Modelos Estadísticos Híbridos en para pronóstico de caudales estacionales en cuencas de Chile Central

Tabla de contenido

[1 Resumen ejecutivo 3](#_Toc134452340)

[2 Introducción 4](#_Toc134452341)

[2.1 Contexto y motivación 4](#_Toc134452342)

[2.2 Objetivos 4](#_Toc134452343)

[3 Fundamentos 5](#_Toc134452344)

[3.1 Modelo hidrológico 5](#_Toc134452345)

[3.2 Índices climáticos: ENSO, ONI y SOI 5](#_Toc134452346)

[3.3 Reanálisis ERA5 5](#_Toc134452347)

[3.4 Modelos estadísticos 5](#_Toc134452348)

[4 Metodología de modelos estadísticos híbridos 5](#_Toc134452349)

[4.1 Modelo de regresión 6](#_Toc134452350)

[4.2 Enfoque híbrido 7](#_Toc134452351)

[4.3 Predictores 7](#_Toc134452352)

[4.4 Coeficientes de Pardé y algoritmo KNN para distribución del volumen 8](#_Toc134452353)

[5 Pronóstico de volúmenes estacionales 8](#_Toc134452354)

[5.1 Pronóstico de caudales medios mensuales 8](#_Toc134452355)

[5.2 Evaluación de los modelos híbridos en hidrología 8](#_Toc134452356)

[6 Evaluación de los pronósticos 8](#_Toc134452357)

[6.1 Métricas determinísticas 8](#_Toc134452358)

[6.2 Métricas de conjunto 8](#_Toc134452359)

[7 Conclusiones 9](#_Toc134452360)

[7.1 Desarrollo futuro 9](#_Toc134452361)

[8 Referencias 10](#_Toc134452362)

# Resumen ejecutivo

En este reporte, se presenta un análisis de los modelos estadísticos híbridos para 45 cuencas de Chile Central, que combinan un modelo hidrológico dinámico y relaciones estadísticas con predictores climáticos.

# Introducción

## Contexto y motivación

El uso de los pronósticos de caudales suele ser más intensivo después de septiembre, sin embargo, en la agricultura hay decisiones del tipo de cultivo o la extensión de las plantaciones que consideran la cantidad de agua disponible durante el periodo de invierno.

## Objetivos

# Fundamentos

## Modelo hidrológico

El año hidrológico comienza en otoño y termina en verano del año siguiente. Esta traslación del año calendario tiene la intención de considerar todas las fuentes de acumulación y derretimiento dentro de un periodo anual. Para el caso específico de Chile, la convención es que el año hidrológico comienza el 1 de abril y termina el 31 de marzo del año siguiente.

## Índices climáticos

## Reanálisis ERA5

## Modelos estadísticos

## Métricas de evaluación del pronóstico

### Métricas Determinísticas

a. Coeficiente de determinación (R²): Esta métrica proporciona una medida de cuánta de la variabilidad total de los datos puede ser explicada por el modelo. Un valor de R² cercano a 1 indica que el modelo puede explicar una gran proporción de la variabilidad, mientras que un valor cercano a 0 indica lo contrario.

b. Error cuadrático medio (RMSE): RMSE es una métrica comúnmente utilizada para medir la diferencia entre los valores predichos por el modelo y los valores observados. Cuanto menor sea el RMSE, mejor será el modelo. Esta métrica es especialmente útil cuando se desean penalizar errores grandes.

c. Error absoluto medio (MAE): El MAE es la media de las diferencias absolutas entre las predicciones y las observaciones reales. Esta métrica es menos sensible a los errores grandes que RMSE.

### Métricas Probabilísticas

a. Puntuación continua de Brier (CRPS): CRPS es una métrica que cuantifica la diferencia entre la distribución de probabilidad prevista y la distribución de probabilidad observada. En el caso de los modelos híbridos en hidrología, CRPS puede utilizarse para evaluar la precisión del pronóstico del volumen estacional de los caudales.

b. Puntuación de habilidad continua de Brier (CRPSS): CRPSS es una métrica que compara el rendimiento de un modelo de pronóstico con un modelo de referencia (generalmente un modelo de pronóstico climatológico u otra versión del modelo). Un CRPSS positivo indica que el modelo de pronóstico tiene un rendimiento superior al modelo de referencia.

