UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"COMPARACIÓN DE LOS MÉTODOS REGRESIÓN MULTIVARIADA ADAPTATIVA UTILIZANDO SPLINES (MARS) Y REDES NEURONALES ARTIFICIALES BACKPROPAGATION (RNAB) PARA EL PRONÓSTICO DE LLUVIAS Y TEMPERATURAS EN LA CUENCA DEL RÍO MANTARO"

Presentado por:

KAREN ALEXANDRA LATÍNEZ SOTOMAYOR

TESIS PARA OPTAR EL TÍTULO DE INGENIERO ESTADÍSTICO E INFORMÁTICO

Lima – Perú

2009

DEDICATORIA

A mi abuelo, César, por estar presente aún cuando ya no este con nosotros. Sé que me esta observando.

A mi familia, por comprender el tiempo tomado para el desarrollo de la presente investigación.

A Alberto, por su apoyo incondicional y por darme fuerzas para seguir adelante.

AGRADECIMIENTO

Esta investigación se ha llevado gracias al financiamiento del Instituto Geofísico del Perú mediante el Sub-proyecto “Pronóstico estacional de lluvias y temperaturas en la cuenca del río Mantaro para su aplicación en la agricultura” Nº 2006-00213-AG-INCAGRO/FDSE.

Gracias a mis asesores: el Ingeniero César Menacho, patrocinador de la UNALM y, al Licenciado Raúl Chávez, co-patrocinador por parte del IGP, por su apoyo durante el desarrollo de la presente Tesis.

Así también quisiera agradecer a la Oficina Académica de Investigación de la UNALM por lanzar el concurso de Subvención de proyectos de tesis de pregrado 2008, en el cual la presente tesis salió elegida, y gracias a la subvención se pudo culminar el trabajo sin contratiempos.

RESUMEN

Muchas actividades agrícolas dependen significativamente de la precipitación y la temperatura, afectando la producción y productividad de los cultivos. La cuenca del río Mantaro, (Junín – Perú), está expuesta a una alta variabilidad climática debido a su ubicación y características geográficas. Además, son escasos los trabajos de investigación sobre la elaboración y utilización de pronósticos climáticos para aprovecharlos en la agricultura, por ello esta investigación se plantea ampliar el conocimiento al respecto. Se utilizaron datos de las estaciones de Huayao, Santa Ana, Jaula y Viques, y una vez que estos datos fueron revisados y se eliminaron los valores atípicos extremos se procedió a su análisis con las técnicas: Regresión Multivariada Adaptativa utilizando *Splines* (MARS) y las Redes Neuronales Artificiales *Backpropagation* (RNAB). Las redes neuronales utilizada para el análisis constan de 17 nodos en el caso de las precipitaciones y 15 para las temperaturas tanto mínimas como máximas. Las variables explicativas que se utilizaron en este estudio son variables globales provenientes de información secundaria, siendo recomendable que para próximos estudios se revise la calidad de estas variables. Los *inputs* utilizados tienen un desfase de tres meses (*lag* = 3). Los resultados mostraron que los pronósticos obtenidos al utilizar el modelo MARS tienen menor error que los obtenidos con las RNAB, a excepción de la variable Temperatura Máxima de Huayao en donde la RNAB resultó con menos errores que el modelo MARS.

Palabras clave: *Multivariate Adaptive Regression Splines (MARS), Redes Neuronales Backpropagation, pronóstico, precipitación, temperatura, Huancayo.*

ABSTRACT

Many agricultural activities depend heavily on rainfall and temperature, affecting the production and productivity of crops. The Mantaro river basin (Junín - Peru), it is exposed to a high variability due to location and geographical characteristics. Moreover, little research work on the development and use of climate forecasts for use in agriculture, so this research is to expand knowledge in this regard. We used data from stations Huayao, Santa Ana, and Viques cage, and once these data were reviewed and extreme outliers were removed were taken analysis techniques: Multivariate Adaptive Regression Splines (MARS) and Artificial Network Neural Backpropagation (RNAB). The neural networks used for the analysis consists of 17 nodes in case of precipitation and 15 nodes for both temperatures. The explanatory variables that were used in this study are global variables from secondary information; which is recommended for future studies will review the quality of these variables. The inputs used have a gap of three months (lag = 3). The results showed that the predictions obtained using the MARS model are less error than those obtained with the RNAB. Except for the variable maximum temperature from Huayao that RNAB has the lowest error that the MARS model

Key words: *Multivariate Adaptive Regression Splines (MARS), Backpropagation Neural Networks, forecast, rainfall, temperature, Huancayo.*

ÍNDICE GENERAL

Pág.

ÍNDICE DE CUADROS

Pág.

Pág.

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE ANEXOS

Pág.

I. INTRODUCCIÓN

Para los agricultores es esencial conocer como se va a desarrollar el año climáticamente en lo que se refiere a precipitación y la temperatura ambiental, porque según estos conocimientos deciden cuando empezar a sembrar sus cultivos o cuanto hay que esperar para sembrarlos. Actualmente, estos conocimientos son empíricos, muchos de ellos llevados por las costumbres ancestrales, mediante la observación del medio que los rodea, ellos pueden saber si se aventura un "año bueno" o un "año malo".

Además, se sabe que en las zonas andinas del Perú la agricultura y otras actividades agrícolas dependen significativamente de la precipitación y la temperatura. Es el caso de la cuenca del río Mantaro, (Junín – Perú). La precipitación y la temperatura son factores que afectan la producción y productividad de los cultivos. Si el agricultor siembra y no hay precipitación, parte o la totalidad de la producción se puede perder, empeorando la situación para una población con índices de pobreza y pobreza extrema del 49.9%[1] y 16.5%[2] respectivamente. La incertidumbre con respecto al tiempo es una de las principales causas del escaso rendimiento y pérdidas de cultivo. Adicionalmente, la cuenca del río Mantaro está expuesta a una alta variabilidad climática debido a su ubicación y características geográficas.

En consecuencia, es importante elaborar pronósticos climáticos oportunos, confiables y validos para la cuenca en su totalidad o parcialmente, que permitan a los agricultores tomar decisiones en el mediano plazo (estacional o anual) con el fin de realizar un adecuado planeamiento agrícola de la cuenca.

En la teoría estadística existen varios métodos que permiten modelar una serie de tiempo para obtener pronósticos como: los modelos de regresión paramétrica y no

[1] Fuente: INEI. Encuesta Nacional de Hogares Continua, 2006, Pág. 10.
[2] Fuente: INEI. Encuesta Nacional de Hogares Continua, 2006, Pág. 12.

paramétrica, los modelos de suavización exponencial, Box y Jenkins, etc. Cabe resaltar que el análisis de series de tiempo actualmente es usado para pronosticar el comportamiento de casi todas las variables meteorológicas. Pero este procedimiento solo relaciona la variable estudiada a través del tiempo, y no con otros variables del entorno que pueden estar relacionadas a esta. Si es de interés describir las relaciones entre variables de varias series de tiempo se deben introducir modelos de series de tiempo vectoriales. En cambio, en esta investigación se quiere determinar pronósticos de precipitación y temperaturas extremas basándose en la influencia de variables globales sobre una determinada zona. Sin la necesidad de ingresar al campo de las series temporales, más bien utilizando algún tipo de regresión.

En el desarrollo de la presente investigación se muestra un modelo de regresión no paramétrico conocido como la Regresión Multivariada Adaptativa usando *Splines* (*Multivariate Adaptive Regression Splines*: MARS) y las Redes Neuronales Artificiales *Backpropagation*. En donde MARS es un método de regresión no paramétrica que no hace ninguna suposición sobre la relación funcional entre las variables respuesta y variables explicativas, como en el caso de la regresión múltiple donde los $\varepsilon_i \sim N\left(0,\sigma^2\right)$. MARS construye esta relación basándose en una serie de coeficientes asociados a las funciones base que son totalmente determinadas a partir de la regresión de los datos. El algoritmo que desarrolla el MARS opera como múltiples pedazos de regresiones lineales. También, se obtendrá un modelo de pronóstico aplicando Redes Neuronales Backpropagation con el fin de comparar los resultados con el MARS.

Los datos analizados provienen de las estaciones meteorológicas: Huayao, Jauja, Santa Ana y Viques. Asimismo, se utilizó información secundaria sobre las variables explicativas obtenidas de instituciones especializadas a través de Internet, cabe resaltar que la cantidad de información es limitada para algunas zonas de la cuenca. Además, la información que se obtiene es solamente aplicable a la cuenca del río Mantaro y alrededores, no siendo aplicable para otras zonas del Perú. Y que las variables explicativas se utilizan con *lag* igual a tres, es decir, tres meses de desfase. Este desfase fue seleccionado a partir de un estudio anterior de la autora [33], donde se demostró que se obtienen mejores resultados para los pronósticos de la cuenca del rió Mantaro utilizando

tres y cuatro meses de desfase, en donde los pronósticos a tres meses tuvieron menos errores que los pronósticos a cuatros meses.

1.1. OBJETIVOS

a. Desarrollar el modelo de pronósticos con los métodos MARS y RNAB para cada estación meteorológica en estudio.

b. Comparar los resultados de pronósticos de los modelos obtenidos mediante los métodos: MARS y Redes Neuronales, utilizados en la estimación de la precipitación y las temperaturas extremas de la cuenca del río Mantaro.

c. Pronosticar los valores absolutos de precipitación y temperaturas extremas para los meses de setiembre, octubre y noviembre de 2008 en la cuenca del río Mantaro con los métodos que muestren mejores resultados a partir de la comparación anterior.

II. REVISIÓN BIBLIOGRÁFICA

Morettin [43], Medeiros [40] y Serinaldi [56]; utilizaron series de tiempo en variables que tienen una relación espacio tiempo, como es el caso de las variables de precipitación y temperaturas extremas vistas en este estudio. Estos autores utilizaron diferentes modelos como los autoregresivos, medias móviles, Box y Jenkins para la resolución de sus problemas.

Menacho [41]; usó series de tiempo (Modelos ARIMA) para realizar pronósticos de precipitación total mensual para la ciudad de Huayao, encontrando pronósticos con errores menores a 10%. Luego, Menacho [42]; aplicó modelos ARIMA estacionales (Box y Jenkins) para pronósticos de temperatura y precipitación en la ciudad de Puno que presentaron parámetros significativos.

Hiromi [26], utilizó el análisis multivariado (método WARD) para estudiar la variabilidad de la precipitación en la ciudad de Santa Catarina en Brasil, con datos de precipitación total mensual. En este análisis se desarrolla la relación de la variable respecto a otras. El método WARD utiliza una aproximación al análisis de la varianza para evaluar la distancia entre clusters, intentando minimizar la suma de los cuadrados de los residuales de cada dos hipotéticos clusters que pueden ser formados en cada paso. Su representación gráfica se llama dendograma.

Flores [11]; aplicó el modelo numérico MM5, este modelo tiene el propósito de aplicar las leyes fisicas usando recursos computacionales para la predicción del estado del tiempo y del clima. El Modelo Meteorológico de Mesoescala de Quinta Generación o MM5 es un modelo numérico desarrollado por Universidad Estatal de Pennsylvania (Penn State University, PSU) y el Centro Nacional de Investigaciones Atmosféricas (Nacional Center for Atmospheric Research, NCAR) para la predicción del estado del tiempo. Su

capacidad de trabajar en altas definiciones lo hace ideal para ser usado en un territorio específico por sus características topográficas. Según los resultados se encontraron diferencias entre los valores pronósticos y los observados. El modelo no sigue adecuadamente la tendencia de los datos observados.

Zapata [63]; aplicó regresión semiparamétrica para una sola variable. Según sus conclusiones los modelos semiparamétricos sirven mejor para conjuntos de datos muy grandes, y están son buenas técnicas cuando no se cumplen los supuestos pero sus parámetros son más difíciles de interpretar.

2.1. ASPECTOS PRELIMINARES AL ESTUDIO

ANÁLISIS EXPLORATORIO DE LOS DATOS (A.E.D.)

El A.E.D. es un conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas. Para conseguir este objetivo el A.E.D. proporciona métodos sistemáticos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recolección de los mismos, tratamiento, evaluación de datos ausentes (valores *missing*), identificación de casos atípicos (*outliers*) y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes (normalidad, linealidad, homocedasticidad). (Salvador, [53]).

ETAPAS DEL A.E.D.

1. Preparar los datos para su aplicación de alguna técnica estadística.
2. Realizar un análisis estadístico gráfico y numérico de las variables del problema con el fin de tener una idea inicial de la información contenida en el conjunto de datos así como detectar la existencia de posibles errores en los mismos. Se representa principalmente mediante histogramas, y *boxplot* con el fin de estudiar la forma de la distribución y analizar, en particular, la posible existencia de diversos grupos homogéneos en la muestra.
3. Realizar un examen gráfico de las relaciones entre las variables analizadas y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre ellas.

Evaluar, si fuera necesario, algunos supuestos básicos subyacentes a muchas técnicas estadísticas como por ejemplo, la normalidad, linealidad y homocedasticidad.

4. Identificar los posibles casos atípicos (*outliers*) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores. Los valores atípicos tienen gran influencia en el cálculo de la media, variancia y otros parámetros estadísticos.
5. Evaluar, si fuera necesario, el impacto potencial que pueden tener los datos ausentes (*missing*) sobre la representatividad de los datos analizados.

DATOS ATÍPICOS

Los datos atípicos son observaciones con características diferentes a los demás. Estos tipos de datos no pueden ser caracterizados categóricamente como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y a su vez se debe evaluar el tipo de información que puedan proporcionar. Los datos atípicos cuantitativos producen aumento en la variabilidad.

DATOS AUSENTES

Los datos ausentes son habituales en el análisis climático; es difícil encontrar una investigación en la que no aparece este tipo de datos. En este caso el investigador debe determinar las razones que subyacen en el dato ausente buscando entender el proceso principal de esta ausencia para seleccionar el curso de acción más apropiado.

