SOCIEDAD CHILENA DE INGENIERÍA HIDRÁULICA

XXIV CONGRESO CHILENO DE INGENIERÍA HIDRÁULICA

EVALUACIÓN DE LA INFLUENCIA DE CLASES DE ATRIBUTOS SOBRE TRES FIRMAS HIDROLÓGICAS MEDIANTE REGRESIÓN LINEAL MÚLTIPLE

ULISES SEPÚLVEDA J.[1]
JOHN MUNRO G.[2]
MIGUEL LAGOS Z.[3]
XIMENA VARGAS M.[4]
NICOLÁS VÁSQUEZ P.[5]

RESUMEN

La estimación de información hidrológica en cuencas no controladas es un desafío común en hidrología, que usualmente se aborda desde el punto de vista de transposición por cercanía espacial. El presente documento estudia la influencia de atributos físicos y climáticos, sobre tres firmas hidrológicas: Coeficiente de Escorrentía, Flujos Altos (cuantil 95) y Descarga Media Diaria, aplicados en 83 cuencas que presentan régimen natural o poco intervenido, ubicadas entre la Región de Coquimbo y la Región de Magallanes y la Antártica Chilena.

Se ajusta una regresión lineal múltiple parsimoniosa a cada firma y examina la naturaleza de los predictores que la constituyen, para interpretar la influencia de cada clase de atributo. Para lograr el ajuste, se aplica una serie de filtros que eliminan progresivamente atributos predictores, dentro de los que destacan: i) Eliminación Gradual, ii) Eliminación por correlación y iii) Error calculado como validación cruzada (*leave one out*).

Los predictores más influyentes sobre las tres variables estudiadas pertenecen a los de clase de tipo de cubierta de suelo, seguidos por la clase de índices climáticos y la clase de ubicación y topografía. El mejor ajuste corresponde a la Descarga Media Diaria, entregando una fórmula robusta para la estimación de este atributo en cuencas sin registro de caudales.

[1] Estudiante, Depto. Ing Civil, Universidad de Chile – ulises.sepulveda@ug.uchile.cl
[2] Estudiante, Depto. Ing Civil, Universidad de la Frontera - johnbmunrog@gmail.com
[3] Profesional Docente, Depto. Ing. Civil, Investigador Asociado AMTC, U. de Chile – mlagosz@uchile.cl
[4] Profesora Asociada, Depto. Ing. Civil, U. de Chile – xvargas@uchile.cl
[5] Profesional, Depto Ing. Civil, U. de Chile – nvasquez.plac@gmail.com

1 INTRODUCCIÓN

Las estimaciones de caudales en cuencas no controladas, es una problemática que en Chile se aborda usualmente desde la transposición de caudales desde cuencas con información fluviométrica hasta la zona de estudio. Desde la publicación de Sivapalan (2003), que generó la iniciativa PUB: Runoff Prediction in Ungauged Basins, cientos de científicos en ciencias hidrológicas trabajaron durante una década para incrementar el conocimiento en este tópico desde un enfoque Darwiniano, que busca comprender la comprensión a nivel global de la hidrología, a diferencia del clásico enfoque Newtoniano basado en el estudio de cuencas individuales. El resultado de esta década de investigaciones se plasmó en la publicación de Blöschl et al. (2013), generando una contribución sin precedentes en este tópico.

Las relaciones hidrológicas que ocurren en una cuenca se pueden estudiar a través de modelos que simplifican y representen los distintos procesos que suceden al interior de ésta. Así, los modelos matemáticos permiten representar un sistema hidrológico por medio de relaciones lógicas y cuantitativas, capaces de ser modificadas para observar cómo el sistema reacciona, siendo los modelos de simulación aquellos capaces de reproducir sistemas altamente complejos (Pizarro et al., 2005). La necesidad de conocimiento específico del funcionamiento e interacción de los fenómenos hidrológicos en una cuenca (Vargas et al., 2012), implica generar modelos de simulación que permitan evaluar la influencia de variables asociadas a procesos naturales.

El hidrograma continuo de una cuenca permite construir diversas componentes típicas de ésta, definidas como firmas hidrológicas[6]. Por ejemplo: caudal medio anual, caudal medio mensual, curvas de duración, entre otros. Un enfoque típico para estimar distintas firmas hidrológicas se basa en métodos regresivos (Blöschl et al., 2013), en donde se busca establecer relaciones entre variables observables y alguna característica hidrológica de interés.

El presente artículo tiene como objetivo desarrollar modelos estadísticos para la estimación de tres firmas hidrológicas: Coeficiente de escorrentía, caudales altos (cuantil 95%) y caudal medio diario; identificando a su vez cuáles son los atributos que explican en mayor medida cada uno de estos. Para ello se consideraron como predictores atributos cuantificables en cuencas sin información fluviométrica como: ubicación y topografía, geología, características de suelos, usos de suelos e índices climáticos sobre las firmas hidrológicas.

