1. Datos del Autor:

Apellidos y Nombres: PARIAMACHI RODRÍGUEZ GISELLA ROXANA

Código de alumno: 98.0316.6.UR Teléfono: 943802345

Correo electrónico: gisellaroxana@hotmail.com DNI: 40820572

2. Modalidad de trabajo de investigación:

( ) Trabajo de Investigación ( ) Trabajo académico

( ) Trabajo de suficiencia personal (X) Tesis

3. Título profesional o grado académico

( ) Bachiller (X) Titulo ( ) Segunda especialidad

( ) Licenciado ( ) Magister ( ) Doctor

4. Título del trabajo de investigación

Capacidad Predictiva de los Métodos No Paramétricos para la Predicción del Caudal Medio Mensual del Río Santa, 2018.

5. Facultad de Ciencias

6. Escuela, Carrera o programa: Escuela Profesional de Estadística e informática

7. Asesor:

Apellidos y Nombres: M.Sc. VARELA ROJAS WALTER ALEJANDRO

Teléfono: 943333089

Correo electrónico: varow_9@hotmail.com DNI: 18117012

A través de este medio autorizo a la Universidad Nacional Santiago Antúnez de Mayolo, publicar el trabajo de investigación en formato digital en el Repositorio Institucional Digital, Repositorio Nacional Digital de Acceso Libre (ALICIA) y el Registro Nacional de Trabajos de Investigación (RENATI).

Asimismo, por la presente dejo constancia que los documentos entregados a la UNASAM, versión impresión y digital, son las versiones finales del trabajo sustentado y aprobado por el jurado y son de autoría del suscrito en estricto respeto a la legislación en materia de propiedad intelectual

FIRMA.....

DNI: 40820572

Fecha: 14 de Noviembre de 2018

UNIVERSIDAD NACIONAL

SANTIAGO ANTÚNEZ DE MAYÓLO

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE ESTADÍSTICA E INFORMÁTICA

"CAPACIDAD PREDICTIVA DE LOS MÉTODOS NO PARAMÉTRICOS PARA LA PREDICCIÓN DEL CAUDAL MEDIO MENSUAL DEL RÍO SANTA, 2018"

TESIS GUIADA

PARA OPTAR EL TITULO PROFESIONAL DE

LICENCIADA EN ESTADÍSTICA E INFORMÁTICA

AUTOR:

Bach. GISELLA ROXANA PARIAMACHI RODRÍGUEZ

ASESOR

M. Sc. WALTER ALEJANDRO VARELA ROJAS

HUARAZ – PERÚ

2018

PROGRAMA DE TITULACIÓN PROFESIONAL

MODALIDAD TESIS GUIADA 2018

N° Registro T002

HOJA DE VISTO BUENO

Dr. Ángel Deciderio Mendoza López
Presidente

M.Sc. Jorge Luis Llanos Tiznado
Secretario

Lic. Hugo Walter Maldonado Leyva
Vocal

CAPACIDAD PREDICTIVA DE LOS MÉTODOS NO PARAMÉTRICOS PARA LA PREDICCIÓN DEL CAUDAL MEDIO MENSUAL DEL RÍO SANTA, 2018

ASESORES Y MIEMBROS DEL JURADO

Asesor:

M.Sc. Varela Rojas Walter Alejandro

Miembros del Jurado:

Dr. Mendoza López Ángel Deciderio

M.Sc. Llanos Tiznado Jorge Luis

Lic. Maldonado Leyva Hugo Walter

DEDICATORIA

Agradezco a Dios por darme la oportunidad de concretar mi grado profesional, por darme fuerza, salud y vida.

A mi madre, por ser mi apoyo y mi empuje, por apoyarme en mi carrera y estar siempre a mi lado incondicionalmente.

A mi pequeña hija Nicolle por su cariño a diario y a toda mi familia por siempre darme la mano de alguna u otra manera, porque siempre están para apoyarme.

Índice

Resumen 1

Introducción 2

Capítulo I:

Planteamiento Del Problema

1.1. Descripción de la Realidad Problemática 3

1.2. Formulación del problema 5

1.3. Objetivo de la Investigación 6

1.4. Justificación de la Investigación 6

1.5. Limitación del estudio 7

1.6. Viabilidad del estudio 8

Capítulo II:

Marco Teórico

2.1. Antecedentes de la investigación 9

2.2. Bases teóricas 12

2.3. Definiciones conceptuales 18

2.4. Formulación de hipótesis 24

Capitulo III:

Metodología

3.1. Diseño metodológico 25

3.2. Población 26

3.3. Operacionalización de la variable 27

3.4. Técnica de recolección de datos 28

3.5. Técnicas para el procesamiento de la información 28

3.6. Aspectos Éticos 28

Capítulo IV:

4.1. Resultados 29

Capítulo V:

5.1. Discusión 42

5.2. Conclusiones 44

5.3. Recomendaciones 45

5.4. Referencias Bibliográficas 46

Resumen

Los métodos no paramétricos han demostrado tener una mayor capacidad predictiva en cuanto se refiere a pronósticos de caudales de ríos, en el estudio se implementan dos métodos MARS y Kernel, ambos métodos permiten trabajar con variables predictoras entre ellas la influencia de fenómenos macroclimáticos, para el desarrollo e implementación de las metodologías se realizan con el registro de los caudales medios mensuales del río Santa en el período comprendido entre el año 1980 hasta el año 2017, dicho río es importante para la generación de energía eléctrica, agricultura y para el proyecto Chavimochic. El estudio concluye que el método no paramétrico MARS presenta mayor capacidad predictiva, esto comparado con los errores cuadráticos medios de ambos modelos.

Palabras Claves. Capacidad Predictiva Caudal Río Santa, Métodos No Paramétricos, Mars y Kernel.

Abstract

Nonparametric methods had a greater predictive capacity in terms of forecasting river flows, in the study, two MARS and Kernel methods are implemented, both methods allow us to work with predictive variables, including the influence of macroclimatic phenomena, for the development and implementation of the methodologies, I worked with the average monthly register Santa river in the period covered between 1980 and 2017, said river is important for the generation of electricity, agriculture and for the Chavimochic project. The study concludes that the MARS nonparametric method has a greater predictive capacity; this is compared with the mean square errors of both models.

Keywords: predictive capacity Santa River, Nonparametric methods, Mars and Kernel.

Introducción

El presente tesis es una investigación que tiene por objetivo analizar la capacidad predictiva de los métodos no paramétricos MARS y Kernel para predecir el caudal medio mensual del río Santa.

Dichos métodos no paramétricos han demostrado ser uno de los principales predictores, sobre todo para la predicción de caudales de los ríos, se tomó como referencia los estudios desarrollados en los ríos de Colombía.

El río Santa es una fuente importante de agua y su caudal es aprovechado para la generación de energía eléctrica, a lo largo de la cuenca del rio están presentes varias empresas, negocios, agricultores que aprovechan su caudal para el riego, además el caudal del río Santa es aprovechado por el proyecto chavimochic.

Conocer el caudal del río Santa es muy importante, para la prevención en caso de sobrecarga y desabastecimiento, conocido su pronóstico se tomarán medidas preventivas que permitirá mitigar su efecto adverso.

Por eso es necesario saber que método usar, se estudiaron dos métodos no paramétrico MARS y Kernel, en la presente Tesis se da el proceso para llegar a las conclusiones y recomendaciones finales.

Capítulo I:

Planteamiento Del Problema

1.1. Descripción de la Realidad Problemática

El hombre desde sus inicios y a través de su evolución ha requerido del agua para su subsistencia, el agua es primordial y vital para el hombre porque su cuerpo necesita del agua para su funcionamiento, el agua también es muy importante para la existencia de otros seres vivos como plantas, animales, también como insumo para la industria, así como también para la generación de energía eléctrica. La actividad agraria ha sido por años y sigue siendo una de las principales actividades del hombre, el agua es usada en la agricultura para los sembríos, plantas y consumo de los animales. Existen distintas fuentes de agua, están los lagos, lagunas, ríos, manantiales, aguas subterráneas, entre otras, las actividades de los hombres se han desarrollado alrededor de estás principales fuentes por muchos años, hasta que se ha visto en la necesidad de trasladar el agua a otras extensiones de terreno donde no hay fuentes de agua, por esta razón ha desarrollado sistemas de riego y reservorios para su aprovechamiento.

El Perú tiene muchas fuentes de agua, como lagos, lagunas, ríos, riachuelos, pero como principal fuente de recurso hídrico son sus cuencas hidrográficas. Los ríos en el Perú sirven para reservorios y potabilización para el consumo del hombre, reservorios para riego, el encausamiento para riego de pequeñas y grandes hectáreas de terreno de sembrío, también para la generación de electricidad a través de sus hidroeléctricas.

Uno de los ríos principales del Perú es el río Santa, cuyo caudal es aprovechado por la Central Hidroeléctrica del Cañón del Pato, para la generación de energía eléctrica, esta energía eléctrica es consumida por los pobladores del departamento de Ancash y el departamento de la Libertad por este motivo es muy importante conocer el caudal del río para la toma de decisiones y no desabastecer de energía a estas regiones. El operador de un sistema eléctrico, es quién se encarga de la generación y distribución de la electricidad y debe mantener un equilibrio dinámico continuo entre la energía que se genera y la que se consume, ya que la energía eléctrica no se puede almacenar de forma masiva. Por tanto, se debe hacer una planificación diaria, y con antelación suficiente, de la energía eléctrica que se va a generar y el tipo de centrales que participarán en ese mix de energía, determinando además la reserva de generación necesaria para mitigar los posibles desvíos.

El caudal del río Santa también sirve para el riego de pequeñas y grandes extensiones de áreas de sembríos, un proyecto importante es el Proyecto Especial Chavimochic este proyecto se extiende en la parte baja de la cuenca del río Santa, en el cual se ubica la bocatoma principal, Chao, Virú, Moche y Chicama. El objetivo del Proyecto Especial es el de garantizar el agua de riego en los perímetros de riego de las partes bajas de las cuencas mencionadas.