VALIDACIÓN

La validación de un modelo se puede definir como la demostración de su exactitud para una aplicación concreta. En este sentido, la exactitud es la ausencia de error sistemático y aleatorio. Todos los modelos son, por su propia naturaleza, representaciones incompletas del sistema del que pretenden ser modelo, pero a pesar de esta limitación pueden ser útiles. Se puede encontrar información general sobre el trabajo con modelos matemáticos en diversos libros. Doucet y Sloep [9] ofrecen una introducción completa de la realización de pruebas de modelos. Estos autores distinguen entre confirmación del modelo (es decir, que se demuestra que es digno de crédito; admisible) y verificación del

modelo (es decir, que se demuestra que es verdadero). En el libro de McCullagh y Nelder [38] sobre modelos lineales se describen algunos principios generales para la aplicación de los modelos matemáticos, destacando tres principios para el creador de modelos:

1. Todos los modelos son erróneos, pero algunos son más útiles que otros.
2. No hay que enamorarse de un modelo con la exclusión de otros.
3. Hay que comprobar cuidadosamente el ajuste de un modelo a los datos.

Además, Law y Kelton [34], al abordar la cuestión de la creación de modelos de simulación válidos, creíbles y debidamente detallados, examinan técnicas para aumentar la validez y la credibilidad del modelo. Conviene señalar, sin embargo, que algunos modelos no se pueden validar plenamente, pero es posible validar componentes o módulos del modelo de manera individual. Dee [8] ha señalado cuatro aspectos importantes asociados con la validación de modelos, como sigue:

1. Validación conceptual
2. Validación de algoritmos
3. Validación de códigos informáticos
4. Validación funcional

A continuación se describen los cuatro aspectos mencionados.

La **validación conceptual** se refiere a la pregunta de si el modelo representa con exactitud al sistema que se está estudiando. Habitualmente la validación conceptual es en gran medida cualitativa y la manera de comprobar una es cotejarla con la opinión de expertos con conocimientos científicos diferentes. Se deben presentar y analizar datos experimentales o de observación en apoyo de los principios y los postulados.

La **validación de algoritmos** es la traducción de los conceptos del modelo en fórmulas matemáticas. Un método para evaluar los efectos de procedimientos numéricos es comparar los resultados de distintos métodos utilizados para estimar la incertidumbre de un parámetro, como la superposición de muestras de parámetros obtenidas por los

procedimientos de Montecarlo o de replicación con intervalos de confianza basados en la verosimilitud. La representación gráfica de los resultados puede ser útil, pero se debe utilizar con cautela.

La **validación de códigos informáticos** se refiere a la aplicación de fórmulas matemáticas en el lenguaje informático. Un requisito previo esencial son las buenas prácticas de programación. Los puntos específicos que requieren atención son los posibles efectos de la precisión de la máquina y los factores informáticos específicos en la obtención del modelo. Los informes sobre errores internos del programa informático son fuentes importantes de información, así como la evaluación del producto intermedio.

La **validación funcional** es la verificación del modelo frente a observaciones obtenidas de manera independiente. La evaluación ideal consiste en obtener los datos pertinentes del mundo real y realizar una comparación estadística de los resultados simulados y las observaciones. Para esto se requiere una información más detallada que la disponible habitualmente. En la mayoría de los estudios realizados hasta ahora se ha considerado que la verificación de una gama de riesgos estimados e incidencias observadas era una "validación" suficiente del modelo. Sin embargo, el carácter mismo de las estimaciones de riesgos (probabilidades estimadas) hace posible su utilización como función de verosimilitud para realizar una prueba más oficial de idoneidad.

PRONÓSTICOS

Los pronósticos son premisas o suposiciones básicas de lo que puede suceder, a partir de los cuales se basan la planeación y la toma de decisiones. Las técnicas de pronósticos permiten disminuir la incertidumbre sobre el futuro, permitiendo estructurar planes y acciones congruentes con los objetivos de la organización y permiten también tomar acciones correctivas apropiadas a tiempo cuando ocurren situaciones fuera de lo pronosticado. El punto fundamental en los pronósticos es ser consistente y lograr la menor desviación respecto a los objetivos.

2.2. ANALISIS DE REGRESIÓN

Núñez y Tusell [44] señalan que en la práctica es frecuente encontrarse con situaciones en las que se cuenta con observaciones de diversas variables, y es razonable pensar en una relación entre ellas. El determinar si existe relación es de sumo interés. Por una parte, ello permitiría que siendo conocidos los valores de algunas variables, efectuar predicciones sobre los valores previsibles de otra. Se podría responder con criterio estadístico a cuestiones acerca de la relación de una variable sobre otra.

Es de interés para el autor señalar que el ajuste de un modelo de regresión no se limita a analizar la relación entre dos variables de forma lineal, sino buscar relaciones del tipo:

$$\mathbf{Y} = f\left(\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{p-1}\right) + \boldsymbol{\varepsilon} \tag{2.1}$$

donde $\mathbf{Y}$ es el vector de valores respuesta (variable dependiente) que se espera relacionar con los valores de otras variables (variables explicativas o variables independientes), $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{p-1}$. Frecuentemente $\mathbf{X}_0$ toma el valor constante uno.

Cada variable explicativa es ponderada de tal forma que dichas ponderaciones indican su contribución relativa a la predicción conjunta de la variable de estudio. Las ventajas de usar ponderaciones radican en que estas aseguran la máxima predicción a la vez que facilitan también la interpretación de la influencia de cada variable en la realización de la predicción, aunque la correlación entre las variables explicativas (multicolinealidad) complica el proceso de interpretación.

Existen dos contextos a la hora de elegir un método de regresión para la estimación de los resultados, que son: la regresión paramétrica y la regresión no paramétrica. En el análisis de regresión paramétrica, el investigador presupone una forma de la función de regresión, de la cual solo se desconoce el valor de los parámetros asociados a la misma. Por el contrario, el análisis de regresión no paramétrica no asume un comportamiento del fenómeno a priori, sino que concibe la forma de la curva o función “a partir de lo que

digan los datos". La curva se escoge de entre un conjunto de curvas con ciertas propiedades de continuidad y diferenciabilidad.

COLINEALIDAD Y MULTICOLINEALIDAD

La colinealidad es la asociación medida como correlación entre dos variables explicativas. Multicolinealidad se refiere a la correlación entre tres o más variables explicativas.

Solución a la Colinealidad y Multicolinealidad [5]

1. Aumentar el tamaño de muestra puede reducir un problema de colinealidad.
2. Si se suprimen variables que están correlacionadas con otras, la pérdida de capacidad explicativa será pequeña y la colinealidad se reducirá.
3. Trabajar con el logaritmo de las variables.
4. Utilizar datos de corte transversal.
5. Desestacionalizar las series y quitarles la tendencia.

2.2.1. REGRESIÓN PARAMÉTRICA

a. Regresión Lineal

Gujarati y Hair [19] señalan que el análisis de regresión lineal en su caso simple y múltiple, es quizá la técnica de dependencia más versátil y ampliamente utilizada, aplicable en cualquier ámbito de la toma de decisiones. Los usos de esta metodología van de los problemas más generales a los más específicos, relacionando en cada caso un factor (o factores) con un resultado específico.

NOTACIÓN

Se considera una variable aleatoria Y la cual se supone que el modelo es:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon \tag{2.2}$$

siendo:

1. $\beta_0, \ldots, \beta_{p-1}$, parámetros fijos desconocidos.
2. $X_0, \ldots, X_{p-1}$, variables explicativas no estocásticas, regresoras, cuyos valores son fijados por el experimentador. Frecuentemente X_0 toma el valor constante de uno.
3. ε es una variable aleatoria estocástica inobservable.

La ecuación (2.2) indica que la variable aleatoria Y se genera como combinación lineal de las variables explicativas, salvo en una perturbación aleatoria ε.

El problema que se aborda es el de estimar los parámetros desconocidos $\beta_0, \ldots, \beta_{p-1}$. Para ello se cuenta con una muestra de n observaciones de la variable aleatoria Y, y de los correspondientes valores de las variables explicativas X. Entonces la forma matricial se escribe así:

$$\hat{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\varepsilon} \tag{2.3}$$

donde:

1. $\hat{y}$ es el vector $n \times 1$ de observaciones de la variable aleatoria Y,
2. X la matriz $n \times p$ de valores de las variables explicativas. Su elemento X_{ij} denota el valor que la j-ésima variable explicativa toma en la i-ésima observación,
3. $\hat{\boldsymbol{\beta}}$ el vector de parámetros $\left(\beta_0, \ldots, \beta_{p-1}\right)^t$,
4. $\hat{\varepsilon}$ el vector $n \times 1$ de valores de la perturbación aleatoria ε.

Se denota $\hat{\beta}$ al vector de estimadores de los parámetros, y por $\hat{\varepsilon}$ al vector de $n \times 1$ residuales, definidos por $\hat{\varepsilon} = \hat{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$; es decir, los residuales recogen la diferencia entre los valores muestrales observados y ajustados de la variable aleatoria Y.

SUPUESTOS

Linealidad del fenómeno

El concepto de correlación está basado en una relación lineal, por tanto, el cumplimiento de este supuesto es vital en la aplicación del análisis de regresión lineal. Primero, se grafica ambas variables para determinar la relación funcional existente entre ellas. Segundo, la linealidad se puede examinar fácilmente en los gráficos de residuales versus valores estimados. Si el modelo que se emplea es adecuado, los residuales deberían carecer de estructura y lucir simplemente como ruido. Si los residuales toman alguna forma específica, por ejemplo una curva, indica que amerita una transformación de los datos, o se puede incluir relaciones no lineales al modelo. Una variante del gráfico anterior es cuando se utilizan los residuales estudentizados[3] r_i, en donde la mayoría de residuales caen dentro del intervalo [-3, 3], y debería haber aproximadamente el mismo numero de residuales positivos y negativos.

La expresión que fuerza este supuesto es: $rango(X) = p < N$. Simplemente fuerza la independencia lineal entre las (p) columnas de X. El requerimiento N > p excluye de esta consideración el caso N = p, pues entonces $\vec{y} = X\hat{\beta}$ es un sistema de ecuaciones lineales determinado, y tiene siempre solución para algún vector $\hat{\beta}$ que hace los residuales nulos. Las estimaciones del vector $\vec{\beta}$ se obtendrían entonces resolviendo dicho sistema.

[3] Residual estudentizado: la estudentización es una vieja palabra que significa estandarización. Es una forma de estandarizar no muy lejana al cálculo de los valores z. En la estandarización z, a cada valor se le resta la media y se divide por una única desviación estándar del conjunto de residuos. En la estudentización de los residuos no es necesario restar la media, ya que la media de los residuos es cero. Se divide por una desviación estándar distinta para cada elemento. La desviación utilizada, se calcula utilizando todos los residuos, salvo el que está siendo considerado.

Varianza constante del término de error

La presencia de varianzas iguales es uno de los supuestos que se incumple frecuentemente. El diagnóstico se realiza mediante la utilización de los famosos gráficos de residuales estudentizados versus valores estimados o pruebas estadísticas simples, las cuales indican si la varianza del error es o no constante a través de las observaciones. Si existe varianza constante, los residuales deberán caer dentro de una banda sin ninguna estructura dentro del intervalo [-3,3]. En caso del incumplimiento de esta suposición se puede aplicar el procedimiento de mínimos cuadrados ponderados. También existen transformaciones que estabilizan la varianza.

Independencia de los términos de error

Este supuesto se refiere a que los valores predichos no estén relacionados unos con otros. Para identificar este hecho, se utiliza el grafico de residuales versus el orden en que se tomaron las observaciones. Si los errores son independientes, la forma que adquiere el gráfico es una nube de puntos de comportamiento aleatorio. Técnicas estadísticas como la prueba de Durbin-Watson, la prueba de Box y Pierce, entre otras se usan para detectar autocorrelación de los residuales.

La expresión que involucra al supuesto de varianza constante y de independencia de errores es: $E\left(\vec{\varepsilon}\vec{\varepsilon}'\right)=\sigma^2 I$, requiere que las perturbaciones sean incorrelacionadas (covarianza cero) y homoscedásticas (idéntica varianza).

Normalidad de la distribución del término de error

La normalidad de los errores: $\vec{\varepsilon} \sim N\left(\vec{0},\sigma^2 I\right)$ es la suposición que permite que se puedan efectuar contrastes de hipótesis diversas. Sin embargo, es uno de los supuestos que se incumple con mayor frecuencia. El diagnóstico más simple se puede realizar a través de la observación de un histograma de residuales, donde se puede comprobar visualmente la normalidad de la distribución. Este método es particularmente útil en muestras pequeñas. Un método más eficaz es el gráfico de probabilidad normal, es un método gráfico que

permite determinar si una muestra de datos se ajusta a una distribución propuesta basándose en una análisis visual subjetivo. Originalmente esta gráfica se realizaba sobre un papel especial llamado papel de probabilidad diseñado con las escalas adecuadas para las diferentes distribuciones, como la distribución normal. Si los puntos obtenidos se desvían significativamente de una línea recta, el modelo propuesto no será el apropiado. La línea recta se construye utilizando diferentes puntos, como por ejemplo lo pares ordenados $(\mu, 0.5)$, $(\mu+\sigma, 0.84)$ y $(\mu-\sigma, 0.16)$; sólo bastan dos puntos para trazar una línea recta.

b. Regresión No Lineal

La hipótesis de linealidad entre dos o más variables es una hipótesis muy fuerte que a menudo no se cumple. En estas circunstancias existen otras funciones de tipo no lineal que podrán ajustarse a los datos. En esta sección se realizarán procedimientos equivalentes a los de regresión lineal pero con otras funciones.

En los modelos lineales se exige que la relación entre las variables sea de tipo lineal, que los residuales sean aleatorios, normales y presenten homocedasticidad. En el caso de regresiones no lineales solo se exigen que la relación entre las variables sea del tipo de la función que se ajusta y que los residuales sean aleatorios y normales.

El modelo más general de una regresión no lineal es:

$$\vec{y} = f\left(\vec{\beta}, x_i^{'}\right) + \vec{\varepsilon}\,. \tag{2.4}$$

En la regresión lineal, $\vec{\beta}$ es un vector de parámetros y cada x_i es un vector de predictores, en la regresión no lineal estos vectores de predictores no necesariamente tienen la misma dimensión. La función $f(\cdot)$ que relaciona la variable respuesta con las variables explicativas no es necesariamente una función lineal.