2 BASE DE DATOS

Para desarrollar este estudio, se utiliza la base de datos $CAMELS{-}CL$[7] (Álvarez-Garretón et al, 2018), en donde existen 105 atributos característicos de 516 cuencas a lo largo de Chile, delimitadas según su estación fluviométrica de salida. Se trabaja sólo con 83 cuencas que presentan un grado de intervención menor a uno, es decir, están en régimen natural o con baja intervención antrópica. En $CAMELS - CL$, la nomenclatura de las firmas hidrológicas analizadas es: Coeficiente de Escorrentía ("runoff_ratio_cr2met"); Flujos Altos (cuantil 95) ("Q95"); y Descarga Media Diaria ("q_mean").

[6] El concepto viene del inglés *Hydrological Signatures*.
[7] http://camels.cr2.cl/

3 ZONA DE ESTUDIO

El dominio del estudio está situado dentro del territorio chileno continental, desde la Región de Coquimbo hasta la Región de Magallanes y la Antártica Chilena. Está limitado al norte por la estación meteorológica Río Lauca En Estancia El Lago, ubicada en el Norte Chico y por el sur por la estación Río Robalo En Puerto Williams, en la zona austral del país. La disposición de las cuencas analizadas se muestra en la Figura 1.

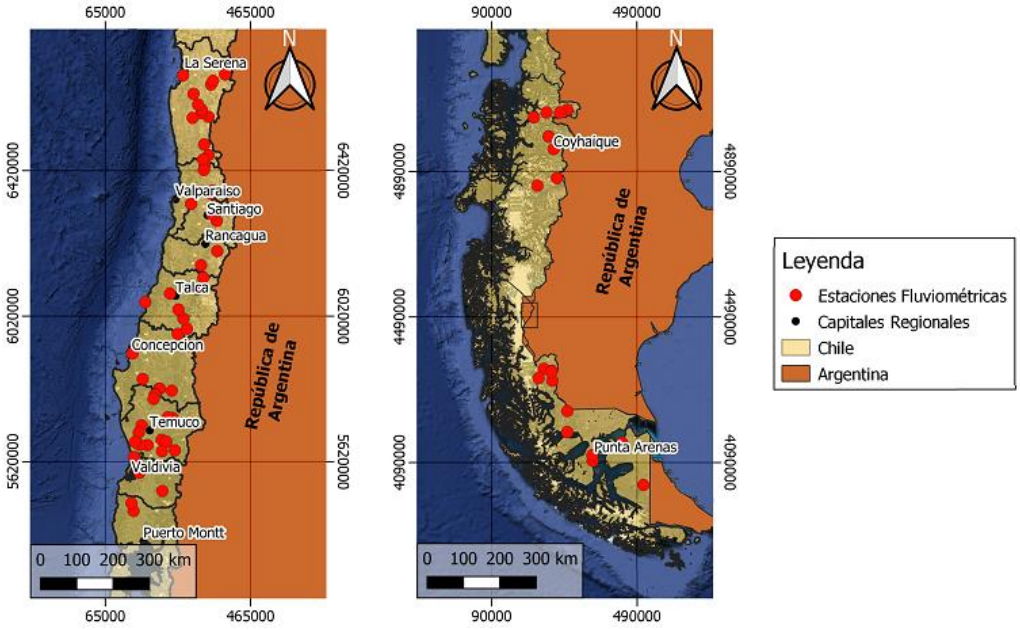

Figura 1: Distribución espacial de las 83 estaciones fluviométricas que delimitan cuencas en régimen natural, pertenecientes a $CAMELS-CL$. Mapa en coordenadas WGS84 / UTM 19S.

4 MARCO TEÓRICO

Las bases teóricas que permiten entender la metodología desarrollada en este estudio son introducidas en los siguientes puntos.

Regresión Lineal Múltiple

Un modelo generado a través de una regresión lineal múltiple (RLM) se desarrolla resolviendo el problema lineal indicado en la Ecuación (1), el que se abrevia matricialmente como muestra la Ecuación (2):

$$y = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_k X_{ik} + \epsilon_i \; ; \qquad i = 1{:}\, n \tag{1}$$

$$y = X\hat{\beta} + \epsilon \tag{2}$$

Donde y corresponde a la variable a estimar, $\hat{\beta}$ a los parámetros asociados a cada predictor definidos por el vector X y el coeficiente ϵ es el residuo obtenido a través de la diferencia de estimación y observación. Además, cada variable está definida como:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \cdots & x_{kn} \end{pmatrix}; \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}; \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Con n el número de predictores y k el número de cuencas. Luego, el vector $\hat{\beta}$ es estimado según la Ecuación (3), mediante mínimos cuadrados.