El área total irrigada beneficiada por el sistema. Es de 144 385 ha, de las cuales se han ganado al desierto 66 075 ha, en las zonas entre los valles. Además

garantiza el suministro de agua a 78 310 ha de tierras de los valles que ya eran cultivados, pero que no tenían el agua garantizada todos los años.

Por eso es muy importante la predicción del caudal del río para la realización de acciones oportunas en las etapas de disminución de caudal y ocasione desabastecimiento. También se tiene que tomar medidas en las épocas de sobrecarga, porque puede producir daños cuando el río se desborda y arrase viviendas y sembríos.

1.2. Formulación del problema

Los métodos no paramétricos han evolucionado y nos dan muestra que son buenos predictores de caudales por los antecedentes del estudio, pero cada río tiene su propia característica y la predicción es distinta una frente a otra por este motivo se necesita evaluar que método es mejor, por este motivo nos hacemos la pregunta:

¿Cuál es la capacidad predictiva de los modelos no paramétricos MARS y Kernel para predecir el caudal medio mensual del río Santa?

1.3. Objetivo de la Investigación

1.3.1. Objetivo General.

Analizar la capacidad predictiva de los métodos no paramétricos MARS y Kernel para predecir el caudal medio mensual del río Santa.

1.3.2. Objetivo Específicos.

- Realizar el diagnóstico del registro de datos de la serie de tiempo del caudal medio mensual del río Santa en el período comprendido desde el año 1980 hasta el año 2017.
- Implementar cada uno de los métodos no paramétricos MARS y Kernel para realizar el pronóstico del caudal medio mensual del río Santa.
- Determinar cuál método no paramétrico (MARS o Kernel) predice mejor del caudal medio mensual del Río Santa.

1.4. Justificación

1.4.1. Justificación Social.

El pronóstico del caudal del río Santa es importante, porque mediante su predicción se podrá, saber épocas de bajo caudal y las épocas de sobrecarga. Con esa información se tomarán medidas para mitigar el impacto. El bajo caudal trae grandes problemas para el sector agropecuario e hidroeléctrico, con pérdida de cultivos desabastecimiento, entre otros.

1.4.2. Justificación Institucional.

Conociendo la problemática y sus consecuencias las autoridades locales, nacionales, empresarios y población puedan tomar medidas para mitigar el impacto, como la realización de proyectos. En épocas de sobrecarga del río, es necesario realizar medidas preventivas como encauzamiento del río, construcción de muros de contención, desalojo de los habitantes de las riveras. Es importante el pronóstico del caudal del río Santa para la generación de energía eléctrica, el operador eléctrico en conjunto con las autoridades debe tomar medidas cuando disminuye el caudal, porque la consecuencia es el desabastecimiento de la energía eléctrica.

1.4.3. Justificación práctica.

Saber con antelación el caudal del río Santa servirá para la toma de decisiones de las empresas privadas, sector público, autoridades locales, autoridades nacionales y población.

1.5. Limitación del estudio

Existe poca investigación y antecedentes de estudio del caudal medio mensual del río Santa, para conseguir y tener acceso a las variables predictoras se usó los antecedentes de otros estudios similares en el país de Colombia. No se trabajó con información de variables predictores de manera libre que ayudaría aún más la eficiencia de la predicción. En la investigación se usó datos accesibles. No se encuentra estudios a profundidad sobre factores que influyen la hidrología del Perú y si existen no son accesibles, los estudios hallados son someros, por eso es

necesario un mayor presupuesto o financiamiento para realizar estudios a profundidad e implementar un sistemas de predicción más eficientes, para ello es necesario contar con investigadores interdisciplinario, con personal especializado, que genere sistemas más específicos para el pronóstico del caudal medio mensual del río y será necesario la búsqueda de financiamiento.

En la investigación no se encontró limitaciones a nivel de recursos humanos porque no se necesitó colaboradores, ni recursos económicos ya que se solventaron los gastos, además se trabajó con investigaciones y metodología usada en el caso de los ríos de Colombia.

1.6. Viabilidad del estudio

El tema de investigación tuvo acceso a información, se trabajó con los registros de las variables, la investigación tuvo como base la data del caudal medio mensual del río Santa, se contó con los recursos y materiales necesarios, con el factor económico, financiero, humano y la disponibilidad de tiempo para desarrollar la tesis en todas sus fases y/o etapas, por los motivos expuestos el estudio tuvo la viabilidad para ejecutarse

.

Capítulo II

Marco Teórico

2.1. Antecedentes de la Investigación

A continuación, se discuten algunos antecedentes que fundamentan el uso del método no paramétricos en la investigación y se analizó la capacidad predictiva a partir de los errores medios cuadráticos.

Sanchez J. y Poveda G. (2004). En su estudio para el VII Seminario Nacional de Hidráulica, llega a la siguiente conclusión:

"Se implementan tres metodologías para el pronóstico de caudales medios mensuales en cuatro ríos de Antioquia (Nare, Guatapé, Guadalupe y Río Grande II): (i) el método no paramétrico MARS ("Multivariate Adaptative Regresion Splines"), (ii) un método ARIMA generalizado no paramétrico, y (iii) el método paramétrico de "Holt–Winters" durante el período 1990-2004. El método MARS permite incorporar la persistencia hidrológica y la influencia de fenómenos macroclimáticos como el ENSO. Se evalúan los métodos de predicción usando intervalos de pronóstico de 1 mes, para horizontes de pronóstico de 12 meses. La bondad del pronóstico se cuantifica a partir del error cuadrático medio y del porcentaje de aciertos por terciles, dependiendo del mes en el cual se inicia

el pronóstico. Los mejores resultados según el criterio del error cuadrático medio los presenta MARS. Se evidencian importantes ganancias en la capacidad de pronóstico al usar la información del Chorro del Chocó (Poveda, 1998) como variable predictora".

Rojo J (2011). En su Tesis presentada expone: *"Se introducen los modelos matemáticos más aplicados en la predicción de caudales medios mensuales en Colombia. Se describen las ecuaciones fundamentales de los diferentes esquemas de regresión y se explica brevemente la estrategia de validación utilizada para comparar la habilidad de pronóstico de los modelos expuestos, además se introduce el procedimiento usado en la estimación los intervalos de confianza y las previsiones probabilísticas. Se proponen algunas técnicas para mejorar la capacidad de pronóstico de los diferentes modelos entre ellos el tratamiento y manejo de la información, la inclusión de esquemas periódicos y el uso de técnicas de descomposición espectral. Finalmente se presenta un protocolo para la predicción de caudales medios mensuales en Colombia usando los métodos de regresión existentes. Los resultados de la aplicación de dicho procedimiento sugieren que se puede lograr un buen ajuste si se combinan esquemas periódicos de pronóstico con métodos espectrales; un modelo híbrido entre el MARS y la transformada en onditas es el más eficiente de los modelos convirtiéndose en el referente ante eventuales mejoras propuestas durante la tesis".*

G. Poveda, et al. (2002). Está investigación afirma lo siguiente: *"Se evalúan diferentes métodos no lineales de predicción de caudales medios mensuales de 6 ríos importantes para la generación de energía eléctrica en Colombia": Alicachín, Betania, Guavio, Nare Salvajina y San Carlos. Se usan ventanas de predicción de mediano y largo plazo de 3, 6 y 12 meses. Los métodos no lineales usados son: MARS ("Multivariate Adaptive Regression ("Predicción en Bandas Espectrales usando Onditas"). Este último es una Splines"), Redes Neuronales (ANN), Regresión Lineal Múltiple (RLM) y PREBEO metodología novedosa usada para potenciar las predicciones y se basa en la Transformada de Onditas ("Wavelets") con el objeto de descomponer las señales en las distintas bandas frecuenciales características de la variabilidad temporal de la hidroclimatología de Colombia. Los resultados indican ganancias muy importantes en la capacidad predictiva en comparación con métodos de predicción tradicionales del tipo lineal y muestran a PREBEO como el método que produce los mejores pronósticos. El análisis de los errores obtenidos permite concluir que PREBEO exhibe los menores errores de predicción, para todos los ríos y todas las ventanas de predicción, siguiendo en su orden RLM y MARS. A la luz del RMS, los métodos ANN y AR(p) muestran los resultados más deficientes. En la mayoría de*

los casos se presenta un sutil incremento del error con el aumento de la ventana de predicción".

2.2. Bases Teóricas

2.2.1. Métodos no Paramétricos

La estimación no paramétrica de la distribución de probabilidades ha llegado a ser un importante objeto de investigación estadística, aunque los primeros intentos de estimación no paramétrica de la densidad comenzaron en la década de los treinta, la preocupación por desarrollar este tema no surge hasta los años ochenta, siendo numerosas las publicaciones de trabajos realizados sobre los aspectos teóricos de este tipo de estimación.

La regresión no paramétrica incluye una fase de validación puesto que una observación dada nunca se utiliza para construir el modelo que se usa para generar la predicción correspondiente. Sin embargo, todavía se puede aislar una sub-muestra que sólo se dedica a la fase de validación, para comprobar la robustez del modelo.

Al contrario de lo que sucede en la regresión lineal clásica, el objetivo no es encontrar un modelo único que describa / explique / prediga un fenómeno, sino obtener un método predictivo eficiente. La regresión no paramétrica es una especie de caja negra. Es numéricamente intensiva, puesto que para cada observación se calcula un nuevo modelo .

2.2.2 Método MARS (Multivariate Adaptative Regresion Splines)

El método MARS ("Multivariate Adaptive Regression Splines"), se basa en la construcción de un modelo multivariado no paramétrico, ajustado adoptivamente por tramos en el tiempo. Friedman (1991). En general, MARS intenta superar algunas de las limitaciones de la modelación no paramétrica y el particionamiento recursivo, planteando algunas generalizaciones a los procedimientos, garantizando modelos continuos con derivadas continuas.