2.2.2. REGRESIÓN NO PARAMÉTRICA

La teoría de los métodos no paramétricos desarrolla procedimientos de inferencia estadística, que no realizan una suposición explícita con respecto a la forma funcional de la distribución de probabilidad de las observaciones de la muestra. Si bien en la estadística no paramétrica también aparecen modelos y parámetros, ellos están definidos de una manera más general que en su contrapartida paramétrica.

El modelo general de la regresión no paramétrica es similar a (2.1), pero la función f es menos específica:

$$\vec{y} = f\left(x_i^{'}\right) + \vec{\varepsilon}$$
$$y_i = f\left(x_{i1}, x_{i2}, \ldots, x_{ip}\right) + \varepsilon_i \qquad (2.5)$$

El objetivo de la regresión no paramétrica es estimar la función $f(\cdot)$, más que estimar los parámetros. Donde f es la curva de respuesta media, llamada también *signal* y el error aleatorio ε es llamado *noise*. La mayoría de los métodos de regresión no paramétrica asumen implícitamente que $f(\cdot)$ es una función continua suavizada. Se pueden utilizar diversas funciones de ponderación, que son los pesos en que se basan los estimadores. La combinación de la función de ponderación y el ancho de la ventana inciden sobre la bondad de la estimación resultante. Es normal suponer que $\vec{\varepsilon} \sim N\left(\vec{0}, \sigma^2 I\right)$, para efectos de hacer inferencias. En regresión no paramétrica, la forma de la función f y la distribución de los errores es determinada usando los datos que se han tomado.

Primero se considera el caso de regresión no paramétrica cuando hay una sola variable predictora y una sola variable respuesta y luego el caso de regresión no paramétrica multidimensional donde hay varias variables explicativas y una sola de respuesta. También existe el caso donde hay varias variables respuesta y varias predictoras, siendo los más conocidos regresión por *projection pursuit* y MARS.

En la regresión no paramétrica simple, donde se emplea un solo predictor:

$$y_i = f(x_i) + \varepsilon_i . \tag{2.6}$$

La regresión no paramétrica simple usualmente es llamada *scatterplot smoothing*, porque una aplicación importante del *scatterplot* es la búsqueda de una curva suavizada de Y contra X.

Ya que en la realidad es muy difícil ajustar un modelo general de regresión paramétrica cuando hay muchos predictores, y, además, es difícil de mostrar el modelo ajustado cuando hay más de dos o tres predictores, se han desarrollado modelos más restrictivos. Uno de estos modelos es el modelo aditivo de regresión:

$$y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \varepsilon_i \tag{2.7}$$

donde la función parcial de regresión $f(\cdot)$ se asume que es suavizada, y es estimada de los datos. Este modelo es substancialmente más restrictivo que el modelo general no paramétrico (2.5), pero menos restrictivo que el modelo de regresión lineal, que asume que todas las variables de la función de regresión parcial son lineales.

Variaciones de este modelo incluyen a los modelos semiparamétricos, en los cuales algunos de los predictores entran linealmente, por ejemplo,

$$y_i = \alpha + \beta_1 x_{i1} + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \varepsilon_i$$

(particularmente útil cuando algunos de las variables son factores), y modelos en los que algunos predictores entran como interacción, que parecen término de mayor dimensión en el modelo, por ejemplo,

$$y_i = \alpha + f_{12}(x_{i1}, x_{i2}) + f_3(x_{i3}) \ldots + f_p(x_{ip}) + \varepsilon_i$$

Todos estos modelos de regresión no paramétrica (y algunos otros, como la *projection-persuit regression*, y la clasificación y regresión de árboles) se discuten en Fox [12] [13].

Entre los modelos más usados en el caso univariado están:

i. El Regresorgrama (Tukey, 1961),
ii. *Running means* (Promedios móviles), *Running medians, Running lines*,
iii. Suavización por Kernels, (Nadaraya-Watson, 1964),
iv. Regresión Polinomial,
v. Regresión local ponderada, LOESS (Cleveland, 1979),
vi. Regresión por *splines*, (Stone y Koo, 1985),
vii. Suavización por *splines*, (Wahba, 1975).

Para el caso multivariado se tienen otros métodos y suavizadores como:

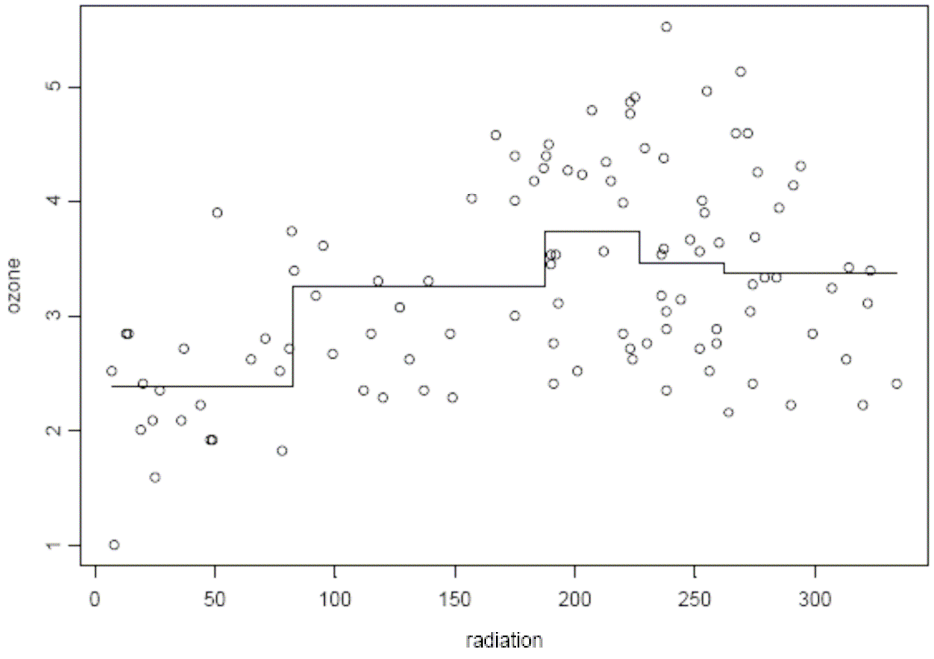
viii. Modelo Aditivos Generalizados, GAM (Hastie y Tibshirani, 1990),
ix. Regresión por Projection Pursuit, PPR (Friedman, Stuelze, 1981),
x. Regresión por árboles, CART (Breiman, Friedman, Olsen y Stone, 1984).

A continuación se describen los diez modelos mencionados.

i. REGRESORGRAMA

En el Regresorgrama se divide el intervalo de los valores de la variable predictora en varios subintervalos (usualmente 5). La amplitud de los subintervalos se elige de tal manera que haya aproximadamente igual número de datos en cada uno de ellos. Luego se promedia los valores de la variable de respuesta en cada subintervalo. Esto determina varios segmentos de línea que al unirse forma el regresorgrama. Lo malo de este estimador es que no es suave porque hay saltos en cada punto de corte. En la Figura 2.1 se muestra un ejemplo de un regresorgrama.

Figura 2.1: Regresorgrama



FUENTE: Edgar Acuña, [1].

Desde un punto de vista práctico, a partir de n observaciones (x_i, y_i), se va a estimar $R(t) = E(Y/X = t)$, agrupando en clases C_i donde se sitúen o no los x_j. Para la clase C_i donde se encuentra el punto t, se efectúa la media y_j correspondientes a los x_j de esta clase C_i.

Denotando como k al número de puntos x_j de la clase C_i, para todo t de C_i, se estima $R(t)$ por:

$$\hat{R}_n(t) = \frac{1}{k}\sum_{j=1}^{k} y_j \quad \text{con} \quad k = \sum_{j=1}^{n} 1_{C_i}(x_j)$$

luego

$$\hat{R}_n(t) = \frac{\sum_{j=1}^{n} 1_{C_i}(x_j) y_j}{\sum_{j=1}^{n} 1_{C_i}(x_j)}$$

En decir, el regresorgrama es una función por escalones que asigna a cada intervalo o escalón el valor medio del mismo. Esta técnica es excesivamente simple ya que asigna el mismo valor a todos los y_i pertenecientes al mismo intervalo disjunto. (Härdle, [21])

ii. *RUNNING MEANS, RUNNING MEDIANS, RUNNING LINES*

Primero, para cada valor de x_i se define una vecindad simétrica $N(x_i)$ que contenga a dicho punto. La simetría esta en el sentido que el número de puntos k es el mismo tanto a la derecha como a la izquierda del punto dado, en los extremos esto no se puede lograr pero se trata de estar lo más cerca posible. Si no es posible tomar k puntos a la izquierda y a la derecha de x_i, se toma tantos puntos como se pueda. Una definición formal de una vecindad simétrica es

$$N(x_i) = \{\max(i-k,1),\ldots,i-1,i,i+1,\ldots,\min(i+k,n)\}.$$

Alternativamente, se puede ignorar la simetría y tomar los r puntos más cercanos a x_i, independientemente del lado en que se encuentren, esto se llama vecinos más cercanos.

Luego se calcula el suavizador por *running means*, *running medians* ó *running lines* en el punto x_i, como se indica en el Cuadro 2.1:

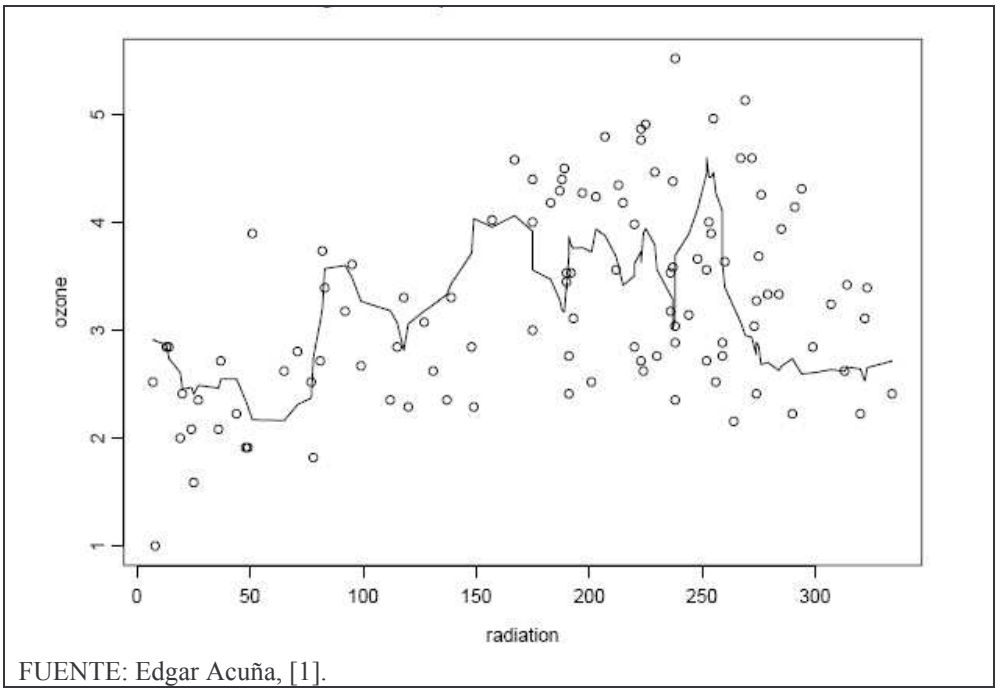
Cuadro 2.1: Suavizadores en el punto x_i

Método	Suavizador $s(x_i)$
Running Means	$s(x_i) =$ promedio de las y's en $N(x_i)$
Running Medians	$s(x_i) =$ mediana de las y's en $N(x_i)$
Running Lines	$s(x_i) =$ valor estimado de la regresión mínimo cuadrática para $x = x_i$ que se obtiene usando los puntos (x_i, y_i) con x_i que cae en $N(x_i)$

FUENTE: Elaboración propia

En la Figura 2.2 se muestra el suavizador por *running means* usando vecindades con $k = 3$ observaciones a cada lado del centro.

Figura 2.2: Suavización por *Running Means*



FUENTE: Edgar Acuña, [1].

Este simple suavizador es también llamado "media móvil", y es popular en series de tiempo uniformemente espaciadas. También es valioso para cálculos teóricos por su simplicidad, pero en la práctica no funciona muy bien. Además, con frecuencia tiende a ser tan rígido que no merece ser llamado suavizador.

iii. SUAVIZACIÓN POR KERNEL

El resultado de un suavizador por *running means* usualmente es poco suave. Esto se debe al hecho que todos los puntos de la vecindad $N(x_i)$, incluyendo los puntos más alejados a x_i, tienen igual peso en el ajuste de y_i. Usando pesos ponderados, donde los pesos más altos son reservados para los puntos cercanos a x_i, la suavización podría mejorar.