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

Finalmente se obtiene la estimación de los valores a partir de la Ecuación (4).

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_j x_{ik} = X\hat{\beta} \tag{4}$$

Criterio de Información de Akaike

El criterio de información de Akaike (AIC) es una medida de parsimonia de un modelo, que relaciona la calidad relativa de un modelo estadístico dado un número de predictores. El índice AIC maneja un *"trade-off"* entre la bondad de ajuste del modelo y la complejidad del modelo (número de predictores). En el caso general, el criterio es:

$$AIC = 2\,k - 2\ln(L) \tag{5}$$

donde k es el número de parámetros en el modelo estadístico, y L es el máximo valor de la función de verosimilitud para el modelo estimado.

Criterio de Correlación de Pearson

El Coeficiente de Correlación de Pearson (R), es una medida de la correspondencia entre dos variables cuantitativas aleatorias. En palabras más simples, se puede definir como un índice utilizado para medir el grado de relación que tienen dos variables, ambas cuantitativas.

$$R = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{6}$$

Donde R es la covarianza entre las variables, y $\sigma_X\ y\ \sigma_Y$ es la desviación estándar de las variables X e Y, respectivamente.

Validación Cruzada Dejando Uno Fuera

La validación cruzada dejando uno fuera $(LOOCV)$[8], es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre

[8] Del inglés *Leave One Out Cross Validation*

datos de entrenamiento y de prueba. Esta metodología consiste en: i) separar un elemento de la muestra, ii) estimar este valor utilizando todo el resto del conjunto de datos y evaluar la diferencia obtenida entre lo observado y lo estimado (Figura 2). El error se calcula como indica la Ecuación (7)

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

Dónde, ECM es el error cuadrático medio de la validación cruzada, y_i es el valor observado e $\hat{y}_i$ corresponde al valor simulado.

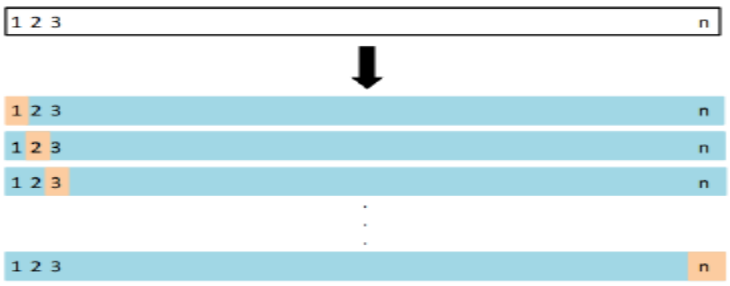

Figura 2. Esquema de $LOOCV$. Un set de n datos es dividido n veces en un set de entrenamiento (azul) y un set de validación con una sola observación (naranjo). (James, 2013)

5 METODOLOGÍA

La determinación de la influencia de las clases de atributos se consigue al ajustar una RLM parsimoniosa a cada firma. Para lograr esto se siguen cinco pasos, descritos a continuación:

Paso 1: Elección Inicial del Set de Predictores

Se aplica una serie de filtros iniciales a la base de datos de $CAMELS - CL$, orientados a mejorar la calidad de los predictores, los que se explicitan a continuación:

- Eliminación de atributos no numéricos.
- Eliminación de atributos que presenten expresión "NA".
- Eliminación de atributos hidrológicos.

Los predictores resultantes se adimensionalizan dividiéndolos por el valor promedio que toma la variable en las 83 cuencas estudiadas.

Paso 2: Eliminación Gradual

El método de Eliminación hacia atrás[9] (Wilks, 2011), implica generar una RLM con todos los predictores candidatos del Paso 1 y eliminar progresivamente el predictor que produce la

[9] Método *Backward Elimination* descrito en Wilks (2011), Capítulo 7, pp. 249.

máxima disminución del AIC. Se escoge el set de predictores tal que cualquiera que se elimine genera un aumento en el valor del AIC.

Paso 3: Eliminación según criterio de correlación de Pearson

A partir de una matriz numérica en base al coeficiente R, se eliminan predictores que estén altamente correlacionados entre ellos ($R > 0{,}7$) y que a la vez, su eliminación empeore levemente el ajuste del modelo, en términos de AIC.

