0.9415	0.1098
Lm5	0.96721	0.6226

Cuadro 8. Estimadores de desempeño para dos métodos de relleno de datos meteorológicos: regresión lineal y razón normal para la Zona Costera y Valle Central. %Data: años porcentuales totales con registro de precipitación anual.

Grupo1: Zona costera	%Data P anual	Estimadores	MRL	Razón normal
Nonguén	91%	Error promedio	131.05	112.82
		SSR	785734.06	48577.10
		MSE	27094.28	16847.49
		RMSE	164.60	129.80
		MRE	0.116	0.095
		MAPE	11.6%	9.5%
		CoD	0.679	0.80
Dichato	94%	Error promedio	73.55	70.50
		SSR	257059.11	246712.74
		MSE	8568.64	8223.76
		RMSE	92.57	90.68
		MRE	0.080	0.076
		MAPE	8.0%	7.6%
		CoD	0.714	0.84
Chiguayante	75%	Error promedio	170.66	141.18
		SSR	1136352.09	775387.92
		MSE	45454.08	31015.52
		RMSE	213.20	176.11
		MRE	0.148	0.121
		MAPE	14.8%	12.1%
		CoD	0.535	0.71

Cuadro 8. Continúa.

Grupo2: Valle Central	%Data P anual	Estimadores	MRL	Razón normal
Bernardo O'Higgins Ch.	91%	Error promedio	98.62	76.10
		SSR	384072.53	253707.01
		MSE	13243.88	8748.52
		RMSE	115.08	93.53
		MRE	0.099	0.074
		MAPE	9.9%	7.4%
		CoD	0.800	0.88
Mayulermo	72%	Error promedio	105.66	79.88
		SSR	344309.57	248662.88
		MSE	14969.98	10811.43
		RMSE	122.35	103.98
		MRE	0.074	0.053
		MAPE	7.4%	5.3%
		CoD	0.893	0.97

Fuente: elaboración propia.

especialmente interesante, teniendo en cuenta que Dichato, a diferencia de Chiguayante y Nonguén, es la estación más alejada de su estación base. La explicación estadística de este hallazgo puede posiblemente encontrarse en la dispersión de los datos de precipitación en la variable independiente. La explicación física de este hallazgo podría relacionarse a la distancia relativa de las estaciones con relación al movimiento de los frentes convectivos que aproximan la humedad y la precipitación al continente desde el océano Pacífico, en concordancia con los esquemas del comportamiento climatológico de la precipitación para Sudamérica (Garreaud, 2011) así como la altitud del lugar de medición (componente orográfico) (Barrett, Garreaud y Falvey, 2009).

De forma semejante, el método que presenta una mejor bondad de ajuste, según el coeficiente de determinación, es el método de razón normal. La estación en la cual los modelos permiten explicar en mayor porcentaje la variabilidad de la precipitación es Mayulermo.

Sugerimos contrastar este método con otros estocásticos como los de kriging espacial o basados en datos como las redes neuronales, que también pueden aportar buenas aproximaciones a la estimación de datos de precipitación faltante en zonas montañosas.

CONCLUSIONES

- La observación de residuales (SSR, MSE, RMSE) indicó que el modelo de regresión lineal está influenciado por los valores atípicos (por ejemplo, en la estación Chiguayante). Sin embargo, estos valores fueron considerados pues pueden representar información relevante en otros análisis de eventos climáticos extremos como la sequía.
- El modelo de regresión está sujeto a reducir sus posibilidades de precisión en inferencias de datos pluviométricos (en la estación incógnita) para valores de precipitación (pasados o futuros) que estén fuera del rango de la variable independiente, en la estación base utilizada al momento del cálculo de la ecuación de regresión.