El método MARS es un método de modelación no paramétrico que extiende el modelo lineal incorporando no linealidades e interacciones de variables, que divide el espacio de las variables predictoras en diferentes subregiones. (Montero 2013, Friedman 1991)

El modelo puede escribirse como:

$$y_t = f(x_t) = \beta_0 + \sum_{i=1}^{K} \beta_i B(x_{it})$$

Donde y_t es la variable respuesta en el instante t y $\beta_1$son los pará-metros del modelo para las respectivas variables x_{it}, que van de $i = 1,. . ., k$. El valor β_0 representa al intercepto, las funciones bases $B(x_{it})$ son funciones que dependen de las respectivas variables x_{it},en donde cada $B(x_{it})$ puede escribirse como $B(x_{it}) = max(0, x_{it} - c)$ o $B(x_{it}) = max(0, c - x_{it})$, c es un valor umbral y k representa el número de explicativas, que incluye interacciones de las variables predictoras. Los puntos de partición del espacio y los parámetros de los modelos se obtienen a partir de los datos analizados. El número de funciones base resultantes indica la complejidad del modelo (Salford, 2001).

MARS genera puntos de corte para las diferentes variables. Los puntos son identificados a través de las funciones basales, las que indican el inicio y el término de una región. En cada región en que se va dividiendo el espacio se ajusta una función base de una variable, la cual es lineal. El modelo final se constituye como una combinación de las funciones base generadas. Para determinar estos puntos de corte usa un algoritmo *forward/backward stepwise* por etapas. Primero, mediante el algoritmo *forward stepwise* se genera un modelo sobreestimado con un gran número de funciones base; posteriormente, mediante el algoritmo *backward stepwise*, se eliminan los nodos que menos contribuyen al ajuste global. El algoritmo se detiene cuando la aproximación construida incluye un número máximo de funciones fijadas por el investigador.

Cuando un investigador tiene varios modelos teóricos y se desea identificar el mejor, pueden utilizarse los siguientes estadísticos (Lewis 1991. Silva 2003).

- Criterio de validación cruzada (GCV), propuesto por Friedman (Friedman 1991), medida de ajuste a los datos y penalización, debido a la complejidad del modelo y el aumento de la varianza. Según este criterio, un modelo más sencillo puede ser preferido frente a otro más complejo.

- Coeficiente de determinación (R^2 ajustado) entre el valor observado y el predicho, el cual permite la adecuación del modelo utilizado para las predicciones.

- La proporción media de error absoluto (mpab), que viene dada por los valores observados y el valor predicho:

$$\sum_{i=1}^{n}\{[valor\ observado_{_i} - valor\ predicho_{_i}]|valor\ observado_{_i}\}$$

$$\div n$$

Muestra el porcentaje de error en que se incurre en la predicción en comparación con los datos observados, considerándose que el mejor modelo es aquel que tiene el porcentaje de error más bajo.

2.2.3 Regresiones Kernel (Poveda et al, 2002)

La Regresión Kernel pertenece a la familia de métodos de regresión no paramétricos. En ocasiones se la relaciona con métodos de suavizado (smoothing).

La regresión Kernel estima la variable dependiente continua a partir de un conjunto limitado de puntos de datos por convolución de las ubicaciones de los puntos de datos con una función kernel - aproximadamente hablando, la función del núcleo especifica la forma de "desenfoque" la influencia de los puntos de datos de modo que sus valores pueden ser utilizados para predecir el valor de localidades cercanas.

La regresión Kernel habitualmente se compone de tres fases:

- un paso de ajuste en el que intentamos encontrar la mejor combinación de tipo de modelo, función kernel, y ancho de banda, utilizando una muestra de prueba.
- una fase de validación que permite validar el modelo sobre nuevas observaciones para las que se conoce una predicción;

- una fase de aplicación, en que el modelo se aplica a un nuevo conjunto de datos para los cuales se desconoce la predicción.

Para estimar con el método Kernel se consideran las variables aleatorias bidimensionales, independientes e idénticamente distribuidas $(X_1, Y_1),. . ., (X_n, Y_n)$. Suponiendo que existe una función r (.) la cual modela la relación entre la respuesta Y_i y la co-variable X_i , así: $Y_i = r(X_i) + \varepsilon_i$, y además que $E(\varepsilon i \,|Xi) = 0$, se puede mostrar que la aproximación optima a Y_i , en el sentido de menor error cuadrático medio, está dada por la esperanza condicional,

$$r(x) = E((Y_i|X_i = x), \qquad x\epsilon\mathbb{R}$$

nótese que no se hace ningún supuesto sobre la linealidad de dicha relación.

Si se definen las funciones:

$$\boldsymbol{f(x) = \int f(x,y)dy},$$

la densidad marginal de X, donde $f(x; y)$ es la densidad conjunta, sea:

$$\boldsymbol{\varphi(x) = \int yf(x,y)dy},$$

entonces, la esperanza condicional de Y_i dado $X_i = x$ es,

$$\boldsymbol{r(x) = \int \frac{f(x,y)y}{f(x)}dy = \frac{\varphi(x)}{f(x)}, \qquad x\epsilon\mathbb{R}}$$

Un estimador kernel de $f(x)$, basado en las n observaciones, es:

$$\boldsymbol{f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)},$$

Donde *K(.)* es una función simétrica alrededor de cero, continua, acotada, no necesariamente positiva en todo su soporte e integrable, análogamente, un estimador de $\varphi(x)$ es (Pagan & Ullah 1999, págs. 83-84):

$$\boldsymbol{\varphi_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) Y_i, \qquad x \epsilon \mathbb{R}.}$$

El estimador kernel de *r(x)*, basado en n observaciones, está definido como:

$$\boldsymbol{r_n(x) = \frac{\varphi_n(x)}{f_n(x)}, x \epsilon \mathbb{R};}$$

Rescribiendo esta función se obtiene:

$$\boldsymbol{r_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i,}$$

Donde (Se debe usar la convención $W_{ni}(x) = 0$ si el denominador es cero):

$$\boldsymbol{W_{ni}(x) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right)}.}$$

Así, el estimador de regresión kernel de $E(Yi \mid Xi = x)$ es:

$$\boldsymbol{r_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)}, \qquad x \epsilon \mathbb{R},}$$

el cual puede verse como un promedio ponderado de los Y , donde el peso depende de la distancia entre Xi y x, la cual es cuantificada por la función K(:); en general el procedimiento de ponderación asigna el mayor peso a puntos cercanos a x, y menor o ningún peso a puntos apartados de x. Este es

precisamente el estimador de Nadaraya-Watson; véase Bosq (1998) para detalles adicionales.

Las ponderaciones kernel definen una vecindad de puntos alrededor de x. El suavizador kernel puede ser entendido como un polinomio de ajuste local constante. Extensiones naturales a esta idea son las regresiones polinómicas locales, esto es, ajustar una regresión lineal local, un ajuste cuadrático local, etc. Propiedades teóricas deseables son obtenidas con esta estrategia. En la práctica, muchas veces resulta suficiente con la regresión lineal local. Véase Fan & Gijbels (1996).

2.3. Definiciones conceptuales.

2.3.1. Caudal. (Portal Wikipedia), Se denomina caudal en hidrografía, hidrología y, en general, en geografía física, al volumen de agua que circula por el cauce de un río en un lugar y tiempo determinados. Se refiere fundamentalmente al volumen hidráulico de la escorrentía de una cuenca hidrográfica concentrada en el río principal de la misma. Suele medirse en m^3/seg lo cual genera un valor anual medido en m^3 o en Hm^3 (hectómetros cúbicos: un Hm^3 equivale a un millón de m^3) que puede emplearse para planificar los recursos hidrológicos y su uso a través de embalses y obras de canalización. El caudal de un río se mide en los sitios de aforo. El comportamiento del caudal de un río promediado a lo largo de una serie de años constituye lo que se denomina régimen fluvial de ese río.

2.3.2. Cuenca Hidrográfica. (Wikipedia). Una **cuenca hidrográfica** es un territorio drenado por un único sistema de drenaje natural, es decir, que sus aguas dan al

mar a través de un único río, o que vierte sus aguas a un único lago endorreico. Una cuenca hidrográfica es delimitada por la línea de las cumbres, también llamada divisoria de aguas. El uso de los recursos naturales se regula administrativamente separando el territorio por cuencas hidrográficas, y con miras al futuro las cuencas hidrográficas se perfilan como una de las unidades de división funcionales con mucha más coherencia, permitiendo una verdadera integración social y territorial por medio del agua. También recibe los nombres de hoya hidrográfica, cuenca de drenaje y cuenca imbrífera. Una cuenca hidrográfica y una cuenca hidrológica se diferencian en que la primera se refiere exclusivamente a las aguas superficiales, mientras que la cuenca hidrológica incluye las aguas subterráneas (acuíferos).

2.3.3. ENSO. "El Niño" y "La Niña" son parte de la oscilación del sur El Niño, (ENSO por sus siglas) (https://www.climas.arizona.edu). ENSO es una fluctuación natural de las temperaturas superficiales del mar y la presión superficial del aire del Océano Pacifico Tropical entre el este y oeste. Durante un evento "El Niño," los vientos alisios del este se debilitan, permitiendo que el agua superficial más cálida del Océano Pacifico Tropical del oeste corra hacia el este. Durante un evento "La Niña," estos vientos alisios se intensifican, causando que el agua cálida del este no pueda correr hacia el oeste y por consiguiente que el agua cálida superficial del este se apile. Grandes áreas de baja presión superficial del aire y precipitación convectiva siguen el agua cálida al migrar a través del Océano Pacifico Tropical, alterando los patrones de circulación atmosférica (por ejemplo, la Circulación Walker), que pueden influir el tiempo de todo el mundo.