Paso 4: Principio de Parsimonia

Se genera una RLM con todos los predictores resultantes del Paso 3 y se prueba la eliminación de una variable, calculando el nuevo valor del error mediante ECM; se descarta el predictor sólo si su eliminación produce una disminución del ECM o un aumento máximo del 5% con respecto al valor inicial de éste. Este proceso se repite hasta que no puedan eliminarse variables sin una pérdida de ajuste estadísticamente significativa.

Paso 5: Ajuste de Regresión Lineal Múltiple

En base a los conjuntos de predictores resultantes de los filtros, se ajusta una RLM final a cada firma hidrológica, con la que se determina cómo influyen las clases de atributos sobre ellas a partir de examinar la naturaleza de los predictores que las constituyen.

6 RESULTADOS

En base a las 83 cuencas en régimen natural y a los 34 atributos resultantes de la Elección Inicial del Set de Predictores, expuestos en la Tabla 1, se muestran los resultados obtenidos para las tres firmas hidrológicas analizadas.

Tabla 1. Predictores de la base de datos de $CAMELS - CL$, resultantes del Paso 1.

Clase de Atributo			
Ubicación y Topografía	Características Geológicas	Cubierta de Suelo	Índices Climáticos
gauge_lat	geol_class_1st_frac	crop_frac	p_mean_cr2met
gauge_lon	geol_class_2nd_frac	nf_frac	pet_mean
area	carb_rocks_frac	fp_frac	aridity_cr2met
elev_gauge		grass_frac	p_seasonality_cr2met
elev_mean		shrub_frac	frac_snow_cr2met
elev_med		wet_frac	high_prec_freq_cr2met
elev_max		imp_frac	high_prec_dur_cr2met
elev_min		lc_barren	low_prec_freq_cr2met
slope_mean		snow_frac	low_prec_dur_cr2met
		lc_glacier	p_mean_spread
		fp_nf_index	
		dom_land_cover_frac	

Eliminación Gradual

La Figura 3 muestra la relación entre el número de predictores usados para los tres modelos y el valor del AIC. Producto de la Eliminación Gradual se pasó desde 34 a 19; 20; y 18 predictores para las variables hidrológicas $runnof\_ratio$, $Q95$ y q_mean, respectivamente.

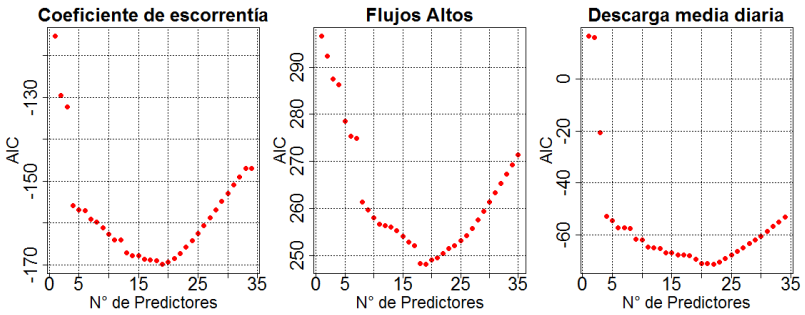

Figura 3. Variación del AIC en función del número de predictores usados.

Eliminación según criterio de correlación de Pearson

Las Figuras 4, 5 y 6 exhiben la correlación entre los predictores seleccionados desde la metodología de Eliminación Gradual y los predictores seleccionados por el criterio de correlación de Pearson. Producto de esta metodología se pasó desde (19, 20 y 18) a (10, 10 y 8) predictores para las variables hidrológicas $runnof\_ratio$, $Q95$ y q_mean, respectivamente.

Coeficiente de Escorrentía (runnof_ratio):

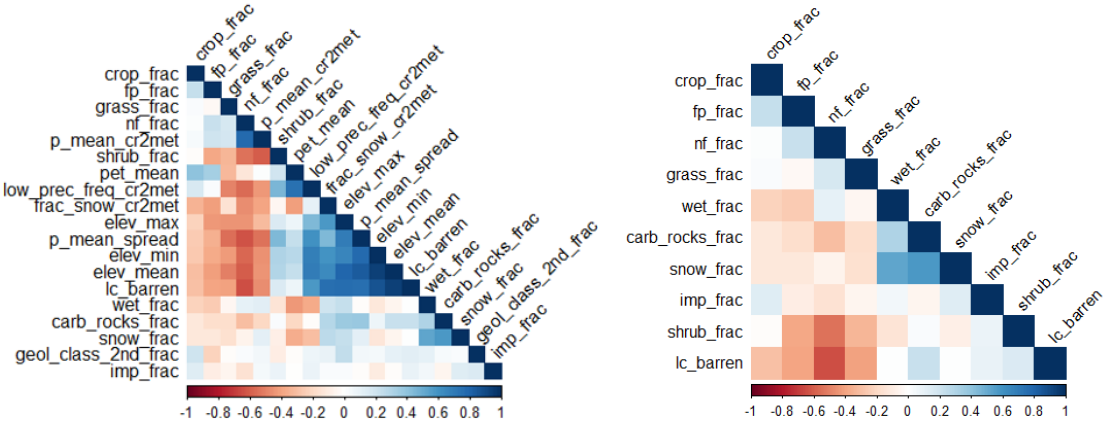

Figura 4. Izquierda: correlograma de predictores obtenido por el filtro de Eliminación Gradual. Derecha: correlograma de predictores filtrado por criterio de correlación de Pearson.