Considerando que tanto x como y son aleatorias, se puede escribir $g(x)=E(y/x)=\int yf(y/x)dy$ donde $f(y/x)$ representa a función de densidad condicional de y dado x. Usando la definición de densidad condicional la fórmula se puede reescribir como:

$$g(x)=\frac{\int yf(x,y)dy}{f(x)}. \tag{2.9}$$

En la suavización por Kernel, la función de densidad de x y la función de densidad conjunta de (x,y) son estimadas usando los datos (x_i,y_i) de la muestra, de la siguiente forma:

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{2.10}$$

y

$$\hat{f}(x,y)=\frac{1}{nh^2}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)K\left(\frac{y-y_i}{h}\right). \tag{2.11}$$

Aquí $K(t)$ es llamado el Kernel y es una función no negativa, simétrica con respecto a cero y con valor máximo en dicho punto. Además $\int_{-\infty}^{\infty}K(t)dt=1$. El Kernel actúa como una función de peso, que otorga mayor peso a los puntos cercanos al punto que se va a suavizar y un menor peso a puntos que estén alejados del mismo. La función más usada es el Kernel Gaussiano, que se define como:

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

Se fija un punto x en el dominio de la función regresora $f(\cdot)$ y se define una "ventana" alrededor de ese punto. Si x está en la recta real, generalmente esa ventana es un intervalo de la forma $(x-h,x+h)$ donde el parámetro h es llamado ancho de banda, *bandwidth*, y se estima usando los datos. Los x_i son puntos dentro de la ventana

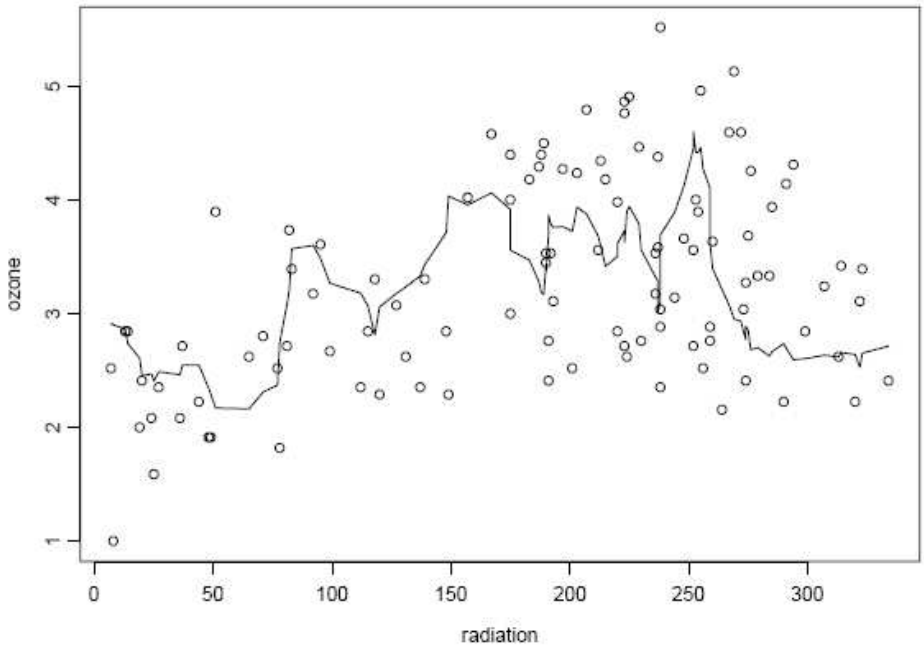
$x-h<x_i<x+h$, es decir $|x-x_i|<h$. Sustituyendo (2.10) y (2.11) en (2.9) se obtiene la estimación por el método de Kernel para g:

$$\hat{g}(x)=\frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} \tag{2.13}$$

El parámetro h puede ser de dos tipos: fijo o variable. Un ancho de ventana fijo, es aquel donde el parámetro h es el mismo en todos los puntos x, y un ancho de "ventana" variable (k-vecinos más cercanos), necesita utilizar una notación más general: $h_k(x)=|x-x_{[k]}|$, donde $x_{[k]}$ es el k-ésimo x_i más cercano a x. En este caso cuanto más densamente estén distribuidos los puntos x_i alrededor de x menor será el ancho de ventana. La longitud de la ventana es aleatoria.

Este estimador de regresión, el suavizador Kernel es también llamado el estimador de Nadaraya-Watson. En la Figura 2.3 se muestra la gráfica de una suavización por Kernels.

Figura 2.3: Suavización por Kernels



FUENTE: Edgar Acuña, [1].

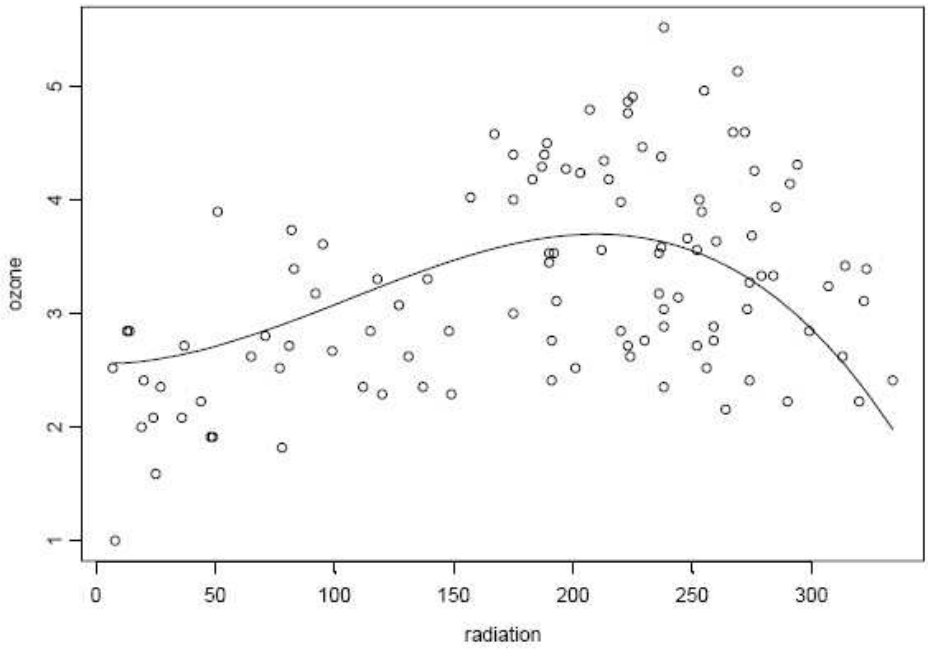
iv. REGRESIÓN POLINOMIAL

En la regresión polinomial, se ajustan los datos (x_i, y_i) para $i = 1, \ldots, n$ a un polinomio de la forma:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon \tag{2.14}$$

n debe ser mayor que $k+1$ de lo contrario se estaría sobreestimando. Donde ε es el error entre el modelo y los datos observados.

Figura 2.4: Ajuste de los datos por un polinomio de grado tres



FUENTE: Edgar Acuña, regresión no paramétrica, Capítulo 9.

Un modelo de regresión polinomial podría tener $x, x^2, x^3, \ldots$ como variables explicativas. El problema con la regresión polinomial es que incurre en un perfecta multicolinealidad tan rápidamente como términos son agregados. En dimensiones bajas, las regresiones polinomiales no son lo suficientemente flexibles para capturar cambios repentinos de la pendiente, en especial en intervalos irregulares. En dimensiones altas, las regresiones polinomiales tienen a fallar debido a la multicolinealidad. (Marsh, [37])

v. REGRESIÓN LOCAL PONDERADA, LOESS

LOESS, originalmente propuesto por Cleveland (1979) y luego desarrollado por Cleveland y Devlin (1988), específicamente denota un método que es más conocido descriptivamente como regresión local ponderada. En cada punto del conjunto de datos un polinomio de grado pequeño es ajustado a un subconjunto de los datos, con los valores de las variables explicativas cerca al punto donde la respuesta está siendo estimada. El polinomio es ajustado usando mínimos cuadrados ponderados, dando más peso a puntos donde la respuesta está siendo estimada y menos peso a los que están más alejados.

En el método de regresión local ponderada, si x_0 es un punto donde se desea hallar la suavización, entonces primero se halla una vecindad usando los k vecinos más cercanos y luego se halla una regresión ponderada en dicha vecindad el valor ajustado de y en x_0 será el valor del suavizador. El valor de la función de regresión para un punto es obtenido evaluando el polinomio local, usando el valor de la variable explicativa para ese punto. El ajuste LOESS es completo después que los valores de la función de regresión han sido calculados para cada uno de los *n* datos. En la Figura 2.5 se muestra el gráfico de la suavización por el método de regresión local ponderada LOESS.

El método trabaja así:

1. Se identifican los k vecinos más cercanos a x_0 y se denota la vecindad por $N(x_0)$.

2. Se calcula la distancia a x_0 del punto más alejado que está dentro de la vecindad $N(x_0)$ y se lo representa por $\Delta(x_0)$.

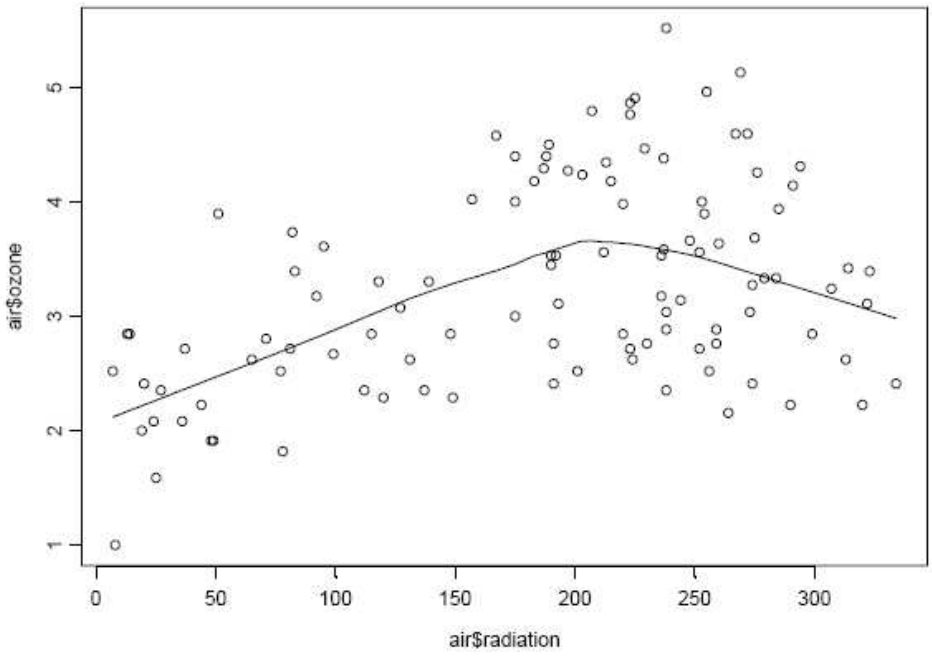
3. Para cada punto x_i en la vecindad $N(x_0)$ se calcula los pesos w_i usando la función peso tri-cúbica definida por:

$$W(t,x_0)=\left[1-\left(\frac{|t-x_0|}{\Delta(x_0)}\right)^3\right]^3, \text{ siempre que } |t-x_0|<\Delta(x_0)$$

4. Se define el suavizador s en x_0 por:

$s(x_0) =$ valor ajustado en x_0 de la regresión ponderada de y versus x en la vecindad $N(x_0)$, usando los pesos definidos en 3

Figura 2.5: Suavización por el método de regresión local ponderada



FUENTE: Edgar Acuña, [1].

vi. REGRESIÓN POR *SPLINES*

La regresión polinomial tiene recursos limitados con respecto a la naturaleza global del ajuste, mientras que en contraste los suavizadores discutidos tienen una naturaleza local explícita. La regresión por *splines* ofrece un compromiso porque representa el ajuste como un polinomio por partes (*piecewise polynomial*[4], Hastie y Tibshirani, [22]). Siendo la regresión por *splines* más flexible que un modelo polinomial y tienen menos probabilidad de generar multicolinealidad en altas dimensiones (Marsh, [37]). Ruppert [51] señala que los métodos *splines* son generalmente más eficientes que los métodos Kernel.

[4] *Piecewise Polinomial*s son funciones que son iguales a los polinomios en intervalos en el dominio. Un ejemplo simple es $|x|$. Donde $y = x$ para $x \geq 0$ y $y = -x$ para $x < 0$.

Los *splines* fueron implementados en estadística por Wahba [61] en 1990, pero sus orígenes se remontan a 1923 gracias a la teoría desarrollada por Whittaker. Un *spline* es simplemente una curva. En matemática, un *spline* es una función especial definida parcialmente por polinomios.

Un *spline* de orden p con k nodos o nudos, $t,\ldots,t_k$ en el intervalo $[a,b]$ es una función que se obtiene dividiendo el intervalo $[a,b]$ en los subintervalos $[X_0,X_1),\ldots,[X_k,X_{k+1}]$ con $X_0=a$ y $X_{k+1}=b$ y usando luego un polinomio de grado menor o igual que p en cada uno de los subintervalos, además, estos pedazos polinomiales deben unirse suavemente en cada uno de los nodos. La función *spline* está definida por:

$$s(x)=\beta_0+\beta_1 x+\ldots+\beta_p x^p+\sum_{j=1}^{K}\beta_{p+j}\left(x-t_j\right)_+^p+\varepsilon \tag{2.15}$$

donde $\beta_0,\beta_1,\ldots,\beta_k$ son constantes a determinar, y $(x-t)_+^p$ se define como:

$$(x-t)_+^p=\begin{cases}(x-t)^p & x\geq t\\ 0 & t<x\end{cases}$$

y es llamada la función potencia truncada de orden p.

Dado que al permitir más nodos, la familia de curvas se vuelve más flexible. Para cualquier conjunto de nodos dados, el suavizador es calculado por regresión múltiple con un conjunto apropiado de vectores base. Estos vectores son las funciones base representando la familia particular de polinomios por partes, evaluado en los valores observados del predictor X.

En resumen, una función *spline* está formada por pedazos de funciones polinomiales de orden p cuyos puntos de unión son los ya mencionado nodos. Una propiedad fundamental de este tipo de funciones es que tienen $p-1$ derivadas continuas.

Funciones Base

Esta clase de métodos incluye a la familia lineal y expansiones polinomiales, pero más importante, incluye una gran variedad de modelos flexibles. El modelo para f es una expansión lineal de funciones base:

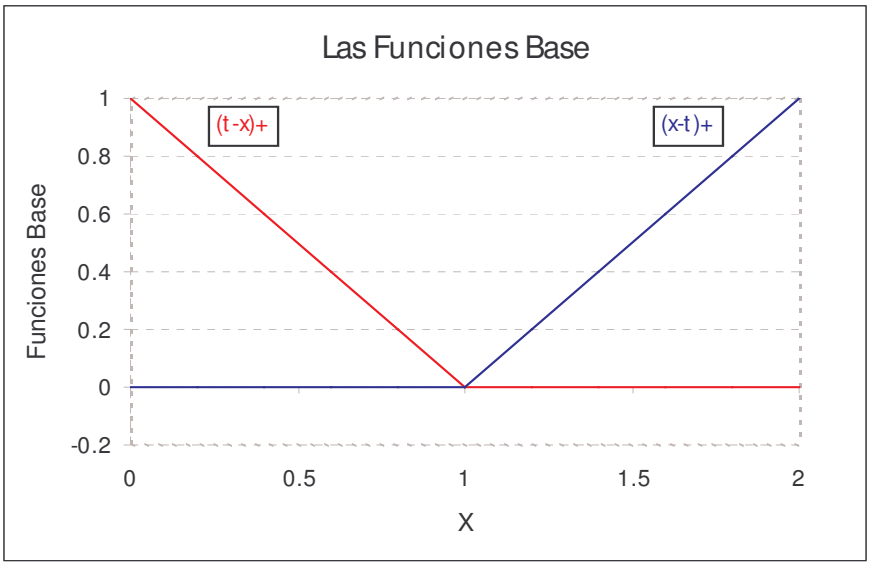
$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x)$$

donde cada h_m es una función del predictor x, y el término lineal aquí se refiere a la acción del parámetro θ. Esta clase cubre una gran variedad de métodos. En algunos casos la secuencia de las funciones base esta preestablecida, como una base para polinomios en x de grado M. Una vez que se determinan las funciones base, los modelos son lineales en estas nuevas variables y el ajuste procede como una regresión lineal. Resumiendo, las funciones base son funciones de dos lados truncados que sirven de base para funciones lineales o no lineales, que se aproxima a las relaciones entre las variables de predicción y respuesta.