- La variación residual dada por el estimador SSR, así como el CoD, evalúan la eficiencia con la que ambos métodos pueden predecir los valores reales evidenciando, para este trabajo, que el método razón normal presenta un mayor coeficiente de determinación y una menor variabilidad en los errores de estimación. Esto nos lleva a concluir que si bien la regresión lineal es un método que sigue siendo ampliamente utilizado y recomendado en Chile y otros países (Barrios, Trincado, y Garreaud, 2018; Gómez, Palarea, y Martín, 2006; Luna y Lavado, 2015; Pizarro *et al.*, 2009)se debe volver la vista al método de la razón normal como un método que presenta mayor precisión para la predicción de la precipitación faltante en estaciones de Chile Centro Sur.
- Los estimadores cuadráticos MSE y RMSE permiten inferir que aquellas estaciones que presentaron menor magnitud promedio del error de estimación y, por tanto, aquellas en las cuales los métodos de predicción analizados obtuvieron mayor éxito, fueron estaciones en las cuales la precipitación mostró un comportamiento más estable alrededor de la media.
- Los estimadores adimensionales MRE y MAPE, permitieron ratificar el método de razón normal y determinar que un mejor comportamiento medio de la predicción no estaba relacionado directamente con la distancia de estas estaciones con la estación base usada para la predicción, sino con la dispersión de los datos.
- Otras explicaciones de índole física y climatológica de este comportamiento deberán ser exploradas conjuntamente, como la conectividad pluviométrica hallada entre estaciones no cercanas en relación con el comportamiento frontal en la precipitación (similar a un efecto Doppler), que permitan confirmar que la orientación relativa de las estaciones respecto al movimiento de los frentes húmedos desde el océano Pacífico, así como su gradiente altitudinal, determinan una mayor conectividad y correlación pluviométrica que la misma separación euclidiana entre estaciones.
- Los dos métodos evaluados representan una manera sencilla de obtener datos meteorológicos cuando no hay información suficiente; sin embargo, el método de la razón normal demostró un mejor comportamiento que LRM para estimación de datos de precipitación faltantes en zonas costeras y del valle central del Centro Sur de Chile.
- El método de razón normal ha demostrado mejores estimaciones en dos zonas morfoestructurales distintas. La implementación en otras áreas requerirá de la prueba y comparación de estos y otros métodos que permitan extrapolar las conclusiones obtenidas en este estudio.

AGRADECIMIENTOS

Los autores agradecen al Programa de Formación de Capital Humano Avanzado PFCHA-CONICYT Doctorado Nacional, Gobierno de Chile, 2016, Folio: 21161069, por la financiación otorgada. Igualmente, al Gobierno de Colombia, Departamento Administrativo de Ciencia, Tecnología e Innovación, COLCIENCIAS.

REFERENCIAS

Aparicio, F. J. (2004). *Fundamentos de Hidrología de Superficie*. México: Limusa.

Barrett, B. S., Garreaud, R. y Falvey, M. (2009). Effect of the Andes Cordillera on Precipitation from a Midlatitude Cold Front. *Monthly Weather Review*, *137*(9), 3092-3109. http://doi.org/10.1175/2009MWR2881.1

Barrios, A., Trincado, G. y Garreaud, R. (2018). Alternative approaches for estimating missing climate data : application to monthly precipitation records in South- Central Chile. *Forest Ecosystems*, *5*(28). https://doi.org/10.1186/s40663-018-0147-x

Benitez G., A. (1998). Taller de Hidrología aplicada a la resolución de solicitudes de derechos de aprovechamiento de aguas superficiales. Santiago, Chile: Dirección General de Aguas. Ministerio de Obras Públicas.

Carrera-Villacrés, D. V., Guevara-García, P. V., Tamayo-Bacacela, L. C., Balarezo-Aguilar, A. L., Narváez-

Rivera, C. A. y Morocho-López, D. R. (2016). Relleno de series anuales de datos meteorológicos mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la precipitación media. *Idesia (Arica)*, *34*(3), 81-90. http://doi.org/10.4067/S0718-34292016000300010

CR2 (Centro de Ciencia del Clima y la Resiliencia). (2016). Datos observados de precipitación acumulada mensual [base de datos]. Recuperado el 11 de enero de 2016 de http://www.cr2.cl/recursos-y-publicaciones/bases-de-datos/

DMC (Dirección Meteorológica de Chile). (2001). *Climatología Regional*. Recuperado de http://164.77.222.61/climatologia/publicaciones/Climatologia_regional.pdf

Fekete, B. M., Vörösmarty, C. J., Roads, J. O. y Willmott, C. J. (2004). Uncertainties in precipitation and their impacts on runoff estimates. *Journal of Climate*, *17*(2), 294-304. http://doi.org/dx.doi.org/10.1175/1520-0442(2004)

Garreaud, R. (2011). Factores del clima en Chile: Una mirada continental. Universidad de Chile. Recuperado de https://goo.gl/8jXYhC

Gómez G., J., Palarea A., J. y Martín F., J. A. (2006). Métodos de inferencia estadística con datos faltantes: estudio de simulación sobre los efectos en las estimaciones. *Estadística Española*, *48*(162), 241-270.

