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 generalmente comienza en otoño y termina en verano del año siguiente. Esta traslación del año calendario tiene la intención de considerar todas las posibles 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 de cada año y termina el 31 de marzo del año siguiente.

## Índices climáticos: ENSO, ONI y SOI

## Reanálisis ERA5

## Modelos estadísticos

# Metodología de modelos estadísticos híbridos

1. Preparación de datos: Seleccionar y preparar los predictores relevantes que se utilizarán en el modelo estadístico. Los predictores pueden incluir índices climáticos como ENSO, ONI, y SOI, así como variables provenientes del modelo hidrológico TUW alimentado por reanálisis meteorológico ERA5, como la precipitación mensual acumulada para el año hidrológico en estudio. Estos predictores deben ser normalizados 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). Utilice el test de Shapiro-Wilks para verificar la normalidad y, si es necesario, aplique transformaciones como logaritmo natural o raíz cuadrada para corregir sesgos en la distribución.

3. Entrenamiento del modelo: Entrene el modelo de regresión lineal múltiple utilizando los predictores seleccionados y normalizados, y la variable dependiente (volumen o logaritmo del volumen). Estime los coeficientes de regresión que minimizan la función de pérdida de mínimos cuadrados ordinarios (MCO).

4. Validación y evaluación del modelo: Valide el modelo utilizando un conjunto de datos de prueba independiente, calculando métricas de evaluación univariable como el coeficiente de determinación (R^2) y la raíz del error cuadrático medio (MSE). También se evaluó el modelo para el conjunto de soluciones (ensembles) ocupando el CRPSS.

5. Pronóstico de volúmenes y caudales: Utilice el modelo entrenado y validado para pronosticar los volúmenes y caudales estacionales en cuencas de Chile Central. Distribuya el volumen utilizando los coeficientes de Pardé de los años más cercanos basados en la distancia euclidiana de los predictores.

6. Aplicación en la toma de decisiones: Los pronósticos de caudales estacionales pueden ser utilizados en la toma de decisiones en diferentes industrias, como la agricultura, para determinar el tipo de cultivo o la extensión de las plantaciones en función de la cantidad de agua disponible durante el periodo de crecimiento.

## 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 intenta establecer 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, también conocida como función de pérdida de mínimos cuadrados ordinarios (MCO).

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.

La utilización de regresiones lineales requiere cumplir supuestos: normalidad, independencia de los predictores. Para verificar la normalidad del predicando (volumen) se aplicó el test de Shapiro-Wilks (SW). Este test 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 (skew) 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 el test de Shapiro-Wilks a la nueva variable. La variable (volumen o logaritmo del volumen) que pase el test 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 2020. 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:

- En el caso de la clasificación, se asigna a la nueva instancia la clase que tenga la mayoría de votos entre sus 'k' vecinos más cercanos.

- En el caso de la regresión, se asigna a la nueva instancia el promedio (o mediana) del 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.

El algoritmo kNN es simple y fácil de implementar, pero puede ser computacionalmente costoso cuando se aplica a conjuntos de datos grandes, ya que requiere calcular la distancia entre la nueva instancia y cada observación del conjunto de datos. Además, la selección del valor de 'k' y la métrica de distancia adecuada son críticas para el rendimiento del algoritmo y pueden requerir ajustes o validación cruzada para encontrar la mejor combinación.

# 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>