2.3.4. MEI. El Niño / Oscilación del Sur (ENSO) es el fenómeno más importante asociada a la atmósfera oceánica que causa la variabilidad climática global en

escalas de tiempo interanuales. Aquí intentamos monitorear el ENSO usando el Multivariate ENSO Index (MEI) en las seis principales variables observadas en el Pacífico tropical. Estas seis variables son: presión del nivel del mar (P), zonal (U) y meridional (V) del viento de superficie, temperatura de la superficie del mar (S), temperatura del aire superficial (A) y fracción de nubosidad total del cielo (C). Estas observaciones se han recopilado y publicado en COADS durante muchos años. El MEI se calcula por separado para cada una de las doce temporadas bimensuales (diciembre / enero, enero / febrero, ..., noviembre / diciembre). Después de filtrar espacialmente los campos individuales en grupos (Wolter, 1987), el MEI se calcula como el primer Componente Principal (PC) sin rotar de los seis campos observados combinados. Esto se logra normalizando primero la varianza total de cada campo y luego realizando la extracción de la primera PC en la matriz de covarianza de los campos combinados (Wolter y Timlin, 1993). Para mantener el MEI comparable, todos los valores estacionales están estandarizados con respecto a cada temporada y al período de referencia de 1950-93. El MEI se extiende durante la primera semana del mes siguiente en base a las observaciones de boyas y boyas marinas casi en tiempo real (cortesía de R.W. Reynolds en NCEP) resumidas en estadísticas mensuales de 2 grados compatibles con COADS en CDC. Se debe tener precaución al interpretar el MEI mes a mes, ya que los datos de entrada para las actualizaciones no son tan confiables como COADS, y el MEI se ha desarrollado principalmente con fines de investigación. Los valores negativos del MEI representan la fase ENSO fría, a.k.a. La Niña, mientras que los valores positivos MEI representan la fase cálida ENSO (El Niño).

2.3.5. Precipitación. (Wikipedia). En meteorología, la precipitación es cualquier forma de hidrometeoro que cae de la atmósfera y llega a la superficie terrestre. Este fenómeno incluye lluvia, llovizna, nieve, aguanieve, granizo, pero no virga, neblina ni rocío, que son formas de condensación y no de precipitación. La cantidad de precipitación sobre un punto de la superficie terrestre es llamada pluviosidad, o monto pluviométrico.

La precipitación es una parte importante del ciclo hidrológico, llevando agua dulce a la parte emergida de la corteza terrestre y, por ende, favoreciendo la vida en nuestro planeta, tanto de animales como de vegetales, que requieren agua para vivir. La precipitación se genera en las nubes, cuando alcanzan un punto de saturación; en este punto las gotas de agua aumentan de tamaño hasta alcanzar una masa en que se precipitan por la fuerza de gravedad.

Si bien la lluvia es la más frecuente de las precipitaciones, no deben olvidarse los otros tipos: la nevada y el granizo. Cada una de estas precipitaciones puede a su vez clasificarse en diversos tipos.

2.3.6. PDO. La Oscilación Decenal del Pacífico (DOP).

http://www.cpc.ncep.noaa.gov. A menudo se describe como un patrón de larga duración de la variabilidad climática del Pacífico similar a El Niño (Zhang et al., 1997). Como se vio con el fenómeno más conocido de El Niño / Oscilación del Sur (ENOS), los extremos en el patrón de DOP están marcados por las variaciones generalizadas en la cuenca del Pacífico y el clima de América del Norte. Paralelamente al fenómeno ENSO, las fases extremas de la DOP se clasificaron como cálidas o frías, tal como lo definen las anomalías de la temperatura del océano en el Océano Pacífico nororiental y tropical. Cuando las

TSM son anormalmente frías en el interior del Pacífico Norte y cálidas a lo largo de la costa del Pacífico, y cuando las presiones del nivel del mar están por debajo del promedio en el Pacífico Norte, el DOP tiene un valor positivo. Cuando los patrones de anomalías climáticas se invierten, con anomalías cálidas de TSM en el interior y anomalías de TSM frías a lo largo de la costa de América del Norte, o por encima de las presiones medias del nivel del mar sobre el Pacífico Norte, el PDO tiene un valor negativo.

2.3.7. **QBO**. La quasi-biennial oscillation (QBO) es una oscilación cuasiperiódica del viento zonal ecuatorial entre los vientos del este y el oeste en la estratosfera tropical con un período promedio de 28 a 29 meses. Los regímenes alternos de vientos se desarrollan en la parte superior de la estratosfera inferior y se propagan hacia abajo a aproximad mente 1 km (0.6 mi) por mes hasta que se disipan en la tropopausa tropical. El movimiento hacia abajo de los vientos del este es generalmente más irregular que el de los vientos del oeste. La amplitud de la fase este es aproximadamente dos veces más fuerte que la de la fase oeste. En la parte superior del dominio QBO vertical, dominan los vientos del este, mientras que en la parte inferior, es más probable que se encuentren los vientos del oeste. En el nivel de 30mb, con respecto a los vientos medios mensuales de la zona, el este registrado más fuerte fue de 29.55 m / s en noviembre de 2005, mientras que el más fuerte registrado en el oeste fue de solo 15.62 m / s en junio de 1995.

2.3.8. **SOI.** El Índice de Oscilación del Sur (SOI). **http://www.cpc.ncep.noaa.gov** **.** Es una medida de las fluctuaciones a gran escala de la presión atmosférica entre el Pacífico tropical occidental y oriental (es decir, el estado de la Oscilación austral) durante los episodios de El Niño y La Niña. Tradicionalmente, este índice se ha calculado sobre la base de las diferencias en la anomalía de la presión del aire

entre Tahití y Darwin, Australia. En general, las series de tiempo suavizadas del SOI se corresponden muy bien con los cambios en las temperaturas oceánicas en el Pacífico tropical oriental. La fase negativa del SOI representa una presión de aire por debajo de lo normal en Tahití y una presión atmosférica superior a la normal en Darwin. Los periodos prolongados de valores negativos de SOI coinciden con las aguas oceánicas anormalmente cálidas en el Pacífico tropical oriental, típicas de los episodios de El Niño. Los periodos prolongados de valores positivos de SOI coinciden con las aguas oceánicas anormalmente frías en el Pacífico tropical oriental, típicas de los episodios de La Niña.

2.3.9. Río Santa. Wikipedia. El río Santa es una corriente de recorrido corto de la vertiente del Pacífico, localizado en los Andes norcentrales del Perú con un pequeño recorrido en la costa de la región Ancash. Nace en la laguna Conococha a 4500 msnm y atraviesa el llamado Callejón de Huaylas de sur a norte, continuando en el nudo cordillerano del Cañón del Pato, ya en su tramo final hace de límite natural con el Departamento de La Libertad y toma un curso nordeste-suroeste hasta avenar el Océano Pacífico, al norte del puerto y ciudad de Chimbote.

Su nombre proviene del quechua saqtan ("lo tumba, lo avienta": haciendo alusión a las caídas de los chimbadores en mulo), que se escribía 'sahta' grafía que cambió a 'santa'.

2.4. Formulación de la hipótesis:

El método no paramétrico MARS tiene mejor capacidad predictiva que el método Kernel, para predecir el caudal medio mensual del río Santa.

Capitulo III:

Metodología

3.1. Diseño Metodológico

3.1.1. Tipo de investigación.

De acuerdo a los objetivos planteados el estudio desarrollado es de tipo:

- Cuantitativa: la investigación desarrollada es cuantitativa, ya que los datos con los que se trabajó la investigación son numéricos.
- Retrospectiva: la ocurrencia de hechos se ha dado en el pasado.
- Correlacional: porque en la investigación los datos fueron usados para predecir, para lo cual se midió el grado de relación de varias variables.

3.1.2. Procedimiento de la investigación

En la investigación desarrollada se comparó dos métodos no paramétricos: MARS y Kernel, dichos métodos fueron usados para predecir el caudal medio del río Santa, los pasos a seguir son los siguientes:

Se analizó la data de la serie de tiempo de las variables. Se examinó posibles transformaciones, diferenciación y estacionalidad.

Para desarrollar los modelos no paramétricos se siguió las siguientes fases:

- Preparación de Datos

- Construcción del Modelo
- Evaluación del Modelo

Para la comparación de las metodologías y elegir la mejor predicción se usó el método que tuvo el menor error de predicción.

3.2. Población

La población de la presente investigación estuvo conformada por el registro del caudal medio mensual del río Santa, comprendido desde el año 1980 hasta el año 2017, esto representa un total de 456 observaciones. Se usaron todos los datos de dicho período por tal razón no fue necesario hallar la muestra.

3.3. Operacionalización de variables

VARIABLES	DEFINICIÓN CONCEPTUAL	DIMENSIONES	INDICADORES	INSTRUMENTOS
INDEPENDIENTE CAUDAL	Caudal medio mensual del Río Santa Comprendido entre los años 1980 hasta 2017	CAUDAL	Caudal medio mensual del río santa comprendido desde 1980 hasta 2017. (m^3)	Registro de datos obtenidos del Ministerio de Agricultura de Ancash y la Libertad– Portal Agroancash
DEPENDIENTE Macroclimáticas	Índice Multivariado del ENSO	MEI	Presión a nivel del mar Componentes zonales y meridionales SST (Temperatura en la superficie del mar) Temperatura del aire en la Superficie	Data encontrada en http://www.cpc.ncep.noaa.gov/
	Índice de Oscilación del sur	SOI	El índice de Oscilación del Sur se calcula como la diferencia estandarizada de presiones atmosféricas entre Tahití y Darwin (Trenberth, 1976)	Data encontrada en http://www.cpc.ncep.noaa.gov/
	Oscilación Decadal del Pacífico	PDO	El índice para la PDO se calcula usando análisis de componentes principales de anomalías mensuales en la temperatura superficial del Océano Pacífico Norte.	Data encontrada en https://www.ncdc.noaa.gov/teleconnections/pdo/
	Oscilación Cuasi-Bienal	QBO	El índice de la QBO utilizado fue construido a partir de datos de viento zonal a 30 hPa (Marquardt y Naujokat, 1997).	Data encontrada en http://www.cpc.ncep.noaa.gov/
Hidrológica	Precipitación	PRECICALLEJON PRECINORTE	Precipitación mm/día Corresponde a las precipitaciones de la cuenca del río Santa, zona Callejón de Huaylas y zona Sihuas	Registro de datos obtenidos del Ministerio de Agricultura de Ancash – Portal Agroancash

3.4. Técnica de recolección de datos

El instrumento de recolección de datos, Los datos del caudal medio mensual y precipitación se encuentran en el Portal Regional Agrario Ancash (http://agroancash.gob.pe/agro), en el portal agrario regional de La Libertad (http://www.agrolalibertad.gob.pe) y los datos macroclimáticos se encuentran en el portal web Climatic Protección Center: http://www.cpc.ncep.noaa.gov/.