Para x unidimensional, los polinomios *spline* de grado K pueden ser representados como una secuencia apropiada de M funciones base *spline*, una vez determinado por $M-K$ nodos. Esto produce funciones que son polinomios por partes de grado K entre nodos, y son unidas por continuidad con grado $K-1$ en los nodos. Unas funciones base intuitivas serían las siguientes funciones: $b_1(x)=1$; $b_2(x)=x$ y $b_{m+2}(x)=(x-t_m)_+$; con $m=1,\ldots,M-2$, donde t_m es el m-ésimo nodo, y $(z)_+$ denota la parte positiva de los productos tensores de la base *spline*, pueden ser usados como predictores (*inputs*) con dimensión mayor a uno. El parámetro θ es el grado total del polinomio o el número de nodos en el caso de *splines*. (Hastie, *et.al*., [23])

En la Figura 2.6 el parámetro t es el nodo de las funciones base; estos nodos también están determinados a partir de los datos. El signo "+" junto a los términos $(t-x)_+$ y $(x-t)_+$ indican que sólo los resultados positivos de las respectivas ecuaciones son considerados, de lo contrario las respectivas funciones se evalúan en cero.

Figura 2.6: Representación de las Funciones Base



FUENTE: Adaptado de Hastie *et al* 2001

CASO LINEAL UNIVARIABLE

Este tipo de modelo de regresión *spline* encuentra pocas aplicaciones debido a que se necesitan muchos nodos para obtener un buen ajuste. Supóngase que se tiene n observaciones de la forma $(x_1, y_1),\ldots,(x_n, y_n)$ y el modelo es:

$$y = s(x) + \varepsilon$$

donde los ε son variables aleatorias independientes y tienen una distribución normal con media cero y varianza σ^2. El objetivo es estimar f basándose en las observaciones, usando un modelo *spline* lineal. Dada una secuencia de nodos $(t_1, t_2,\ldots, t_k)$, se puede escribir el modelo *spline* lineal como sigue:

$$s(x) = \beta_0 + \beta_1 x + \sum_{j=1}^{K} \beta_{j+1} (x - t_j)_+ \qquad x \in R \tag{2.16}$$

Por tanto, dado un conjunto de nodos estimados, esta colección de modelos (2.16), son un espacio lineal que se denotará como G. Las funciones $\{1, x, (x-t_1)_+,\ldots,(x-t_k)_+\}$ constituyen una base para este espacio. La ventaja de esta representación es que permite conectar cada nodo con una función.

CASO POLINOMIAL UNIVARIABLE

Este caso se representa de la forma más general como en (2.15). La función *spline* más utilizada en los diferentes estudios es el *spline* cúbico debido a que tiene una primera y segunda derivadas continuas.

En particular el *spline* cúbico esta dado por:

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{K} \beta_{3+j} \left(x - t_j\right)_+^3 + \varepsilon \tag{2.17}$$

Las funciones $1, x, x^2, \ldots, x^p, \left(x - t_1\right)_+^p, \ldots, \left(x - t_k\right)_+^p$, forman una base de funciones del *spline* que lamentablemente tiende a crear problemas de multicolinealidad, por lo que se recomienda explorar otras bases, como los B-*spline* o los *natural splines*. Para referencias básicas de B-*spline* revisar De Boor [7] y Schumaker [55].

Entonces la regresión por *spline* utilizando K nodos t_j se define por:

$$y = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{j=1}^{K} \beta_{p+j} \left(x - t_j\right)_+^p + \varepsilon$$

El modelo puede ser linealizado mediante transformaciones y hay que estimar $p + K + 1$ parámetros. El problema es determinar el número de nodos K. La idea básica es añadir el máximo número de nodos posibles y luego ir eliminando uno por uno, tratando de maximizar la bondad de predicción del modelo minimizando su complejidad.

Decidir el número y posición de los nodos, así como el orden de la regresión polinomial en cada segmento no es una tarea fácil. Es recomendable que el número de cortes sea el menor posible. Se debe tener en especial cuidado en este punto ya que la gran flexibilidad del método puede ocasionar problemas de sobrestimación.

vii. SUAVIZACIÓN POR *SPLINES*

Barrientos, Olaya y González [2]; menciona que en su mayoría, los *splines* han sido estudiados más en el marco del análisis numérico como método de interpolación que en el estadístico como método de suavización.

El suavizador por *splines* se obtiene minimizando:

$$\sum_{i=1}^{n}\left(y_i - s\left(x_i\right)\right)^2 + \lambda \int \left[s''(t)\right]^2 dt \,. \tag{2.18}$$

El primer término de (2.18) es una media de la bondad de ajuste del modelo y el segundo término es una media del grado de suavidad (medida en la que se desea suavizar el gráfico de dispersión de los datos). El parámetro de suavidad λ es positivo y rige el intercambio entre la suavidad y bondad de ajuste del suavizador. Cuando $\lambda = \infty$ se obtiene una aproximación polinomial y cuando $\lambda = 0$ se obtiene una regresión por *spline*.

Considerando que $X_i^t = \left\{1, X_i, \ldots, X_i^p, \left(X_i - t_1\right)_+^p, \ldots, \left(X_i - t_k\right)_+^p\right\}$

$$\mathbf{X} = \begin{bmatrix} X_1^t \\ \vdots \\ X_n^t \end{bmatrix} \qquad \text{y} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{k+p} \end{bmatrix}$$

Entonces la ecuación (2.18) se puede escribir como:

$$\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^t \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) + \lambda \boldsymbol{\beta}^t \boldsymbol{\Omega} \boldsymbol{\beta} \tag{2.19}$$

donde $\boldsymbol{\Omega}$ es una matriz tal que $\left\{\boldsymbol{\Omega}\right\}_{jk} = \int X_j''(t)\, X_k''(t)\, dt$

Reinsch [47] mostró que existe un único mínimo de (2.18), y que este es un *spline* cúbico natural con nodos en los únicos valores de x_i.

Minimizando la expresión (2.19) con respecto a β se obtiene que

$$\hat{\boldsymbol{\beta}}(\lambda) = \left(\mathbf{X}'\mathbf{X} + \lambda\boldsymbol{\Omega}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{2.20}$$

este es un resultado muy similar a la Regresión Ridge.

Recordando que $f = \mathbf{X}\boldsymbol{\beta}$ se tendría que:

$$\hat{f} = \mathbf{X}'\left(\mathbf{X}'\mathbf{X} + \lambda\boldsymbol{\Omega}\right)^{-1}\mathbf{X}'\mathbf{y}$$

donde $\mathbf{X}'\left(\mathbf{X}'\mathbf{X} + \lambda\boldsymbol{\Omega}\right)^{-1}\mathbf{X}' = \mathbf{H}(\lambda)$, $\mathbf{H}(\lambda)$ es la llamada matriz HAT.

Los grados de libertad de suavización son igual a la traza de $\mathbf{H}(\lambda)$. Esto es bastante similar al número de variables explicativas en un modelo de regresión.

ELECCIÓN DEL PARÁMETRO λ

Se pueden utilizar dos métodos para encontrar el valor óptimo de λ

1. Usando Validación Cruzada, *Cross Validation* (CV).

   Sea $s\left(x;\hat{\boldsymbol{\beta}}(\lambda)\right)$ el *spline* ajustado con parámetro de suavización λ.

   Sea $s_{-i}\left(x;\hat{\boldsymbol{\beta}}(\lambda)\right)$ el *spline* ajustado con parámetro de suavización λ pero sin usar la observación (x_i, y_i) entonces se define la función de validación cruzada como:

$$CV(\lambda) = \sum_{i=1}^{n}\left\{s_i - s_{-i}\left(x_i, \hat{\beta}(\lambda)\right)\right\}^2$$

   El valor de λ que minimice a $CV(\lambda)$ es el valor que se escoge como parámetro de suavización.

El problema con el CV es que es computacionalmente complejo de calcular. Una mejor alternativa es el criterio GCV.

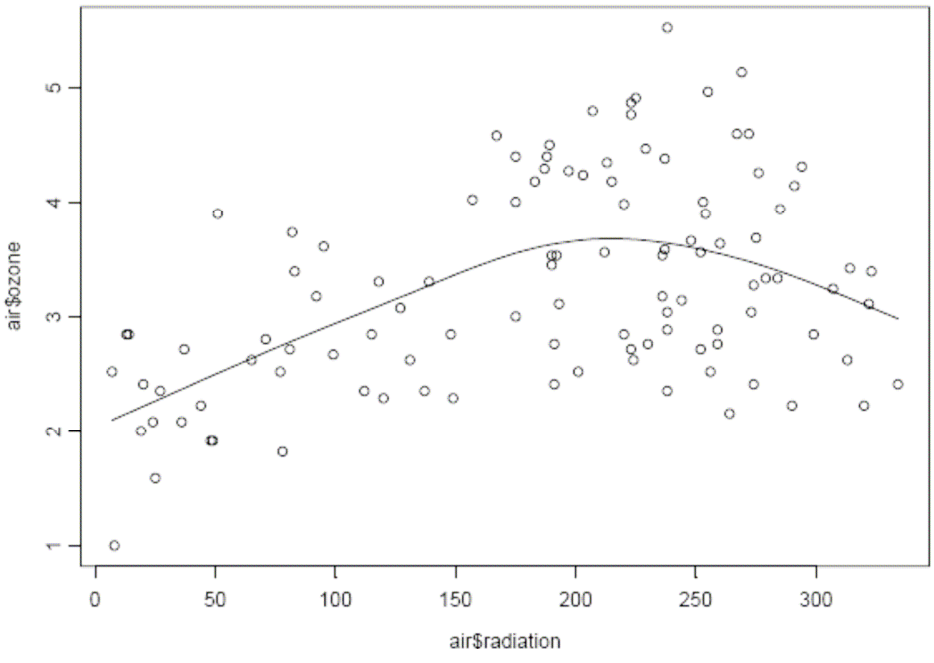
2. Usando Validación Cruzada Generalizada, *Generalized Cross Validation* (GCV).

El GCV en realidad no es una generalización del CV sino es una aproximación, y se define como:

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}\left\{y_i - s\left(x_i, \hat{\beta}(\lambda)\right)\right\}^2}{\left[1 - tr\left(H(\lambda)/n\right)\right]^2}$$

El valor de λ que minimice el GCV es el que se escoge como parámetro de suavización.

Figura 2.7: Suavización por el método *spline*



FUENTE: Edgar Acuña, [1].

viii. MODELOS ADITIVOS GENERALIZADOS, *Generalized Aditive Models* (GAM)

Un modelo aditivo generalizado es de la forma:

$$y = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) + \varepsilon$$

donde los f_j son estimados utilizando cualquiera de los suavizadores bivariados.

El modelo es ajustado usando el algoritmo *local scoring*, el cual iterativamente ajusta modelos aditivos ponderados usando *backfitting*. El algoritmo *backfitting* es un método de Gauss-Seidel para ajustar modelos aditivos usando residuales parciales de suavización iterativos.

Algoritmo *backfitting*

1. En el paso inicial se definen las funciones $f_j^{(0)} \equiv 1$
2. En la i-ésima iteración, se estima $f_j^{(i+1)}$ como:

$$f_j^{(i+1)} = s\left(y - \sum_{k \neq j} f_k^i(x_k) \right) \text{ para } j = 1, \ldots, p$$

3. Cotejar si $\left| f_j^{(i+1)} - f_j^i \right| < \delta$ para todo $j = 1, \ldots, p$, donde δ es una constante de tolerancia. Si no se cumple la condición, se vuelve al paso 2. En caso contrario, se para y se usa $f_j^{(i)}$ como f_j en el modelo aditivo.

ix. REGRESIÓN POR *PROJECTION PURSUIT*

Supóngase que se tiene una nube de puntos observados en un espacio de alta dimensión. Para muchos propósitos, entre los cuales está la exploración y presentación de datos, es útil reducir la dimensionalidad de los datos, proyectándola linealmente en una o dos dimensiones. Una técnica clásica para reducción de dimensión de este tipo es el Análisis de Componentes Principales (ACP), reducir las variables en una o dos

componentes principales corresponde precisamente a una proyección en el subespacio escogido para maximizar la variancia de los datos proyectados. Una descripción más detallada sobre el ACP la encontrará en Chatfield *y Collins* [4] y Mardia *et.al.* [36].

Una estrategia más general para la reducción de dimensiones es definir una estructura "interesante" y luego construir un índice de "interés", que es satisfactoriamente maximizado. Esta es la idea detrás de la Regresión por *Prejection Pursuit* (PPR), con índices diferentes se llega a diferentes escenarios. Entonces, lo primero que se debe hacer es definir qué es lo que se quiere decir por estructura "interesante" del conjunto de datos y luego seleccionar el subespacio en donde la proyección de los datos es más interesante. (Silverman, [58])

En la Regresión por *Projection Pursuit* se tiene un vector $\mathbf{X}$ con p componentes, y un *target* Y. Se deja que w_m, $m=1,2,\ldots,M$ sea el vector que contiene los parámetros desconocidos. Entonces se define el modelo de regresión por *Projection Pursuit* (PPR) como sigue:

$$f(x)=\sum_{m=1}^{M} g_m\left(w_m^T\mathbf{X}\right)$$

Este es un modelo aditivo, pero con la característica que tiene a $V_m=w_m^T\mathbf{X}$ en lugar de los *inputs* propios. Las funciones g_m no son específicas y son estimadas a través de las direcciones w_m usando algún método flexible de suavización.