Flujos Altos (Q95):

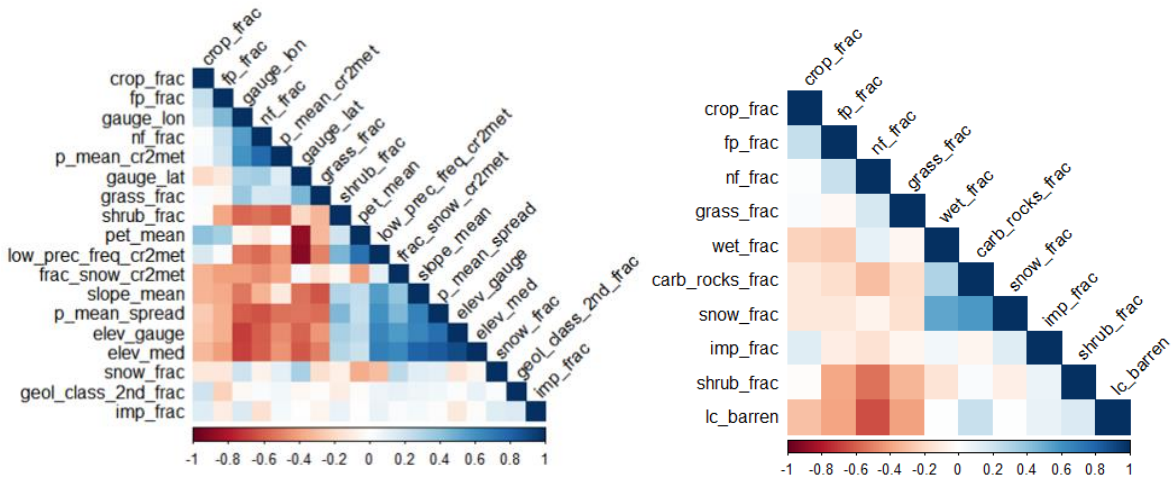

Figura 5. Izquierda: correlograma de predictores obtenido por el filtro de Eliminación Gradual. Derecha: correlograma de predictores filtrado por criterio de correlación de Pearson.

Descarga media diaria (q_mean):

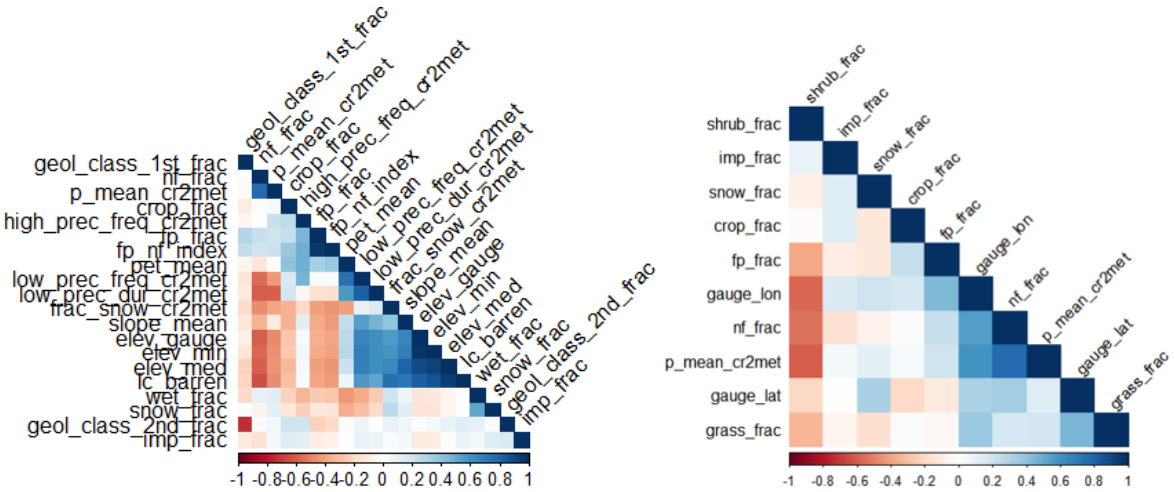

Figura 6. Izquierda: correlograma de predictores obtenido por el filtro de Eliminación Gradual. Derecha: correlograma de predictores filtrado por criterio de correlación de Pearson.