3.5. Técnicas para el procesamiento de la información

Para el desarrollo del modelo se usó el Software Estadísticos: Software Minitab, Software Salfor Predictive Modeler (SPM Versión 8.2), MatlLab 2018 y la hoja de cálculo Excel.

3.6. Aspectos Éticos

El investigador desarrolló la presente investigación con datos reales y sin alteración alguna de los mismos.

Capítulo IV:

Resultados

4.1. Análisis de la serie de tiempo: Caudal Río Santa:

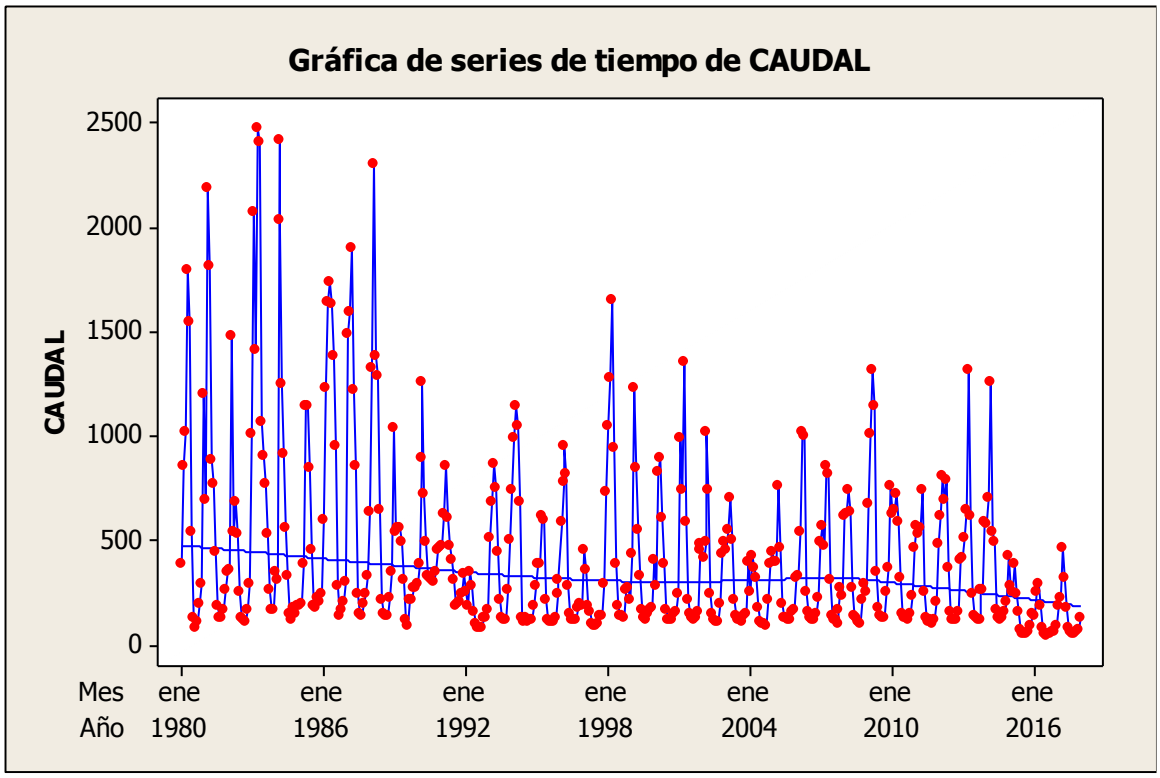

Figura1. Series de tiempo de la Variable: Caudal Río Santa

Estación Cóndor Cerro (Registro de 1980 hasta 2017)

(Promedio mensual, Millones de m3)

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario)..

La gráfica muestra estacionalidad y una tendencia decreciente, el caudal aumenta en los primeros trimestres de cada año, en los años 80 se nota mayor variabilidad y valores máximos de caudal.

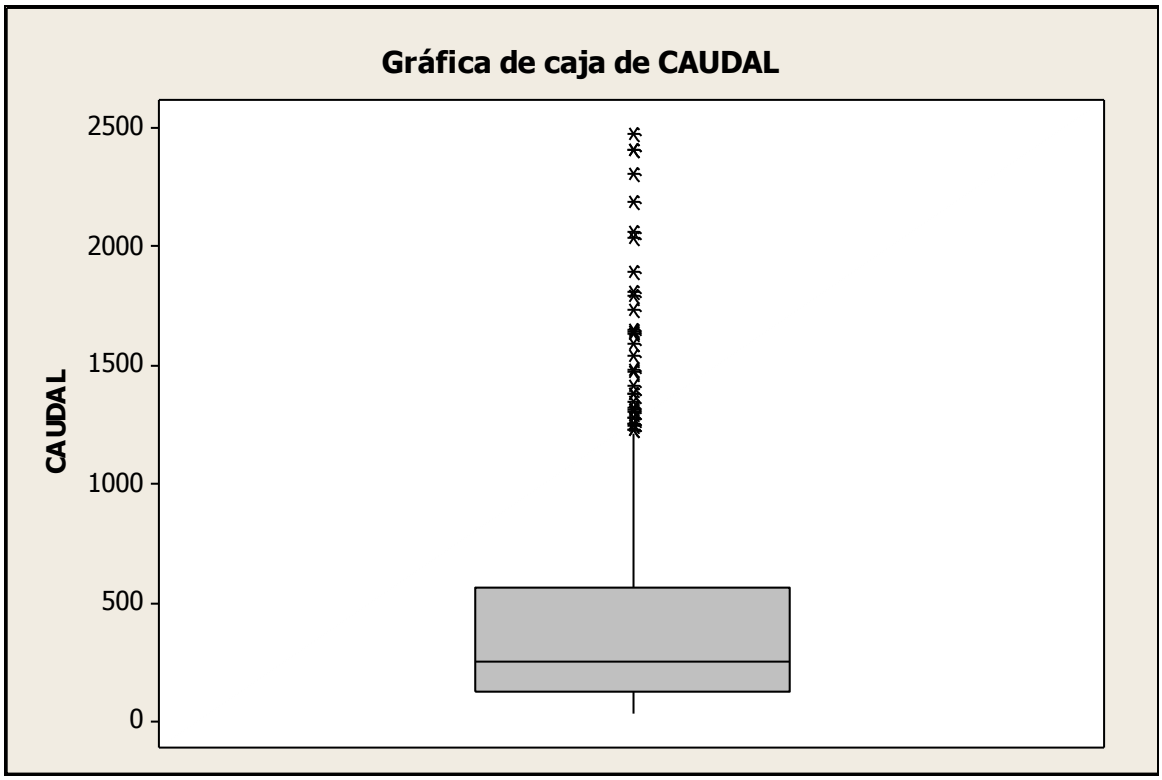

Figura 2. Gráfico de Cajas de la Variable: Caudal Río Santa

Fuente: Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario).

El gráfico de cajas nos muestra que no existe mucha variabilidad, existen datos outliers generados por fenómenos como el niño.

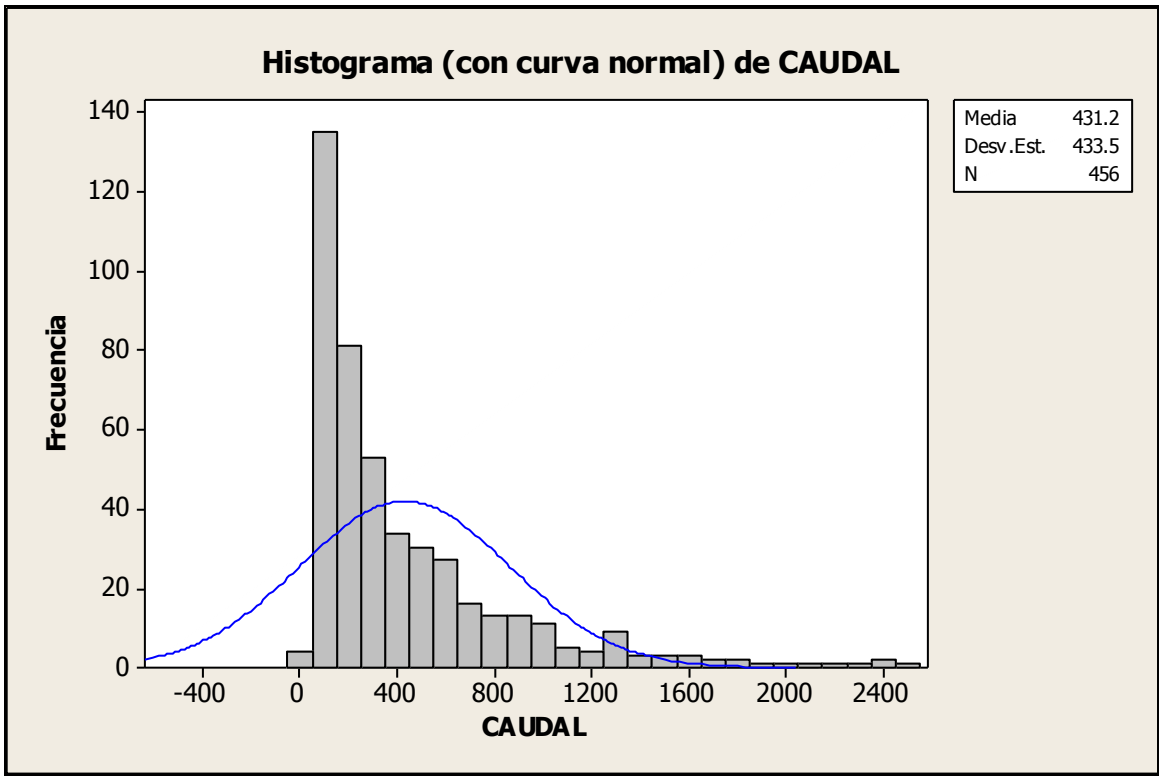

Figura 3. Histograma de Frecuencia de la Variable: Caudal Río Santa

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario).

La curva normal muestra asimetría positiva, la cola de la distribución se alarga para valores superiores a la media.