La función $g_m\left(w_m^T\mathbf{X}\right)$ es llamada la función Ridge en $\mathbb{R}^p$. Varía sólo en la dirección definida por el vector w_m. La variable escalar $V_m=w_m^T\mathbf{X}$ es la proyección de X en el vector w_m, y se busca w_m para que el modelo ajuste bien, es por ello el nombre "*Projection Pursuit*".

La manera en que se ajusta un modelo PPR, es dándole datos de entrenamiento $\left(x_i,y_i\right)$; $i=1,2,\ldots,N$. Se busca la aproximación que minimice el error de la función:

$$\sum_{i=1}^{N}\left[y_i - \sum_{m=1}^{M} g_m\left(w_m^T \mathbf{X}\right)\right]^2$$

sobre funciones g_m y vectores de dirección w_m, $m = 1, 2, \ldots, M$.

x. REGRESIÓN POR ÁRBOLES CART

En el caso de la regresión por árboles CART, la superficie de regresión es estimada usando el siguiente modelo aditivo

$$s(\mathbf{x}) = \sum_{i=1}^{n} c_i I_{N_i}(\mathbf{x})$$

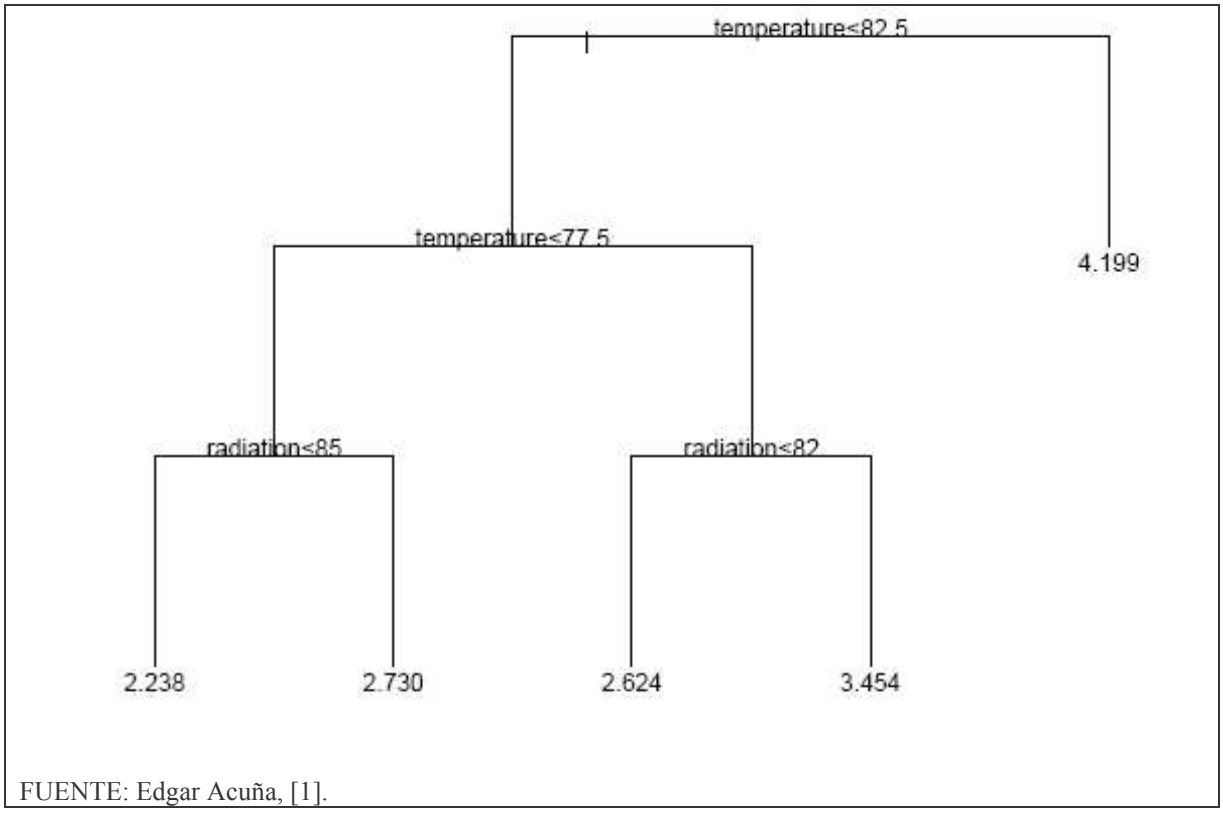
donde:

1. c_i son constantes
2. $I_{N_i}(\mathbf{x}) = 1$ si $\mathbf{x} \in N_i$, sino es igual a cero en otro caso.
3. Los N_i son hiperrectángulos disjuntos con lados paralelos a los ejes coordenados. Los hiperrectángulos son construidos por partición recursiva y pueden ser representados como un árbol.

En la Figura 2.8 se puede observar cómo se construye una regresión basándose en los datos, la cual se puede representar en forma de árbol. En este caso se ha podado[5] el árbol hasta un número final de 5 nodos [1], para su mayor comprensión. En la Figura 2.9 se representan las variables estudiadas en forma tridimensional.

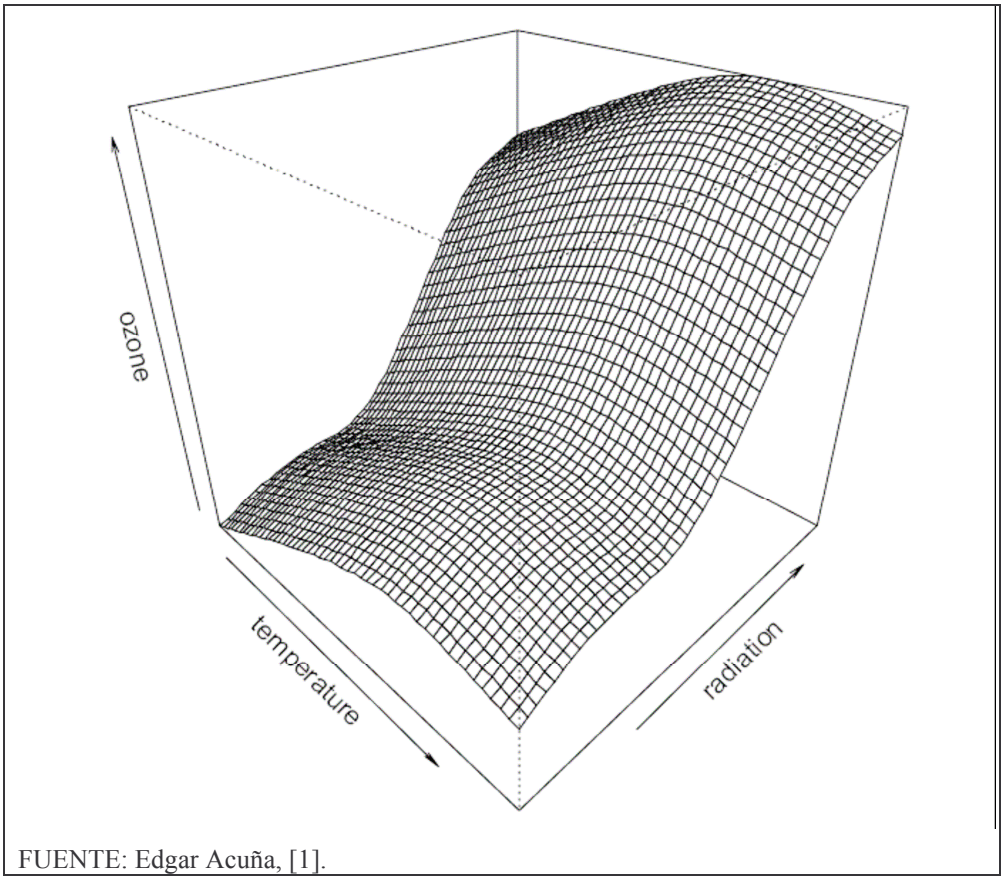
[5] Cuando se habla de podar un árbol de decisión se refiere al hecho de que se busca minimizar el sobreajuste del modelo. La poda consistente en restringir el numero de nodos que presente el árbol, eliminar ramas que no proporcionen información importante. Simplificar el árbol. Hay muchos métodos de poda, como por ejemplo el costo de complejidad.

Figura 2.8: Ejemplo de árbol con 5 nodos terminales



FUENTE: Edgar Acuña, [1].

Figura 2.9: Superficie de la suavización por árboles tridimensional



FUENTE: Edgar Acuña, [1].

III. MATERIALES Y MÉTODOS

3.1. MATERIALES Y EQUIPOS

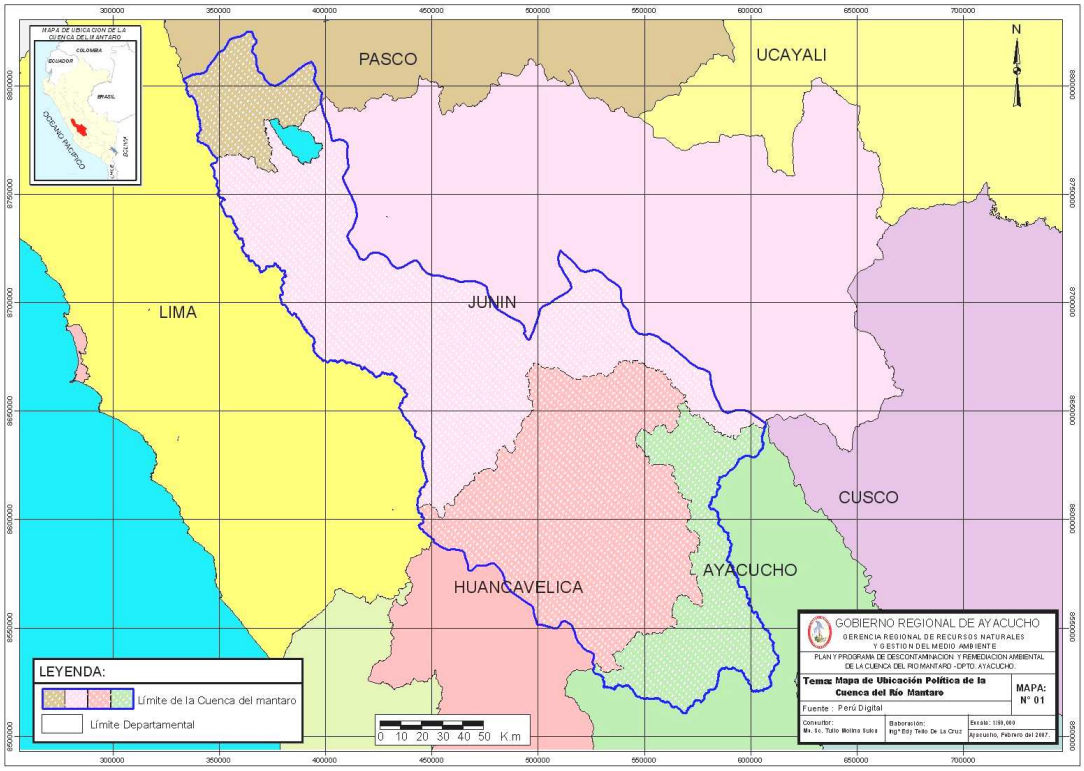
Los datos analizados provienen de las estaciones meteorológicas: Huayao, Jauja, Santa Ana y Viques e información secundaria sobre las variables explicativas obtenidas de instituciones especializadas a través de Internet, cabe resaltar que la cantidad de información es limitante para algunas zonas de la cuenca. Además, la información que se obtenga es solamente aplicable a la cuenca del río Mantaro y alrededores, no siendo utilizable para otras zonas del Perú.

DESCRIPCIÓN DE LA ZONA DE ESTUDIO

La Cuenca Hidrográfica del río Mantaro está ubicada en la sierra central del país (Figura 3.1), abarca 23 provincias y se encuentra circunscrita dentro del territorio de los departamentos: Pasco, Junín, Huancavelica y Ayacucho, constituyendo el principal tributario[6] para unirse con el río Apurímac y formar el río Ene. La cuenca se encuentra ubicada entre los paralelos 10º34'30'' y 13º35'30'' de latitud sur y entre los meridianos 73º55'00'' y 76º40'30'' de longitud oeste. Tiene un área total de 34 550,08 km^2. Un mapa más detallado sobre el relieve de la cuenca se encuentra en el Anexo 1.

[6] Adj. Se dice de un curso de agua con relación al río o mar adonde va a parar. Fuente: Diccionario de la lengua española de la Real Academia Española, vigésima segunda edición.

Figura 3.1. Ubicación Política y Geográfica de la Cuenca del Río Mantaro



FUENTE: Gobierno Regional de Ayacucho

INFORMACIÓN METEOROLÓGICA

A continuación se presenta información acerca de las estaciones de donde provino cada variable respuesta y el rango de tiempo de cada una, para el análisis de precipitación (pp) [ver Cuadro 3.1] y en el análisis de temperatura máxima (tx) y temperatura mínima (tm) [ver Cuadro 3.2].

Cuadro 3.1: Características de las estaciones utilizadas en el análisis de precipitación

Estaciones Meteorológicas	**Longitud (O)**	**Latitud (S)**	**Altura (m.s.n.m.)**	**Período**
Huayao	75.30°	12.00º	3350	1960 - 2008
Jauja	75.50°	11.75°	3410	1960 - 2008
Santa Ana	75.22°	12.00°	3295	1992 - 2004
Viques	75.23°	12.17°	3184	1963 - 2008

FUENTE: Elaboración propia

Cuadro 3.2: Características de las estaciones utilizadas en el análisis de temperatura				
Estaciones Meteorológicas	**Longitud (O)**	**Latitud (S)**	**Altura (m.s.n.m.)**	**Período**
Huayao	75.30°	12.00º	3350	1952 - 2008
Jauja	75.50°	11.75°	3410	1960 - 2008
Santa Ana	75.22°	12.00°	3295	1992 - 2002

FUENTE: Elaboración propia

VARIABLES GLOBALES

Se utilizan como variables explicativas un conjunto de índices globales, a continuación se brindará una descripción de cuales son las variables globales utilizadas: Índices globales de TSM en el Pacifico ecuatorial, conocidos como las regiones Niño 1+2, Niño 3, Niño 4 y Niño 3.4 que provienen del *Earth System Research Laboratory*[7] (ESRL); la presión atmosférica a nivel del mar en Darwin, Tahití y el Índice de Oscilación del Sur (SOI), provienen del *Climate Prediction Center*[8] (CPC) de la NOAA[9]; también se utilizaron otros índices de distintas zonas del mundo; un resumen de todas las variables explicativas se encuentra en el Anexo 2. La recolección de las variables explicativas se hizo a través de Internet, accediendo desde cada portal antes mencionado. Se utilizaron 17 variables explicativas correspondientes al mismo periodo de análisis. La ubicación geográfica de las regiones en donde se ubican los índices se presenta en el Anexo 3.