Principio de Parsimonia

La Figura 7 indica la variación del ECM al eliminar el predictor menos significativo del modelo. Producto de esta metodología, se determina que el número de predictores a usar para las variables hidrológicas $runnof\_ratio$, $Q95$ y q_mean, son 8, 8 y 5 respectivamente.

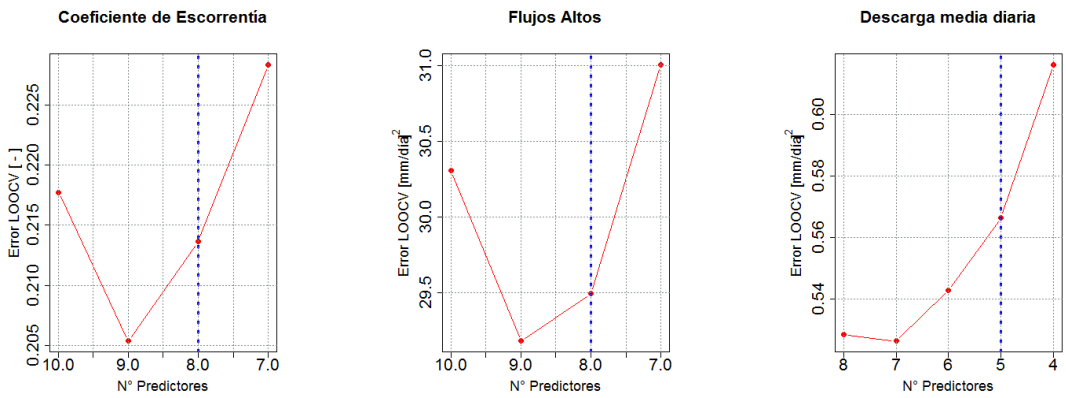

Figura 7. Variación del ECM obtenido mediante $LOOCV$, para las tres firmas hidrológicas.

Regresiones Lineales Múltiples

Tres RLM parsimoniosas se ajustan a las firmas hidrológicas $runnof\_ratio$, $Q95$ y q_mean, las que son explicitadas en las Ecuaciones (8), (9) y (10), respectivamente. La Tabla 2 muestra un resumen de la naturaleza de los predictores usados en estas ecuaciones.

$$\boldsymbol{runnof\_ratio} = 4{,}43 - 0{,}06\ carb_{rock} - 0{,}09\ crop - 0{,}74\ nf - 0{,}25\ fp - 0{,}58\ grass - 1{,}03\ shrub - 0{,}08\ wet - 0{,}75\ barren \quad (8)$$

$$\boldsymbol{Q95} = -318{,}57 + 18{,}45\ lat + 313{,}86\ lon - 4{,}33\ nf - 3{,}85\ grass - 2{,}17\ shrub - 1{,}55\ imp - 0{,}80\ snow + 7{,}48\ P_{mean}CR2MET \quad (9)$$

$$\boldsymbol{q\_mean} = -2{,}14 + 1{,}34\ slope + 1{,}08\ nf + 0{,}19\ imp + 0{,}30\ snow + 2{,}31\ P_{mean}CR2MET \quad (10)$$

Dónde:

$carb_{rock}$ $[-]$: Fracción de cuenca caracterizada como “rocas sedimentarias carbonatadas”.
$crop$ [%]: Porcentaje de cuenca cubierta por tierras de cultivo.
nf [%]: Porcentaje de la cuenca cubierta por bosque clasificado como natural.
fp [%]: Porcentaje de la cuenca cubierta por bosque clasificado como plantación.
$grass$ [%]: Porcentaje de la cuenca cubierta por praderas.
$shrub$ [%]: Porcentaje de la cuenca cubierta por matorrales.
wet [%]: Porcentaje de la cuenca cubierta por humedales y cuerpos de agua.
$barren$ [%]: Porcentaje de la cuenca cubierta por tierras áridas.
lat $[°S]$: Latitud de la cuenca.
lon $[°N]$: Longitud de la cuenca.
imp [%]: Porcentaje de la cuenca cubierta por superficies impermeables.

$snow$ [%]: Porcentaje de captación cubierta por nieve y hielo.
$P_{mean}CR2MET$ $[mm/d]$: Precipitación media diaria del producto $CR2MET$.
$slope$ $[m/km]$: Pendiente media de la cuenca.

Tabla 2. Predictores agrupados en función de la naturaleza de su clase de atributo.