### Métricas de Clasificación

a. Exactitud (Accuracy): Esta métrica proporciona una medida general de cuántas predicciones fueron correctas.

b. Precisión: La precisión es la proporción de predicciones verdaderas positivas sobre el total de verdaderas positivas y falsas positivas. Es una indicación de cuántos de los años clasificados como húmedos, normales o secos fueron clasificados correctamente.

c. Sensibilidad (Recall): También conocida como sensibilidad o tasa de verdaderos positivos, mide la proporción de verdaderos positivos que fueron identificados correctamente. Es una indicación de cuántos de los años verdaderamente húmedos, normales o secos fueron identificados correctamente.

Estas métricas proporcionan una perspectiva sobre el rendimiento del modelo, y la elección de las métricas dependerá de los objetivos específicos. Es importante recordar que no existe una única "mejor" métrica para todos los contextos, por lo que generalmente se consideran múltiples métricas al evaluar el rendimiento.

Tabla 1. Métricas de evaluación del pronóstico

|  |  |  |
| --- | --- | --- |
| Métrica | Símbolo | Ecuación |
| Coeficiente de determinación | R² |  |
| Error cuadrático medio | RMSE |  |
| Error absoluto medio | MAE |  |
| Puntuación continua de Brier | CRPS |  |
| Puntuación de habilidad continua de Brier | CRPSS |  |
| Exactitud | Accuracy |  |
| Precisión | Precision |  |
| Sensibilidad | Recall |  |

Donde

: valor observado

: valor pronosticado

: media de los valores observados

: número de observaciones

: función de distribución acumulativa de las predicciones

: función de distribución acumulativa de las observaciones

: verdaderos positivos

: verdaderos negativos

: falsos positivos

: falsos negativos

# Metodología de modelos estadísticos híbridos

## Introducción

De manera simplificada el modelo estadístico tiene una secuencia de pasos que se realizan de forma iterativa hasta conseguir un modelo final de acuerdo con criterios de calidad. Los pasos para conseguir el modelo son:

1. Preparación de datos: Seleccionar y preparar los predictores relevantes que se utilizarán en el modelo estadístico. Los predictores incluyen índices climáticos como ENSO, PDO, ONI, y SOI, así como variables diarias provenientes del modelo hidrológico TUW alimentado por reanálisis meteorológico ERA5, como la precipitación mensual acumulada para el año hidrológico en estudio. Estos predictores se normalizan utilizando métodos como centrar/escalar o minmax, utilizando los datos de entrenamiento de los años hidrológicos 1980 a 2020.

2. Verificación de supuestos: Antes de aplicar la regresión lineal múltiple, se deben verificar los supuestos estadísticos, como la normalidad de la distribución del predicando (volumen). Se utiliza el test de Shapiro-Wilks para verificar la normalidad y se aplica una transformación de logaritmo para corregir sesgos en la distribución en caso de que no se pase el test.

3. Entrenamiento del modelo: se entrena el modelo de regresión lineal múltiple utilizando los predictores mensualizados y normalizados, además de la variable dependiente: volumen o logaritmo del volumen. Se estiman los coeficientes de regresión que minimizan la función de pérdida de mínimos cuadrados ordinarios (MCO). Se generan 1000 miembros (ensembles) a partir de una perturbación con distribución normal del error cuadrático medio de la regresión lineal.

4. Validación y evaluación del modelo: el modelo se verifica usando validación cruzada para un año fuera (LOOCV, leave one out cross validation), calculando métricas de evaluación univariable como el coeficiente de determinación (R^2) y la raíz del error cuadrático medio (MSE). También se evalua el rendimiento los ensembles ocupando el CRPSS (Continuous Ranked Probability Skill Score).

5. Pronóstico de volúmenes y caudales: se utiliza el modelo entrenado y validado para pronosticar los volúmenes y caudales estacionales con nuevos predictores. Se obtiene el caudal medio mensual distribuyendo el volumen utilizando los coeficientes de Pardé de los años más cercanos basados en la distancia euclidiana de los predictores.

6. A partir de los resultados de las métricas se exploran nuevas regresiones, predictores, horizontes del pronóstico, tiempos de emisión, forzantes meteorológicas, modelos hidrológicos, métodos de corrección de sesgo, etc.

A continuación, se detallan los elementos más relevantes del proceso de modelación estadística.

## Datos de entrada

## Modelo de regresión

El modelo estadístico consiste en una regresión lineal múltiple de los diferentes predictores hídricos. La regresión lineal múltiple es un modelo estadístico que establece una relación lineal entre una variable dependiente (predictando) y varias variables independientes (o predictores). La función de optimización en la regresión lineal múltiple se utiliza para encontrar los coeficientes de regresión que minimizan la suma de los errores cuadráticos, generalmense se ocupa la función de pérdida de mínimos cuadrados ordinarios (MCO), aunque existen otras formas de optimizar la regresión lineal especialmente para un gran número de atributos (predictores o muestra).