La serie de tiempo de las variables predictoras, estuvieron listas para su análisis al buscar la data se encontró que la data ya estaba lista para ser trabajada, no fue necesario un mayor análisis o transformación.

4.2. Desarrollo de los modelos no paramétricos,

4.2.1. Preparación de Datos.

La base original se encuentra en archivo Excel, para la construcción del modelo fueron exportados a otros formatos

4.2.2. Construcción y Evaluación del Modelo

METODOLOGÍA MARS

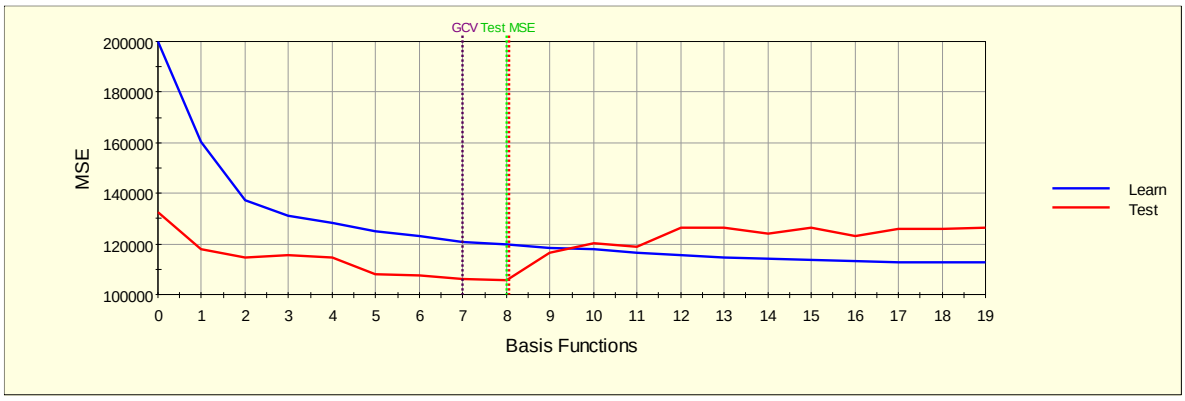

Figura 4. Test MSE

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario) y NOAA Earth System Research Laboratory (portal web)

La gráfica muestra la iteración del modelado, aquí observamos las funciones base que se agregan una a la vez aunque se ha solicitado 30 funciones base, apreciamos que Mars funciona internamente y decidió parar una vez que agregó 19 funciones base debido a algún problema potencial de colinealidad, las dos curvas de rendimiento presentes en la gráfica muestran el error cuadrático medio de color azul y a medida que avanza logra mejores resultados, la curva roja representa el error cuadrático medio de la prueba que tomamos al 20%, vemos que la curva va tomando valores menores pero llega un momento que empieza a empeorar, por este motivo es muy importante el haz verde

que representa el modelo con mejor rendimiento basado en la muestra de la prueba independiente, en nuestro caso recae en la función base 8.

Tabla 1

Medidas de error del modelo

Name	Learn	Test
RMSE	346.02727	324.76879
MSE	119,734.87449	105,474.76695
GCV	134,286.90634	n/a
MAD	238.18264	239.42159
MAPE	0.92320	1.30790
SSY	74,583,716.78981	10,684,248.25004
SSE	44,661,108.18404	8,754,405.80139
R-Sq	0.40119	0.18063
R-Sq Norm	0.40119	0.21431
GCV R-Sq	0.33202	n/a
MSE Adjusted	116,845.82926	n/a
R-Sq Adjusted	0.38803	n/a

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario) y NOAA Earth System Research Laboratory (portal web)

El r-cuadrado en la muestra de prueba del modelo actual es de aproximadamente 18.06%, mientras que el r-cuadrado obtenido es mejor porque obtiene un 40.12%, Por este motivo el MSE es 119,734.8744

Tabla 2
Importancia de las variables

Variable	Importance	-gcv
MEI	100.00000	.151856E+06
PRECINORTE	78.44006	.145097E+06
PDO	42.42480	.137449E+06
QBO	26.85810	.135554E+06
SOI	0.00000	.134287E+06
PRECICALLEJON	0.00000	.133802E+06

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario) y NOAA Earth System Research Laboratory (portal web)

Missing value indicators are nested with their source predictors

La variable que contribuye más al modelo es MEI seguido de la variable PRECINORTE

Construcción del Modelo:

Funciones Bases:

```
BF1 = ( PRECINORTE ne . );
BF4 = max( 0, 4.66442 - PRECINORTE) * BF1;
BF5 = max( 0, MEI - 68);
BF7 = max( 0, QBO - 0.53);
BF8 = max( 0, 0.53 - QBO);
BF10 = ( PRECICALLEJON ne . );
BF13 = max( 0, 0.399908 - PRECICALLEJON) * BF10;
BF14 = max( 0, PDO + 1.49012e-008);
BF19 = max( 0, PDO - 1.74);
BF21 = max( 0, PDO - 1.44);
```

Modelo Mars:

```
Y = 818.628 - 96.4399 * BF4 + 1174.1 * BF5 - 269.811 * BF7
            - 78.7421 * BF8 - 298.163 * BF13 + 217.723 * BF14
            + 1903.8 * BF19 - 1510.99 * BF21;
```

MODEL CAUDAL = BF4 BF5 BF7 BF8 BF13 BF14 BF19 BF21;

Tabla 3.

Modelo final

Basis Function	Coefficients	Variable	Sign	Parent Sign	Parent	Knot
0	818.62811					
4	-96.43989	PRECINORTE	-	+	PRECINORTE_mis	4.66442
5	1,174.10388	MEI	+			68.00000
7	-269.81116	QBO	+			0.53000
8	-78.74208	QBO	-			0.53000
13	-298.16290	PRECICALLEJON	-	+	PRECICALLEJON_mis	0.39991
14	217.72348	PDO	+			-0.00000
19	1,903.79871	PDO	+			1.74000
21	-1,510.98682	PDO	+			1.44000

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario) y NOAA Earth System Research Laboratory (portal web)

La gráfica muestra el modelo de la regresión Mars, las funciones bases, los coeficientes y las variables.

METODOLOLOGÍA KERNEL

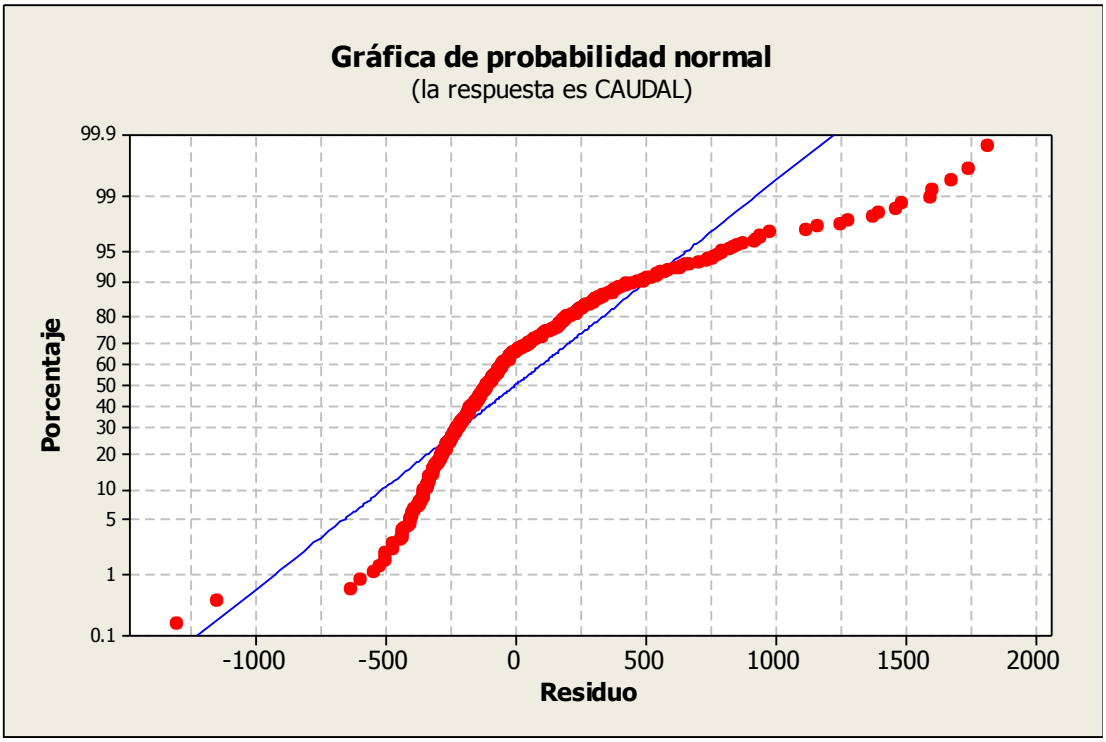

Figura 5. Probabilidad normal: Caudal Río Santa

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario) y NOAA Earth System Research Laboratory (portal web)

Como la variable tiende a una distribución normal, por este motivo se usará el Kernel Gaussiano para la estimación de la densidad.