Firma Hidrológica	Número de predictores	Clase de atributo			
		Ubicación y Topografía	Características Geológicas	Cubierta de Suelo	Índices Climáticos
runnof_ratio	8	0	1	7	0
Q95	8	2	0	5	1
q_mean	5	1	0	3	1

Las Figuras 8, 9 y 10 muestran el desempeño de las RLM para cada firma hidrológica analizada, además de las métricas: coeficiente de determinación (R^2); Índice de eficiencia de Nash-Sutcliffe (NSE); e Índice de eficiencia de Kling-Gupta (KGE).

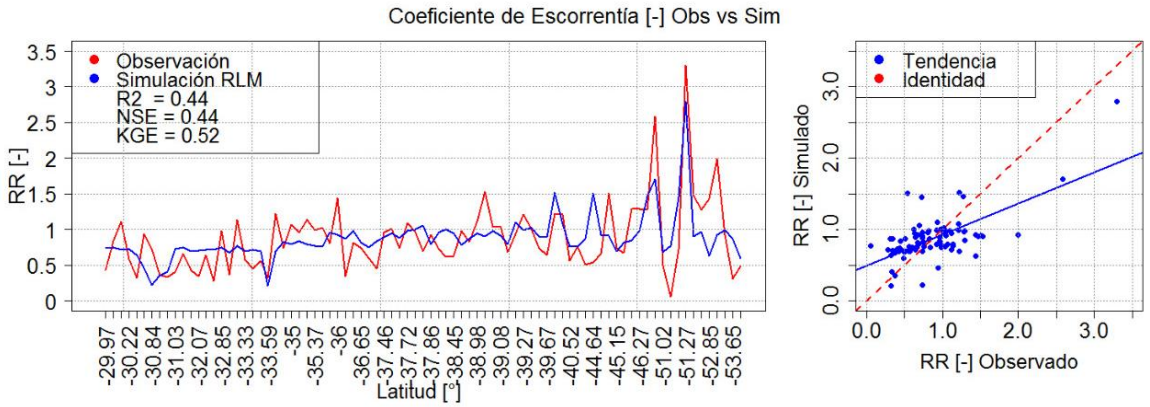

Figura 8. Diagramas de líneas, dispersión y métricas hidrológicas, para la firma $runnof\_ratio$

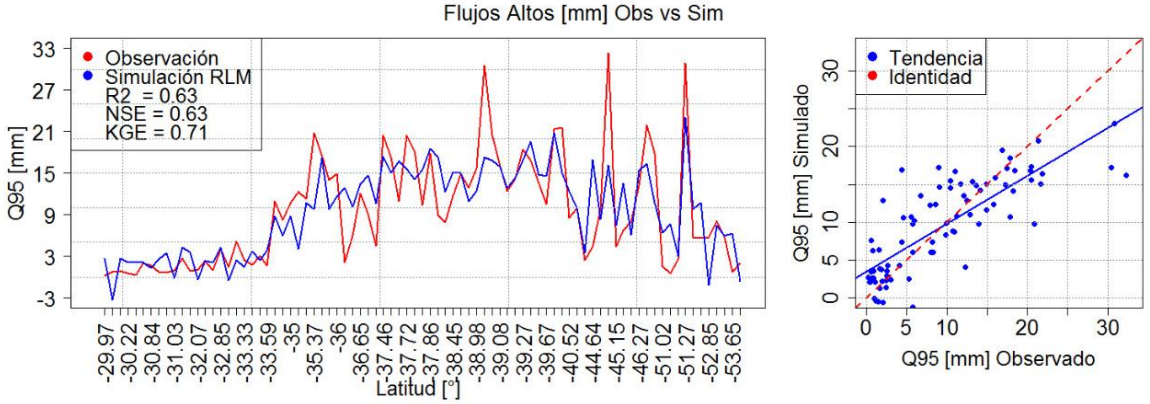

Figura 9. Diagramas de líneas, dispersión y métricas hidrológicas, para la firma $Q95$.