Kazmier, L. J. (1998). *Estadística aplicada a la administración y a la economía* (3a. ed). México: McGraw-Hill.

Linsley, R. K., Kohler, M. A. y Paulhus, J. (1958). *Hydrology for Engineers*. Nueva York: McGraw-Hill.

Lu, G. Y. y Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers and Geosciences*, *34*(9), 1044-1055. http://doi.org/10.1016/j.cageo.2007.07.010

Luna, E. y Lavado, W. (2015). Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la cuenca Jetepeque, Perú. *Revista Tecnológica ESPOL – RTE*, *28*(3), 42–52. http://doi.org/10.1089/ees.2013.0409

Martínez de Azagra, A. y Navarro H., J. (2007). *Hidrología forestal, el ciclo hidrológico*. Valladolid, España: Secretariado de Publicaciones, Universidad de Valladolid.

McCuen, R. (1998). *Hydrologic Analysis and Design* (2a. ed.). Englewood Cliffs, NJ.: Prentice Hall.

Mendenhall, W. (1990). *Estadística para administradores*. (2a. ed.) (pp. 442-473). Belmont, California: Grupo Editorial Iberoamérica.

Monsalve, G. (2009). *Hidrología en la Ingenieria* (2a. ed.). Colombia: Escuela Colombiana de Ingeniería.

Pizarro, R., Ausensi, P., Aravena, D., Sangüesa, C., León, L. y Balocchi, F. (2009). Evaluación de Métodos Hidrológicos para la Completación de Datos Faltantes de Precipitación en estaciones de la Región del Maule, Chile. *Aqua-LAC*, *1*(2), 172-185.

Pizarro T., R., Ramirez B., C. y Flores V., J. P. (2003). Análisis comparativo de cinco métodos para la estimación de precipitaciones areales anuales en períodos extremos. *Bosque*, *24*(3), 31-38. https://doi.org/10.4067/S0717-92002003000300003

Rojo, J. M. A. (2007). Regresión Lineal Simple. Madrid, España: Instituto de Economía y Geografía .Consejo Superior de Investigaciones Científicas. Recuperado de http://humanidades.cchs.csic.es/cchs/web_UAE/tutoriales/PDF/Regresion lineal simple_3.pdf

Ruiz-Maya, L. y Martín Pliego, J. (1999). *Fundamentos de Inferencia Estadística*. Madrid: Editorial AC.

Sarricolea E., P., Herrera O., M. y Araya E., C. (2013). Análisis de la concentración diaria de las precipitaciones en Chile central y su relación con la componente zonal (subtropicalidad) y meridiana (orográfica). *Investigaciones Geográficas*, *45*, 37-50. https://doi.org/10.5354/0719-5370.2013.27595

Teegavarapu, R. S. V. (2009). Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. *Journal of Hydroinformatics*, *11*(2), 133–146. http://doi.org/10.2166/hydro.2009.009

Teegavarapu, R. S. V y Chandramouli, V. (2005). Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, *312*(1-4), 191-206. http://doi.org/10.1016/j.jhydrol.2005.02.015

Toro T., A. M., Arteaga R., R., Vázquez P., M. A. y Ibáñez C., L. A. (2015). Relleno de series diarias de precipitación, temperatura mínima, máxima de la región norte del Urabá Antioqueño. *Revista Mexicana de Ciencias Agrícolas*, *6*(2), 577-588. https://doi.org/10.29312/remexca.v6i3.640

Tung, Y. (1983). Point Rainfall Estimation for a Mountainous Region. *Journal of Hydraulic Engineering*, *109*(10), 1386-1393. http://doi.org/10.1061/(ASCE)0733-9429(1983)109:10(1386).

UNESCO-ROSTLAC. (1982). Guía metodológica para la elaboración del balance hídrico de América del Sur. Recuperado de https://hydrologie.org/BIB/Publ_UNESCO/SR_999_S_1982.pdf

Wooldridge, J. (2013). *Introductory Econometrics: A modern approach* (5a. ed.). South-Western, Cengage Learning.