$$K(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}$$

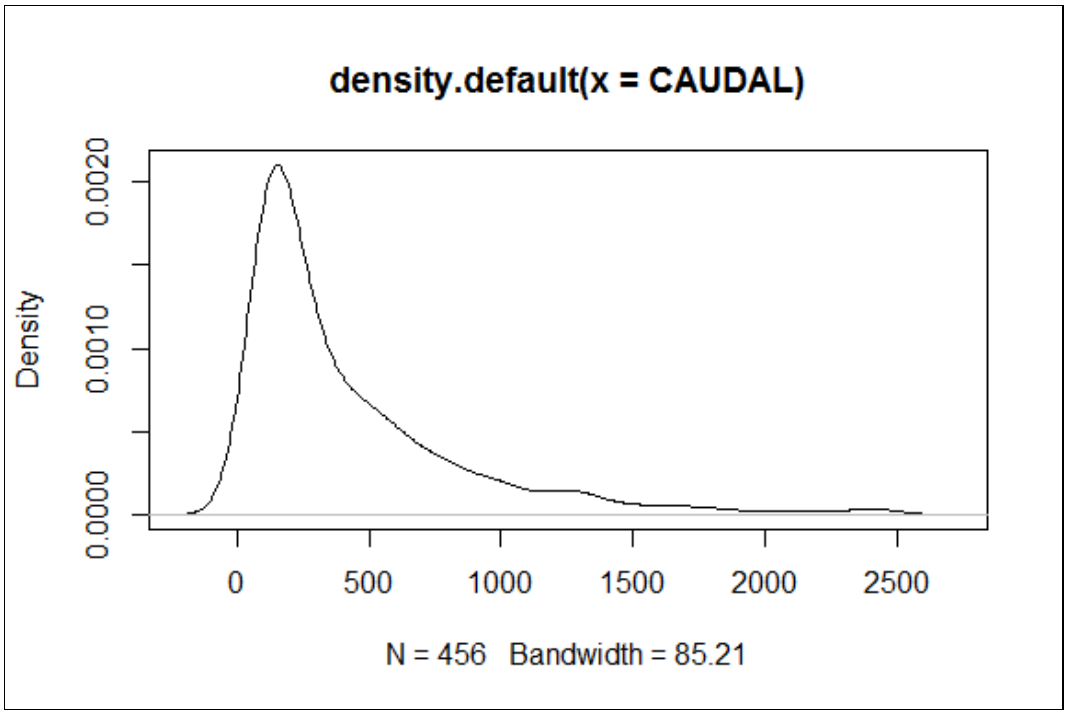

Figura 6. Análisis de Densidad de Kernel: Caudal Río Santa

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario) y NOAA Earth System Research Laboratory (portal web).

La gráfica muestra la función de densidad de Kernel con un ancho de banda de 85.21 con 456 observaciones.

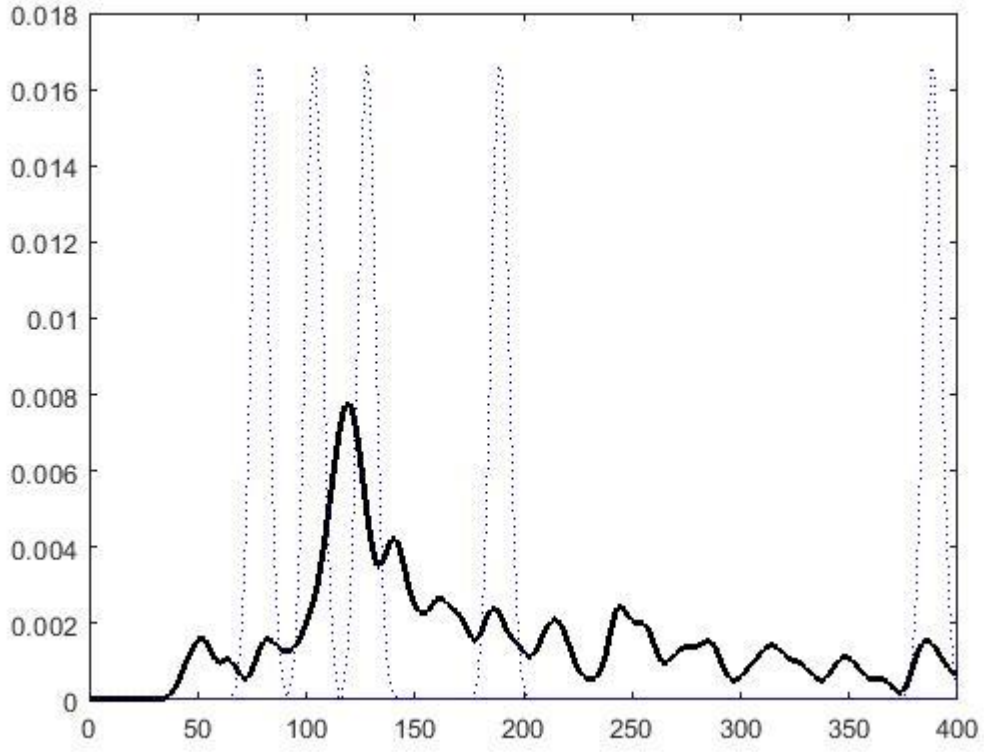

Figura 7. Análisis de Densidad de Kernel: Caudal Río Santa

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario) y NOAA Earth System Research Laboratory (portal web).

La distribución del Kernel construye la función de densidad de probabilidad, creando una curva de densidad de probabilidad individual para cada valor de datos, luego sumando las curvas suaves. Este enfoque crea una función de densidad de probabilidad uniforme y continua para el conjunto de datos.

Las curvas de puntos más pequeñas son las distribuciones de probabilidad para cada valor en los datos de muestra, ajustados para la gráfica. La curva sólida más grande es la distribución general.

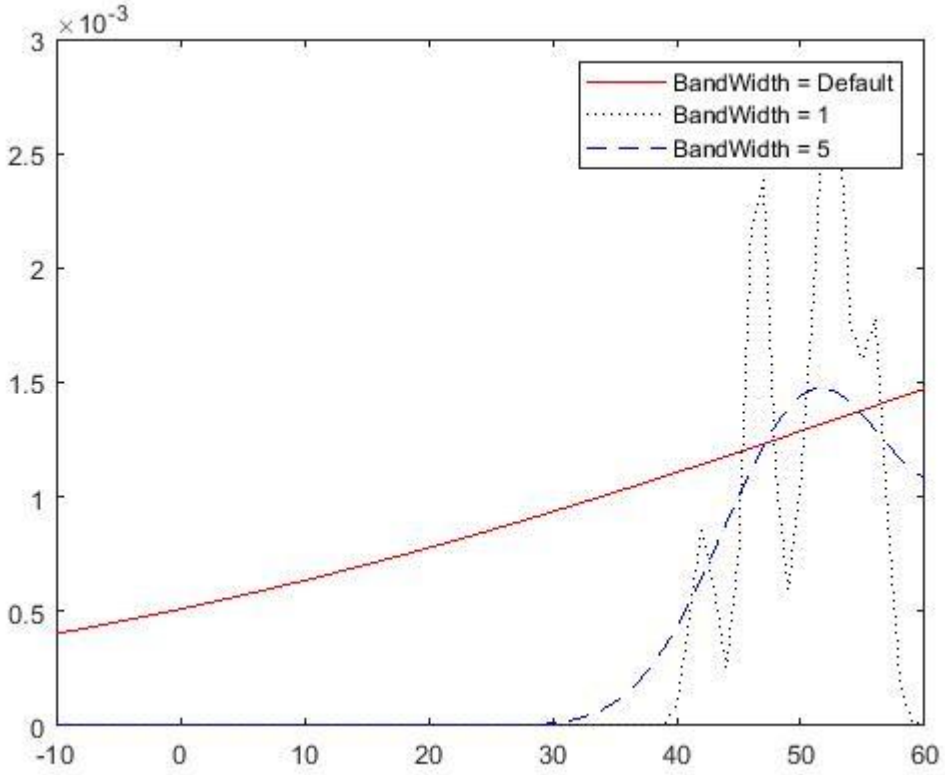

Figura 8. Análisis de Densidad de Kernel: Caudal Río Santa

Fuente: Ministerio de Agricultura de la Región Ancash y La Libertad (Portal Agrario) y NOAA Earth System Research Laboratory (portal web).

La gráfica muestra la curva de densidad de distribución normal para distintos anchos de banda y la que más se aproxima a una función gaussiana es el ancho de Bando 5.

Construcción del Modelo Kernel:

Matlab

```
function kernel
a=xlsread('kernel.xls');% leemos los datos en excel
%%identificamos las variables del archivo excel:
CAUDAL=a(:,2);
MEI=a(:,3);
SOI=a(:,4);
PDO=a(:,5);
QBO=a(:,6);
PRECICALLEJON=a(:,7);
PRECINORTE=a(:,8);
%%%%%%%%%%%%%%%%%%%%%%
%Identificamos las variables dependiente e independientes para el
modelo
Y = CAUDAL;
X = [MEI,SOI,PDO,QBO,PRECICALLEJON,PRECINORTE];
%%%%%%%%%%%%%%%%%%%%%%%
%Reserve el 10% de las observaciones como una muestra retenida.
%Extraiga los índices de entrenamiento y prueba de la definición de
partición.
rng(10)                  %Controla la generación de números aleatorios
N = length(Y);           % longitud de la variable Y, numero de datos
cvp = cvpartition(N,'Holdout',0.1);
idxTrn = training(cvp);
idxTest = test(cvp);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Estandarizamos los datos de entrenamiento y
%entrenar el modelo de kernel de regresión.
Xtrain = X(idxTrn,:);
Ytrain = Y(idxTrn);
[Ztrain,tr_mu,tr_sigma] = zscore(Xtrain); % Standardize the training
data
tr_sigma(tr_sigma==0) = 1;
Mdl = fitrkernel(X,Y)
%%%%%%%%%%%%%%%%%%%%%%%%
%Estandarice los datos de prueba usando la misma media y desviación
estándar de las columnas de datos de entrenamiento.
%Predice respuestas para el conjunto de prueba.
Xtest = X(idxTest,:);
Ztest = (Xtest-tr_mu)./tr_sigma; % Estandariza los datos de prueba
Ytest = Y(idxTest);
YFit = predict(Mdl,Ztest);
%%%%%%%%%%%%%%%%%%%%
%Crea una tabla que contenga los primeros 10 valores de respuesta
%observados y valores de respuesta predichos:
table(Ytest(1:10),YFit(1:10),'VariableNames',...
    {'ObservedValue','PredictedValue'})
%%%%%%%%%%%%%%%%%%%%%
% L Estima la pérdida de regresión del conjunto de prueba utilizando
%la función de pérdida cuadrática media de error.
L = loss(Mdl,Ztest,Ytest)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Modelo de Regresión kernel

```
Mdl = RegressionKernel
```

```
             ResponseName: 'Y'
                  Learner: 'svm'
           Transformation: 'FastFood'
  NumExpansionDimensions:  256
              KernelScale:  1
                   Lambda:  0.002192982456140
            BoxConstraint:  1
                  Epsilon:  32.104658977020009