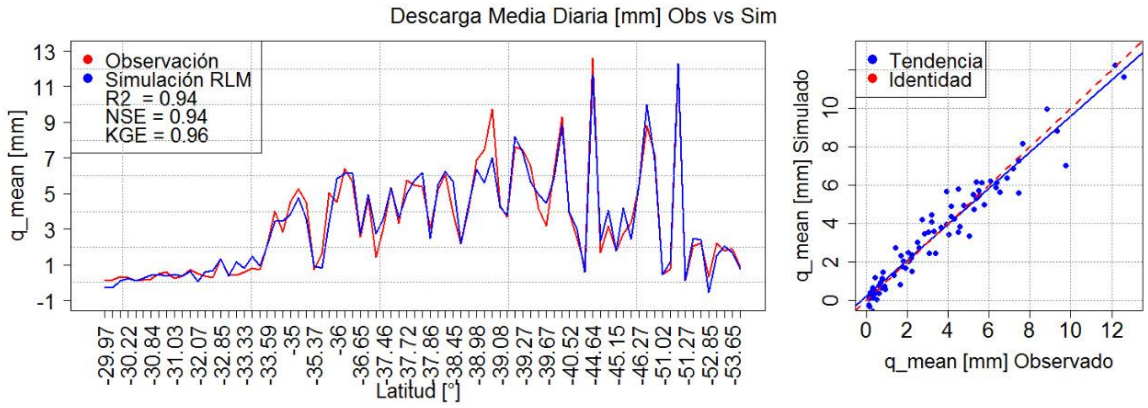

Figura 10. Diagramas de líneas, dispersión y métricas hidrológicas, para la firma q_mean.

Claramente el modelo de la Descarga Media Diaria (Figura 10) es el más relevante, ya que las métricas hidrológicas aplicadas indican que puede reproducir correctamente magnitud y variabilidad, entregando una fórmula robusta para la estimación de este atributo en cuencas sin registro de caudales.

7 DISCUSIÓN Y CONCLUSIONES

El proceso de filtración de atributos ayuda a que las RLM generadas tengan mejor desempeño en estimar la variable hidrológica de interés, ya que, si existen muchas variables predictoras, probablemente un número considerable de ellas no tendrá relación con la firma hidrológica a estimar, lo que lleva a que sólo sean "ruido" o a que eventualmente se produzca un sobre ajuste. En la medida de lo posible, es recomendable reducir la cantidad de predictores.

Mediante el filtro inicial, Eliminación Gradual, Coeficiente de correlación de Pearson y el error calculado como $ECM - VC$ se eliminaron hasta 100 predictores, dando a paso a tres RLM parsimoniosas de ocho; ocho; y cinco variables para las firmas hidrológicas $runnof\_ratio$, $Q95$ y q_mean, respectivamente. El modelo asociado a la firma q_mean es el más destacado ya que sólo depende de cinco predictores y su error calculado como ECM toma el valor de 0,565 $\left[\frac{mm}{día}\right]^2$, el que es relativamente bajo e indica estimaciones de calidad.

Apoyados sobre el total de predictores que pertenecen a una clase de atributo específica, definida en $CAMELS - CL$, los atributos más influyentes en general sobre las tres variables estudiadas son los de la clase de tipo de cubierta de suelo, seguidos por la clase de índices climáticos y la clase de ubicación y topografía. Además, los resultados muestran que la clase de atributo de tipo geológico es poco significativa en los ajustes.

De los resultados de este estudio, se cuenta con tres regionalizaciones de coeficiente de escorrentía, caudal medio anual y caudal perteneciente al 95% más alto que pueden ser utilizadas en estudios de líneas bases, así como de forma comparativa con otras metodologías para estimación de caudales en cuencas no controladas.

8 REFERENCIAS

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., . . . and Ayala, A. (2018). The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset. *Hidrology and Earth System Sciences*, 5817-5846.

Blöschl, G., Sivapalan, M., Savenije, H., Wagener, T., & Viglione, A. (Eds.). (2013). Runoff prediction in ungauged basins: synthesis across processes, places and scales. Cambridge University Press.

James, G. W. (2013). *An Introduction to Statical Learning.* Nueva York: Springer.

Kutner , M., Nachtsheim , C., Neter, J., & Li, W. (1996). *Applied Linear Statistical Models.* New York: The McGraw-Hill Companies.

McKee, T., Doesken, N. J., & Kleist, J. (1993). *The Relationship of Drought Frequency and Duration to time scales.* Anaheim: Eighth Conference on Applied Climatology.

Montgomery, D., & Runger, G. (2002). *Probabilidades y Estadística aplicadas a la Ingeniería.*

Murrel, P. (2006). *R Graphics: Computer Science and Data Analysis Series.*

Pizarro T., R., Soto B., M., Farias D., C., & Jordan D., C. (2005). *Aplicación de dos Modelos de Simulación Integral Hidrológica, para la estimación de caudales medios mensuales, en dos cuencas de Chile central.*

Stowhas, L. (2016). *Fundamentos de la Hidrología Aplicada.* Valparaíso: Editorial USM.

Sivapalan, M. (2003). Prediction in ungauged basins: a grand challenge for theoretical hydrology. Hydrological Processes, 17(15), 3163-3170.

Vargas, J., De La Fuente, L., & Arumí, J. (2012). *Balance hídrico mensual de una cuenca Patagónica de Chile: Aplicación de un modelo parsimonioso.*

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences.* San Diego: Elsevier.