La ecuación general de la regresión lineal múltiple es:

Donde, es la variable dependiente, son las variables independientes, es el término constante o intercepto son los coeficientes de regresión y es el término de error aleatorio.

La función de optimización en la regresión lineal múltiple busca estimar los coeficientes que minimizan la función de pérdida, que en este caso es la suma de los errores cuadráticos:

Dónde es la función de pérdida, es el valor observado de la variable dependiente, son los valores observados de las variables independientes y representa el i-ésimo predictor en el conjunto de datos.

### Supuestos regresión lineal

Los modelos de regresión lineal múltiple se basan en varios supuestos. Aquí están los supuestos más comunes y cómo se pueden comprobar:

1. Relación lineal: Los modelos de regresión lineal asumen que la relación entre las variables independientes y la variable dependiente es lineal. Esto se puede comprobar visualizando los datos utilizando gráficos de dispersión.
2. Independencia de las observaciones: Las observaciones deben ser independientes entre sí, lo que significa que las observaciones no deben influirse mutuamente. En el caso de los datos temporales o espaciales, este supuesto a menudo se viola. El estadístico Durbin-Watson se usa comúnmente para detectar la autocorrelación en los residuos de la regresión.
3. Homocedasticidad: Los residuos (errores) deben tener una varianza constante en todos los niveles de las variables independientes. Si la varianza de los residuos varía a lo largo de los niveles de las variables independientes, esto se llama heterocedasticidad. Puedes comprobar la homocedasticidad visualizando los residuos frente a los valores ajustados: los residuos deben estar distribuidos aleatoriamente y no formar un patrón discernible. También puedes usar pruebas como la prueba de Breusch-Pagan o la prueba de White para comprobar la heterocedasticidad.
4. Normalidad de los errores: Los errores (residuos) deben estar normalmente distribuidos. Esto se puede comprobar utilizando un histograma de los residuos, un gráfico Q-Q o realizando una prueba de normalidad como la prueba de Shapiro-Wilk o la prueba de Kolmogorov-Smirnov.
5. No colinealidad multicolinealidad: Las variables independientes no deben ser altamente correlacionadas entre sí, ya que esto puede hacer que sea difícil separar los efectos individuales de las variables independientes sobre la variable dependiente. Esto se puede comprobar calculando el factor de inflación de la varianza (VIF) para cada variable independiente. Un VIF mayor de 5 o 10 indica generalmente alta multicolinealidad.
6. Ausencia de puntos de apalancamiento y valores atípicos influyentes: Los modelos de regresión lineal pueden ser sensibles a los puntos de apalancamiento (observaciones con valores extremos de las variables independientes) y a los valores atípicos influyentes (observaciones con valores extremos de la variable dependiente). Estos pueden ser identificados a través de gráficos de residuos y medidas como la distancia de Cook.

Para verificar la normalidad del volumen (variable independiente) se aplicó la prueba de Shapiro-Wilks (SW). Esta prueba tiene como hipótesis nula que la variable viene de una distribución normal. Por lo tanto, si el p-value es menor al nivel alfa (generalmente 0.05) la hipótesis nula se rechaza y por lo tanto la distribución no sería normal. Cabe mencionar que este tipo de metodología no identifica la distribución, sino que aprueba o rechaza la hipótesis de normalidad. En caso de ser rechazada la hipótesis, si la variable en el histograma se encuentra sesgada a la derecha o la izquierda, la muestra se puede transformar aplicando alguna transformación de orden de magnitud, por ejemplo, logaritmo natural o raíz cuadrada. Para los volúmenes se observaron principalmente volúmenes sesgados (skewed) a la derecha, es decir, una mayor frecuencia de caudales bajos comparado con medios y altos. Por lo que se aplicó, si la hipótesis nula se rechazaba, una transformación logarítmica y nuevamente se realizó la prueba de Shapiro-Wilks a la nueva variable. La variable (volumen o logaritmo del volumen) que pase la prueba de Shapiro-Wilks se utiliza como predicando en la regresión lineal múltiple.

## Enfoque híbrido

Los modelos estadísticos híbridos en hidrología son una combinación de enfoques basados en procesos físicos de la cuenca y en relaciones estadísticas, utilizados para pronosticar el volumen estacional de los caudales. Estos modelos integran la información de varias fuentes como índices climáticos, por ejemplo, ENSO (Índices de El Niño) y SOI (Índice de Oscilación del Sur), junto con las variables de estado de modelos hidrológicos.