  Properties, Methods
```

Evaluación del Modelo:

ans = 10×2 table

ObservedValue	PredictedValue
852.68	263.236136758141
118.54	263.885509668451
345.48	263.981506306568
327.92	262.634901707503
188.25	264.009400173661
219.91	263.150853398594
949	263.730067981814
1592.53	264.668277161911
1895.12	264.04321799329
86.75	263.390991129223

L = 1.930048218465660e+05

Error Cuadrático Medio:

193,004.8218

4.3. Comparación de Metodologías.

Las metodologías se comparan a través del error cuadrático medio:

ECM MARS = 119,734.8744

ECM Kernel = 193,004.8218

Observamos que el menor error cuadrático medio nos da el método no paramétrico MARS.

Capítulo V:

5.1. Discusión

- Existen distintos métodos para predecir los caudales medios mensuales de los ríos, la manera de saber que método tiene mejor capacidad predictiva representa el objetivo de la presente investigación, por estudios anteriores desarrollados en otros países podemos apreciar que los métodos no paramétricos logran dicho objetivo como en la investigación desarrollada por Sánchez J. y Poveda G. (2004), indica al igual que en nuestra investigación que el método MARS presenta la mejor capacidad predictiva además le permitió incorporar la persistencia hidrológica y la influencia de fenómenos macro climáticos como el ENSO. También como bondad de pronóstico se usó el error cuadrático medio, los mejores resultados según el criterio del error cuadrático medio los presenta MARS similar a lo concluido en nuestro estudio.

- Existen métodos de mayor complejidad que mejoran la capacidad predictiva, para ello es necesario la búsqueda de financiamiento, para el desarrollo de estos métodos presentan limitaciones en nuestra investigación, por ello se usan métodos no paramétricos que han demostrado ser eficientes para la predicción de caudales, en el estudio desarrollado por Rojo J (2011)., observamos que realiza una combinación de métodos y uno de estos métodos corresponde al método no paramétrico Mars, donde indica que el método MARS es uno de los métodos que predicen mejor el caudal.

- No sólo una limitante es económica también existe limitación de profesionales de otras áreas porque se requieren de conocimientos hidrológicos y de ingeniería, los

métodos no paramétricos pueden ayudar a solucionar este problema porque no es necesario la contratación de profesionales porque el método no paramétricos es menos complejo y sólo requieren las variables respuesta y predictoras para su elaboración. Esto no significa que es menos eficiente, porque es un método que se situa por encima de métodos tradicionales de predicción. Según G. Poveda, et al. (2002). “El análisis de los errores obtenidos permite concluir que PREBEO exhibe los menores errores de predicción, para todos los ríos y todas las ventanas de predicción, siguiendo en su orden RLM y MARS. A la luz del RMS, los métodos ANN y AR(p) muestran los resultados más deficientes”. PREBEO requiere de mayor especialización en otras áreas, se puede apreciar que MARS le sigue en el orden, dicha comparación se realiza con el error medio cuadrático

5.2. Conclusiones

- Observamos que la serie del caudal medio mensual del río Santa en la Figura 1 presenta una tendencia decreciente y estacional, se visualiza el fuerte efecto estacional en cada uno de los 37 años de la serie, con valores modales en febrero-marzo de cada año y depresiones en los meses de invierno.

- Se ha implementado dos metodologías no paramétricas, para la predicción del caudal medio mensual del río Santa MARS y Kernel, Los métodos incorporan explícitamente la influencia del ENSO y otros fenómenos macroclimáticos, que mejoran el pronóstico.

- Los resultados favorecen de manera contundente las metodologías MARS sobre el método Kernel. Específicamente, los resultados obtenidos con el método MARS, muestran menores valores de los errores de predicción (ECM=119,734.8744) en comparación del método Kernel (ECM=193,004.8218).

- El método Mars tiene mayor capacidad predictiva que el método Kernel, porque presenta menores errores cuadráticos medios.

5.3. Recomendaciones

MARS se ve fortalecido por su versatilidad de uso; y su capacidad predictiva depende de la información secundaria utilizada para potenciarlo y de las periodicidades en las que se hace la descomposición de la información hidrológica y macro-climática. Estos resultados deberán servir al sector eléctrico y agrícola de la zona norte del Perú, para un manejo y operación optimas del recurso hídrico del país.

Los resultados también indican que en un clima como el de Perú, que es altamente influenciado por fenómenos macroclimáticos como el ENSO, se deben utilizar metodologías de predicción no paramétricas tales como MARS, las cuales incluyen la influencia de estos fenómenos en sus predicciones de caudales medios mensuales.

Como resultado final, se espera la adopción de la metodología MARS como una herramienta para el pronóstico de series de tiempo (caudales, afluentes) por parte de los investigadores y su aceptación por parte de la comunidad académica como una nueva línea de investigación a desarrollarse, debido a su gran potencial para resolver un sin número de problemas de predicciones, el modelo MARS es muy fácil de implementar y no requiere mayores costos, de esta manera el sector agrícola y la empresa privada podrán definir mejor sus proyecciones y prever cualquier sobrecarga o desabastecimiento del Caudal.

Capítulo VI

Referencias Bibliográficas

- Bosq, D. (1998). *"Nonparametric Statistics for Stochastic Processes*" [Estadísticas no paramétricas para procesos estocásticos]. Springer-Verlag, New York. Recuperado de: http://www.springer.com/la/book/9780387985909.
- Fan, J. & Gijbels, I. (1996), "Local Polynomial Modeling and Its Applications" [Modelado polinomial local y sus aplicaciones]. London. Chapman and Hall.
- Friedman J. (1991) "*Multivariate Adaptive Regression Splines*" [Regresión Multivariante Adaptativo, Splines]. Ann Statist. 19:1–141. Recuperado de: https://projecteuclid.org/euclid.aos/1176347963
- G. Poveda, O. J. Mesa, L. F. Carvajal, C. D. Hoyos, J. F. Mejía, L. A. Cuartas y A. Pulgarín (2002). *"Predicción de Caudales Medios Mensuales en Ríos Colombianos Usando Métodos no Lineales".* Universidad Nacional de Colombia, Esta investigación ha sido financiada con fondos del Instituto Interamericano para la Investigación del Cambio Global (IAI). Recuperado de: http://ciencias.bogota.unal.edu.co/fileadmin/content/geociencias/revista_meteorologia_colombiana/numero06/06_11.pdf
- Lewis P, Stevens J. (1991*). "No linear modeling of time series using multivariante adaptiveregression splines (MARS)".* J Am Stat Assoc. 86:1–36.
- Marquardt, C., y B. Naujokat. (1997). "*An update of the equatorial QBO and its variability*". [Una actualización de la QBO ecuatorial y su variabilidad]. Melbourne Australia. [1st SPARC Gen. Assemb. WMO/TD-No. 814, Vol. 1, 87-90].
- Montero R. (2013). "*Variables no estacionarias y cointegración. Documentos de Trabajo en Economía Aplicada".* España: Universidad de Granada.

- Rojo J. (2011). *"Desarrollo de un sistema experto para la predicción de caudales medios mensuales en Colombia*". Tesis presentada como requisito parcial para optar al título de: Maestría en Ingeniería – Recursos Hidráulicos. Recuperado de: http://www.bdigital.unal.edu.co/5493/1/71536359.2011_1.pdf.
- Salford Systems. (2001). "*Multivariate Adaptive Regression Splines (MARS)"* [Regresión Multivariante Adaptativo, Splines]: user guide.Chapter 3. MARS Basics - Smoothing, splines and knot selection. p. 9-34.
- Sanchez y Poveda (2004), "*Predicción de Caudales medios mensuales en Ríos de Antioquia*". Colombia. VII Seminarío Nacional de Hidráulica. Recuperado de: http://www.bdigital.unal.edu.co/4576/1/AA_4042.pdf.
- Silva C, Alvarado S, Montaño R, et al. (2003). "*Modelamiento de la contaminación atmosférica por partículas: comparación de cuatro procedimientos predictivos en Santiago*". Chile. Biomatemática XIII. 113–127. Recuperado de: http://www.ime.unicamp.br/~biomat/bio13art_12.pdf
- Trenberth, K. E. (1976). *"Spatial versus noise in the Southern Oscillation"* [Espacial versus ruido en la Oscilación del Sur]. [*Mon. Wea. Rev*., 112, 323-332, 1976].

o

UNIVERSIDAD NACIONAL
SANTIAGO ANTÚNEZ DE MAYOLO
"Una Nueva Universidad para el Desarrollo"
ESCUELA PROFESIONAL DE ESTADÍSTICA E INFORMÁTICA

"Año del Diálogo y la Reconciliación Nacional"

ACTA DE SUSTENTACIÓN DE TESIS

Los **Miembros del Jurado** de la Revisión y Sustentación de Tesis de la Escuela Académico Profesional de Estadística e Informática de la Facultad de Ciencias, designados mediante **Resolución de Consejo de Facultad Nº 192-2018-UNASAM-FC**, se reunieron el día lunes 13 de agosto de 2018, a horas 3:00 p.m. en el centro de cómputo Aula G302 de la Facultad de Ciencias en acto público para evaluar la Sustentación de Tesis, presentado por el:

Bachiller : **PARIAMACHI RODRIGUEZ, Gisella Roxana**

Tesis Titulada : **"CAPACIDAD PREDICTIVA DE LOS MÉTODOS NO PARAMÉTRICOS PARA LA PREDICCIÓN DEL CAUDAL MEDIO DEL RIO SANTA, 2018"**

Después de la Sustentación y las respuestas a las preguntas, el Jurado lo declara **Aprobada** para optar el Título Profesional de Licenciado en Estadística e Informática, con el calificativo de **Quince (15)** En señal de conformidad y para constancia, firmamos la presente ACTA, siendo las 4.02 pm del mismo día y año.

Huaraz, 13 de agosto de 2018.

Dr. MENDOZA LÓPEZ, Ángel Deciderio
PRESIDENTE

M.Sc. LLANOS TIZNADO, Jorge Luis
SECRETARIO

Lic. MALDONADO LEYVA, Hugo Walter
VOCAL