MATERIALES, EQUIPOS Y SOFTWARES

- R cran[10] versión 2.8.1
- SPSS 16 Trial Versión
- Weka 3.4[11]
- Computadora de escritorio (Intel Core 2, 2.13 GHz, 1.98 GB RAM), *laptop* (AMD Turion, 2 GB de RAM) y materiales de escritorio.

[7] Se pueden bajar desde l siguiente dirección: http://www.esrl.noaa.gov/
[8] http://www.cpc.ncep.noaa.gov/
[9] NOAA- National Oceanic and Atmospheric Administration
[10] http://cran.r-project.org/doc/FAQ/R-FAQ.html
[11] http://www.cs.waikato.ac.nz/ml/weka/

R cran

R es un software para análisis estadísticos y gráficos creado por Ross Ihaka y Robert Gentleman. R tiene una naturaleza doble de programa y lenguaje de programación, Es considerado como un dialecto del lenguaje S creado por los Laboratorios *AT&T Bell*. S está disponible como el programa S-PLUS comercializado por *Insightful*. Existen diferencias importantes en el diseño de R y S: aquellos interesados en averiguar más sobre este tema pueden leer el artículo publicado por Ihaka & Gentleman [30] o las Preguntas Más Frecuentes en R, que también se distribuyen con el software.

R se distribuye gratuitamente bajo los términos de la GNU *General Public Licence*; su desarrollo y distribución son llevados a cabo por varios estadísticos conocidos como el Grupo Nuclear de Desarrollo de R.

R en el contexto de esta investigación se ha utilizado para el análisis del conjunto de datos mediante el paquete mda[12] para el desarrollo de los modelos MARS, este software, además, se utiliza para el análisis discriminante mixto y flexible, entre otros modelos. Los autores de este software son Trevor Hastie and Robert Tibshirani.

WEKA

Weka es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Es un software que ha sido desarrollado en la universidad de Waikato (Nueva Zelanda) bajo licencia GPL lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años.

Weka contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, *clustering*, asociación y visualización. Weka está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla.

[12] http://cran.r-project.org/web/packages/mda/index.html

Sin embargo, y pese a todas las cualidades que Weka posee, tiene un gran defecto y éste es la escasa documentación orientada al usuario que tiene junto a una usabilidad bastante pobre, lo que la hace una herramienta difícil de comprender y manejar sin información adicional. Se recomienda revisar manuales para el mejor entendimiento[13].

SPSS

Statistical Package for the Social Sciences (SPSS) es un programa estadístico informático muy usado en las ciencias sociales y las empresas de investigación de mercado. Originalmente SPSS fue creado como el acrónimo de *Statistical Package for the Social Sciences*. En la actualidad, la sigla se usa tanto para designar el programa estadístico como la empresa que lo produce.

Fue creado en 1968 por Norman H. Nie, C. Hadlai (Tex) Hull y Dale H. Bent. Entre 1969 y 1975 la Universidad de Chicago por medio de su *National Opinión Research Center* estuvo a cargo del desarrollo, distribución y venta del programa. A partir de 1975 corresponde a SPSS Inc.

SPSS tiene un sistema de ficheros en el cual el principal son los archivos de datos (extensión. SAV). Además de este tipo existen otros dos tipos de uso frecuente:

- Archivos de salida (*output*, extensión. SPO): en estos se despliega toda la información de manipulación de los datos que realizan los usuarios mediante las ventanas de comandos. Son susceptibles de ser exportados con varios formatos (originalmente HTML, RTF o TXT, actualmente la versión 15 incorpora la exportación a PDF junto a los formatos XLS y DOC que ya se encontraban en la versión 12)
- Archivos de sintaxis (extensión. SPS): Casi todas las ventanas de SPSS cuentan con un botón que permite hacer el pegado del proceso que el usuario desea realizar. Lo anterior genera un archivo de sintaxis donde se van guardando todas las

[13] Manual sugerido en http://metaemotion.com/diego.garcia.morate/download/weka.pdf

instrucciones que llevan a cabo los comandos del SPSS. Este archivo es susceptible de ser modificado por el usuario. Muchos de los primeros usuarios del SPSS suelen escribir estos archivos en vez de utilizar el sistema de pegado del programa.

Existe un tercer tipo de fichero: el fichero de *scripts* (extensión. SBS). Este fichero es utilizado por los usuarios más avanzados del software para generar rutinas que permiten automatizar procesos muy largos y/o complejos. Muchos de estos procesos suelen no ser parte de las salidas estándar de los comandos del SPSS, aunque parten de estas salidas. Buena parte de la funcionalidad de los archivos de scripts ha sido ahora asumida por la inserción del lenguaje de programación *Python* en las rutinas de *sintax* del SPSS.

Este programa se utilizó para el desarrollo de las estadísticas descriptivas, tablas y gráficos de esta investigación.

3.2. MÉTODOS

El período de análisis para todos los métodos es desde que empieza la serie hasta diciembre de 2000. El período de enero de 2001 hasta agosto de 2008 se utilizó como datos para la validación de los modelos MARS y RNAB.

3.2.1. ANALISIS EXPLORATORIO DE DATOS (AED)

Las técnicas utilizadas en el AED fueron: (a) histogramas y (b) diagrama de cajas *(Boxplot)*. Estas técnicas se utilizarán debido a que los datos en algunos casos pueden tener errores de medición, ya que las observaciones son tomadas con instrumentos que al ser observados por el ser humano se corre el riesgo de cometer algún tipo de error.

Para la detección de valores atípicos se utilizó la diferencia entre el primer cuartil Q_1 y el tercer cuartil Q_3, es decir, el rango intercuartílico (IQR). En un diagrama de caja se consideran dos tipos de valores atípicos: (a) valor atípico leve y (b) valor atípico extremo. A continuación se detalla el cálculo de estos valores y para que son calculados.

Valor atípico leve

Siendo Q_1 y Q_3 el primer y tercer cuartil, y IQR el rango intercuartílico ($Q_3 - Q_1$), un valor atípico leve será aquel que sea:

$$\text{LI:} < Q_1 - 1.5 \times IQR, \qquad \text{ó} \qquad \text{LS:} > Q_3 + 1.5 \times IQR.$$

Los resultados determinan, pues, los llamados limites interiores $\langle LI_i, LS_i \rangle$, a partir de los cuales la observación se considera un atípico leve.

Valor atípico extremo

Los atípicos extremos son observaciones más allá de los siguientes límites externos $\langle LI_e, LS_e \rangle$:

$$\text{LI:} < Q_1 - 3 \times IQR, \qquad \text{ó} \qquad \text{LS:} > Q_3 + 3 \times IQR.$$

Las series analizadas resultaron no presentar valores atípicos extremos. Los valores atípicos leves no son removidos, debido a que las variaciones de las variables respuesta, en este caso la temperatura y la precipitación tienden a tener algún punto fuera de rango asociado a algún fenómeno físico ocurrido en el planeta. Y considerando que algunas variables climáticas ciertas veces presentan grandes fluctuaciones, se ha visto la necesidad de dejar sólo los valores atípicos leves mas no los valores atípicos extremos.

3.2.2. REGRESIÓN MULTIVARIADA ADAPTATIVA *SPLINE*. MARS

La Regresión Multivariada Adaptativa *Spline* (*Multivariate Adaptive Regression Spline*, MARS) fue desarrollada por Friedman [17] en 1991, con el fin de ajustar modelos con relaciones aditivas y posibles interacciones. Esta técnica está inspirada en la regresión con particionamiento recursivo (Friedman, [15]).

MARS es un método de regresión no paramétrica que no hace ninguna suposición sobre la relación funcional entre las variables respuesta e independientes así como los

supuestos que requiere la Regresión Lineal Múltiple. Por esto, MARS construye esta relación basándose en una serie de coeficientes asociadas a las funciones base que son totalmente determinadas a partir de la regresión de los datos. Puede pensarse que el algoritmo para desarrollar el MARS opera como múltiples pedazos de regresiones lineales.

EL MODELO MARS

El modelo MARS se construye utilizando las llamadas funciones base, y junto con los parámetros del modelo (estimados a través de mínimos cuadrados) se combinan para producir las predicciones dadas por los *inputs*. Este modelo es visto como una función de las variables explicativas X y sus posibles interacciones. La ecuación general del modelo MARS (Friedman, [17]) se expresa así:

$$\hat{f}(\mathrm{x}) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_{km} \right) \right]_+ \tag{3.1}$$

donde la sumatoria es sobre las M funciones base definidas en el modelo.

Luego:

1. a_0 es el intercepto de la función
2. a_m son los coeficientes de las funciones base, que ponderan a la productoria de las funciones base
3. $\left[s_{km} \cdot \left(x_{v(k,m)} - t_{km} \right) \right]_+$ es una función base, siendo $s_{km} = \pm 1$.
4. $\prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_{km} \right) \right]_+$, es el producto de las Km funciones base.
5. Siendo $x_{v(k,m)}$ los valores de las v variables explicativas en el k-ésimo nodo de la m-ésima función base.

Se puede pensar en este modelo como "la selección de" una suma ponderada de funciones base del conjunto de funciones de base que abarcan todos los valores de cada predicción.

El algoritmo MARS hará las búsquedas sobre el espacio de todos los valores de las variables explicativas y variables respuesta, así como las interacciones entre las variables explicativas.

Funciones base

Las funciones base dentro del marco del modelo MARS son producidas por el Algoritmo MARS 1, las regiones correspondientes no son disjuntas sino que se sobreponen. Es por eso que al remover una función base no se produce un vacío en el espacio de los predictores.

Inicialmente (línea 1) el modelo esta compuesto por todo el conjunto de funciones base J^* proveniente del Algoritmo MARS 1. Cada iteración del bucle externo *For* del Algoritmo MARS 2 produce la eliminación de una función base. El bucle interior elige cual función base se elimina. Esta es la única que se remueve y su salida mejora o degrada el ajuste del modelo. Nótese que la función base constante $B_1(\mathrm{x}) = 1$ nunca es seleccionada para ser removida. El Algoritmo MARS 2 construye una secuencia de $M_{\max} - 1$ modelos (donde M es la cantidad de términos no constantes del modelo), cada uno tiene una función base menos que el modelo previo en la secuencia. El mejor modelo en esta secuencia se devuelve al momento de supresión.

Durante esta investigación, un número cada vez mayor de funciones base se añaden al modelo (Algoritmo MARS 1), para aprovechar al máximo criterio de bondad de ajuste. Como resultado de estas operaciones, MARS determina automáticamente las variables explicativas más importantes (Algoritmo MARS 2), así como las más importantes interacciones entre ellos.

Algoritmo MARS 1

$B_1(\mathrm{x}) \leftarrow 1; M \leftarrow 2$

Loop until $M > M_{\max}$: $lof^* \leftarrow \infty$

For $m = 1$ to $M - 1$ do:

For $v \notin \{v(k,m) \mid 1 \le k \le K_m\}$

For $t \in \{B_m(x_j) > 0\}$

$g \leftarrow \sum_{i=1}^{M-1} a_i B_i(\mathrm{x}) + a_M B_m(\mathrm{x})\left[+(x_v - t)\right]_+ + a_{M+1} B_m(\mathrm{x})\left[-(x_v - t)\right]_+$

$lof \leftarrow \min_{a_1,\ldots,a_{M+1}} LOF(g)$

if $lof < lof^*$ then $lof^* \leftarrow lof$; $m^* \leftarrow m$; $v^* \leftarrow v$; $t^* \leftarrow t$ end if

end for

end for

end for

$B_M(\mathrm{x}) \leftarrow B_{m^*}(\mathrm{x})\left[+(x_{v^*} - t^*)\right]+$

$B_{M+1}(\mathrm{x}) \leftarrow B_{m^*}(\mathrm{x})\left[-(x_{v^*} - t^*)\right]+$

$M \leftarrow M + 2$

end loop

end algorithm

Fuente: Friedman, J. [17]

Algoritmo MARS 2

$J^* = \{1, 2, \ldots, M_{\max}\}$; $K^* \leftarrow J^*$

$lof^* \leftarrow \min_{\{a_j \mid j \in J^*\}} LOF\left(\sum_{j \in J^*} a_j B_j(\mathrm{x})\right)$

For $M = M_{\max}$ to 2 do: $b \leftarrow \infty$; $L \leftarrow K^*$

For $m = 2$ to M do: $K \leftarrow L - \{m\}$

$lof \leftarrow \min_{\{a_k \mid k \in K\}} LOF\left(\sum_{k \in K} a_k B_k(\mathrm{x})\right)$

if $lof < b$ then $b \leftarrow lof$; $K^* \leftarrow K$ end if

if $lof < lof^*$ then $lof^* \leftarrow lof$; $J^* \leftarrow K$ end if

end for

end for

end algorithm

Fuente: Friedman, J. [17]

VARIABLES RESPUESTA

El algoritmo MARS se puede aplicar a múltiples variables respuesta. Este algoritmo determina un conjunto de funciones base a partir de las variables explicativas, pero los coeficientes se calculan de manera diferente para cada una de las variables respuesta. Este método trabaja con múltiples variables respuesta y no se diferencia de las arquitecturas de las redes neuronales, donde hay múltiples variables respuesta. En el caso de MARS a múltiples variables respuesta se prevé comunes funciones base, con diferentes coeficientes.