El enfoque híbrido se basa en combinar las ventajas de ambos tipos de modelos: las representaciones físicas detalladas de los modelos hidrológicos y la capacidad predictiva de las relaciones estadísticas.

Los índices climáticos mencionados están vinculados a las variaciones en la temperatura superficial del océano y la presión atmosférica, que influyen en la precipitación y la evaporación y en consecuencia, en los volúmenes de agua almacenados en la atmosfera.

ENSO es un fenómeno climático que implica fluctuaciones en la temperatura superficial del océano en el Pacífico ecuatorial, afectando los patrones de precipitación y temperatura en todo el mundo. ONI es una medida estandarizada de la anomalía de la temperatura superficial del océano en una región específica del Pacífico ecuatorial, mientras que SOI mide la diferencia normalizada entre la presión atmosférica en Tahití y Darwin, Australia.

Por otro lado, el modelo hidrológico TUW es un modelo conceptual que simula los procesos del ciclo hidrológico, como la humedad del suelo, la infiltración, el flujo subsuperficial y el caudal. Este modelo se alimenta con datos meteorológicos diarios del reanálisis ERA5, que proporciona información detallada y precisa sobre las condiciones meteorológicas pasadas a nivel global.

Estas relaciones pueden ser lineales, no lineales o basadas en técnicas de aprendizaje automático, como las redes neuronales artificiales, árboles de decisión o máquinas de vectores de soporte.

## Predictores

Los predictores utilizados se normalizan utilizando el método centrar/escalar que consiste en restar a cada individuo de un predictor el promedio de la muestra (centrar) y luego dividir por la desviación estándar (escalar). Tanto el promedio como la desviación se calculan ocupando los años de la muestra de entrenamiento que corresponden a los años hidrológicos 1980 a 2019. Existen otras alternativas de normalización como minmax (restar el mínimo y dividir por el máximo), sin embargo, el tipo de normalización no suele ser influyente en el resultado final de la regresión.

## Coeficientes de Pardé y algoritmo KNN para distribución del volumen

El volumen simulado se distribuye a caudales medios mensuales ocupando los coeficientes de Pardé de los años más cercanos (kNN) como se muestra en la ecuación 1.

|  |  |
| --- | --- |
|  | Ec. 1 |
|  | Ec. 2 |

Donde representa el caudal pronosticado para el mes i. El mes i debe estar dentro del horizonte del pronóstico (septiembre a marzo). El volumen pronosticado es *.*

El coeficiente de Pardé para el mes i y el año hidrológico j, se define ocupando que corresponde al caudal observado para el mes i y el año hidrológico j y que es el volumen observado para el año hidrológico j. es la distancia entre los predictores del año pronosticado y el periodo histórico (1980-2020). k corresponde al número de años considerados cercanos.

El método de k-Nearest Neighbors (kNN) es un algoritmo de aprendizaje supervisado no paramétrico utilizado para clasificación y regresión. La idea principal detrás de kNN es que las observaciones similares tienden a tener valores de salida similares. Este algoritmo se basa en medir la similitud o la distancia entre los puntos de datos en un espacio multidimensional, generalmente utilizando la distancia euclidiana u otras métricas de distancia.

El algoritmo kNN funciona de la siguiente manera:

1. Determinar el valor de 'k': Seleccionar un número entero 'k' que representa el número de vecinos más cercanos que se considerarán en el proceso de clasificación o regresión.
2. Calcular la distancia: Calcular la distancia entre la observación de interés (nueva instancia) y todas las demás observaciones en el conjunto de datos. La medida de distancia más común es la distancia euclidiana, aunque también se pueden utilizar otras métricas, como la distancia de Manhattan o la distancia de Minkowski.
3. Encontrar los vecinos más cercanos: Ordenar las observaciones según las distancias calculadas y seleccionar las 'k' observaciones con las distancias más cortas. Estos 'k' puntos de datos son los vecinos más cercanos de la instancia de interés.
4. Agregar las respuestas: se asigna a la nueva instancia el valor de la variable objetivo de sus 'k' vecinos más cercanos.
5. Predecir la etiqueta o valor de la nueva instancia basándose en la agregación de las respuestas de los vecinos más cercanos.

# Pronóstico de volúmenes estacionales

## Pronóstico de caudales medios mensuales

## Evaluación de los modelos híbridos en hidrología

# Evaluación de los pronósticos

## Métricas determinísticas

## Métricas de conjunto

# Conclusiones

## Desarrollo futuro

# Referencias

Parajka, J., Merz, R. and Blöschl, G. (2007), Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments. Hydrol. Process., 21: 435-446. <https://doi.org/10.1002/hyp.6253>