MODELO DE SELECCIÓN Y PODA

En general, los modelos no paramétricos son adaptables y pueden exhibir un alto grado de flexibilidad que en última instancia podría resultar sobrestimado si no se toman medidas para contrarrestarlo. Aunque tales modelos pueden lograr un error igual a cero en los datos ajustados, tienen la tendencia a predecir mal si se presentan nuevas observaciones. MARS, al igual que la mayoría de los métodos de este tipo, tienden a sobrestimar a los pronósticos. Para combatir este problema, MARS utiliza la técnica de la poda para limitar la complejidad del modelo, reduciendo el número de sus funciones base.

Como en Friedman y Silverman [18] y Friedman [16] se usa una forma modificada del criterio original de validación cruzada generalizada propuesto por Craven y Wahba [6] El criterio del GCV es el promedio cuadrado residual del ajuste de los datos (numerador) y una penalidad (denominador inverso) que cuenta para el incremento de la variancia asociada con el incremento de la complejidad del modelo (número de funciones base M), se expresa:

$$LOF\left(\hat{f}_M\right) = GCV\left(M\right) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}_M\left(\mathrm{x}_i\right)\right]^2 \Bigg/ \left[1 - \frac{C\left(M\right)}{N}\right]^2 . \qquad (3.2)$$

Aquí la dependencia de $\hat{f}$ (3.1), y el criterio, sobre el número de las M funciones base es indicada explícitamente.

Si los valores de los parámetros de las funciones base (número de factores K_m, variables $v(k,m)$, ubicación de los nodos t_{km} y el signo s_{km}) asociados con el modelo MARS fueron determinados independientemente de los valores de la variable respuesta ($y_1,\ldots,y_N$), entonces sólo los coeficientes ($a_o,\ldots,a_M$) son ajustados a los datos.

Consecuentemente la función costo de complejidad en relación con el número de parámetros ajustados es:

$$C(M) = \text{traza}\left(\mathrm{B}\left(\mathrm{B}^T\,\mathrm{B}\right)^{-1}\mathrm{B}^T\right)+1 \tag{3.3}$$

donde B es la matriz de datos de dimensión $M \times N$, de las M funciones base (no constantes) $\left(B_{ij} = B_i\left(\mathrm{x}_j\right)\right)$. Esto es igual al número de funciones base linealmente independientes en (3.1) y, por tanto, $C(M)$ en (3.3) es justo el número de parámetros que serán ajustados. Usando (3.3) en (3.2) lleva al criterio de GCV propuesto por Craven y Wahba [6].

El procedimiento MARS hace uso intensivo de los valores de la variables respuesta para construir el conjunto de funciones base. Es así como logra su poder y flexibilidad. Esto reduce (usualmente en forma dramática) la precisión del modelo estimado, pero al mismo tiempo incrementa la variancia desde que parámetros adicionales son estimados para ayudar al mejor ajuste de los datos. La disminución de la precisión es directamente reflejada en la reducción del promedio cuadrado residual [numerador (3.2)]. El denominador (3.2) (3.3), sin embargo, ya no refleja la variancia debida al número adicional de parámetros así como su naturaleza no lineal.

3.2.3. REDES NEURONALES ARTIFICIALES *BACKPROPAGATION* (RNAB)

El propósito principal de las redes neuronales artificiales es resolver problemas que no admiten un tratamiento algorítmico. Una red neuronal es un modelo artificial y simplificado del cerebro humano, que es capaz de adquirir conocimiento a través de la

experiencia. Una red neuronal es "un nuevo sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona".

Las redes neuronales artificiales se comenzaron a estudiar en 1936, por Alan Turing que estudiaba al cerebro como una forma de ver la computación, pero los primeros que concibieron los fundamentos de la computación neuronal fueron Warren McCulloch, neurofisiólogo, y Walter Pitts, matemático, quienes en 1943 lanzaron la teoría de la forma de trabajar de las neuronas (McCulloch, [38]).

Las redes neuronales se basan en una estructura de neuronas unidas por enlaces que transmiten información a otras neuronas, las cuales entregan un resultado mediante funciones matemáticas. Las redes neuronales aprenden de la información histórica a través de un entrenamiento, proceso mediante el cual se ajustan los parámetros de la red, a fin de entregar la respuesta deseada, adquiriendo entonces la capacidad de predecir respuestas del mismo fenómeno. El comportamiento de las redes depende entonces de los pesos para los enlaces, de las funciones de activación que se especifican para las neuronas.

DEFINICIÓN DE UNA RED NEURONAL ARTIFICIAL

A lo largo de los años han surgido muchas formas de definir a una red neuronal, se mencionarán algunos ejemplos (ver Hilera y Martínez, [25]):

> *"...un sistema de computación hecho por un gran número de elementos simples, elementos de proceso muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas".* (Hecht-Niesen, [24])

> *"Redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico".* (Kohonen, [32])

Una red neuronal es una estructura de procesamiento de información paralela y distribuida, que intenta emular las funciones computacionales elementales de la red nerviosa del cerebro humano, basándose en la interconexión de multitud de elementos de procesamiento, cada uno de los cuales presenta un comportamiento completamente local. (Introducción a las Redes Neuronales.- ISA-UMH © T-98-012V1.0)

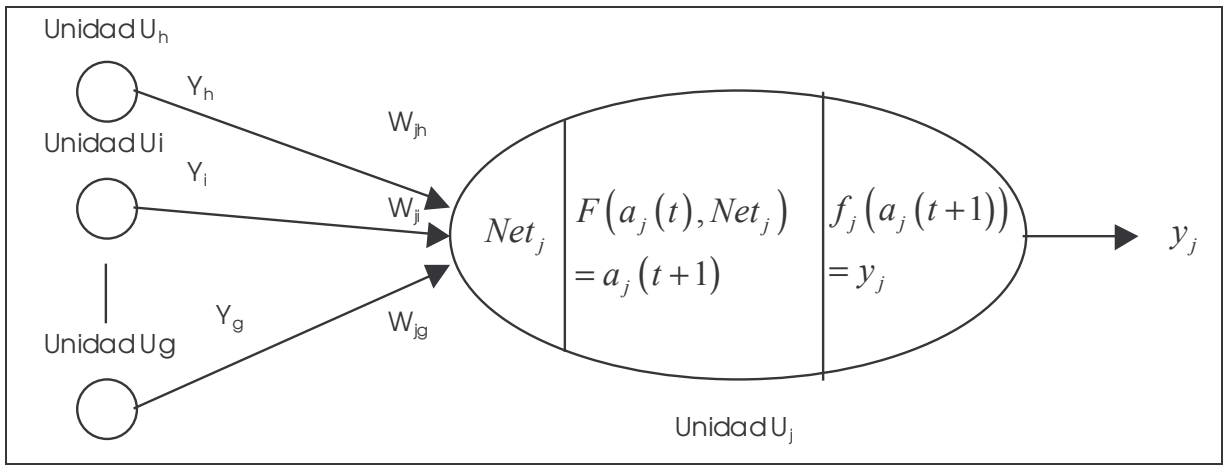
FUNDAMENTOS DE UNA RED NEURONAL ARTIFICIAL

a. Unidades de proceso: La neurona artificial.

Si se tienen N unidades (neuronas), se pueden ordenar arbitrariamente y designar la j-ésima unidad U_j (Ver Figura 3.2). Su trabajo es simple y único, y consiste en recibir las entradas de las células vecinas y calcular el valor de salida, el cual es enviado a todas las células restantes.

En todo sistema, se caracterizan tres tipos de unidades: las entradas, las salidas y las ocultas. Las unidades de entrada reciben señales desde el entorno del sistema. Las unidades de salida envían la señal fuera del sistema. Las unidades ocultas son aquellas cuyas entradas y salidas se encuentran dentro del sistema.

Figura 3.2: Entradas y salidas de una neurona U_j



FUENTE: Hilera et.al [25]

b. Estados de activación

Además del conjunto de unidades de proceso, la representación necesita los estados del sistema en un tiempo t. Esto se especifica por un vector de N números reales $A(t)$, que representa el estado de activación del conjunto de unidades de procesamiento. Cada elemento del vector representa la activación de la unidad en el tiempo t. La activación de una unidad U_i en el tiempo t se designa por $a_i(t)$:

$$A(t) = (a_1(t), a_2(t), \ldots, a_i(t), \ldots, a_N(t))$$

Todas las neuronas que componen la red se hallan en cierto estado. Se puede decir que hay dos posibles estados, reposo y excitado, a los que se denominan generalmente estados de activación, y a cada uno de los cuales se le asigna un valor. Los valores de activación pueden ser continuos o discretos. Además, pueden ser limitados o ilimitados.

c. Función de salida o transferencia

Entre las unidades o neuronas que forman la red neuronal artificial existe un conjunto de conexiones que unen unas con otras. Cada unidad transmite señales a aquellas que están conectadas con su salida. Asociada con cada unidad U_i hay una función de salida $f_i(a_i(t))$, que transforma el estado actual de activación $a_i(t)$ en una señal de salida $y_i(t)$ de la forma:

$$y_i(t) = f_i(a_i(t))$$

El vector que contiene las salidas de todas las neuronas en un instante t es

$$Y(t) = (f_1(a_1(t)), f_2(a_2(t)), \ldots, f_i(a_i(t)), \ldots, f_N(a_N(t)))$$

Existen cuatro funciones de transferencia típicas que determinan distintos tipos de neuronas: (a) Función escalón, (b) Función lineal y mixta, (c) Función sigmoidal y (d) Función gaussiana.

d. Conexiones entre neuronas

Las conexiones que unen a las neuronas que forman una RNA tiene asociado un peso, que es el que hace que la red adquiera conocimiento. Consideremos y_i como el valor de salida de una neurona i en un instante dado. Una neurona recibe un conjunto de señales que le dan información del estado de activación de todas las neuronas con las que se encuentra conectada. Cada conexión (sinapsis) entre la neurona i y la neurona j está ponderada por un peso w_{ji}. Normalmente, se considera que el efecto de cada señal es aditivo, de tal forma que la entrada neta que recibe una neurona net_j es la suma del producto de cada señal individual por el valor de la sinapsis que conecta ambas neuronas.

$$net_j = \sum_{i}^{N} w_{ji} \cdot y_i$$

Esta regla muestra el procedimiento a seguir para combinar los valores de entrada a una unidad con los pesos de las conexiones que llegan a esa unidad y es conocida como regla de propagación.

Suele utilizarse una matriz W con todos los pesos w_{ji} que reflejan la influencia que sobre la neurona j tiene la neurona i. W es un conjunto de elementos positivos, negativos o nulos. Si w_{ji} es positivo, indica que la interacción entre las neuronas i y j es excitadora; es decir, siempre que la neurona i este activada, la neurona j recibirá una señal de i que tenderá a activarla. Si w_{ji} es negativo, la sinapsis será inhibidora. En este caso, si i esta activada, enviará una señal a j que tenderá a desactivar a esta. Finalmente, si $w_{ji} = 0$, se supone que no hay conexión entre ambas.

e. Función o regla de activación

Asimismo hay la necesidad de tener una regla que combine las entradas con el estado actual de la neurona para producir un nuevo estado de activación. Esta función F produce un nuevo estado de activación en una neurona a partir del estado (a_i) que existía y la combinación de las entradas con los pesos de las conexiones (net_i).

Dado el estado de activación $a_i(t)$ de la unidad U_i y la entrada total que llega a ella, Net_i, el estado de activación siguiente, $a_i(t+1)$, se obtiene aplicando una función F, llamada función de activación.

$$a_i(t+1) = F(a_i(t), Net_i)$$

En la mayoría de los casos, F es la *función identidad*, por lo que el estado de activación de una neurona $t+1$ coincidirá con el Net de la misma en t. En este caso, el parámetro que se le pasa a la función de salida, f, de la neurona será directamente el Net. El estado de activación anterior no se tiene en cuenta. Según esto, la salida de una neurona i (y_i) quedará así:

$$y_i(t+1) = f(Net_i) = f\left(\sum_{j=1}^{N} w_{ij} y_j(t)\right)$$

Por tanto, y en lo sucesivo, se considera únicamente la función f, que se denominará indistintamente de transferencia o de activación. Además, normalmente la función de activación no está centrada en el origen del eje que representa el valor de la entrada neta, sino que existe cierto desplazamiento debido a las características internas de la propia neurona y que no es igual en todas ellas. Este valor se denota como θ_i y representa el umbral de activación de la neurona i.

$$y_i(t+1) = f(Net_i - \theta_i) = f\left(\sum_{j=1}^{N} w_{ij} y_j(t) - \theta_i\right)$$

f. Regla de aprendizaje

En el caso de las RNA, se puede considerar que el conocimiento se encuentra representado en los pesos de las conexiones entre neuronas. Todo proceso de aprendizaje implica cierto número de cambios en estas conexiones. En realidad, puede decirse que se aprende modificando los valores de los pesos de la red.

LA RED NEURONAL *BACKPROPAGATION*

En 1986, Rumelhart, Hinton y Williams [50], basándose en los trabajos de otros investigadores [Werbos (1974) y Parker (1982)] formalizaron un método para que una red neuronal aprendiera la asociación que existe entre los patrones de entrada a la misma y las clases correspondientes, utilizando más niveles de neuronas que los que utilizó Rosenblatt para desarrollar el Perceptron[14].

El funcionamiento de una red *backpropagation* consiste en un aprendizaje de un conjunto predefinido de pares de entradas-salidas dados como ejemplo, empleando un ciclo propagación-adaptación de dos fases: primero se aplica un patrón de entrada como estímulo para la primera capa de neuronas de la red, se va propagando a través de todas las capas superiores hasta generar una salida, se compara el resultado obtenido en las neuronas de salida con la salida. A continuación, estos errores se transmiten hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la intermedia que contribuyan directamente a la salida, recibiendo el porcentaje de error aproximado a la participación de la neurona intermedia en la salida original. Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido un error que describa su aportación relativa al error total. Basándose en el valor del error recibido, se reajustan los pesos de conexión de cada neurona, de manera que en la siguiente vez que se presente el mismo patrón, la salida este más cercana a la deseada; es decir, el error disminuya.

[14] Perceptron: Tipo de red neuronal artificial, con conexiones hacia delante. La más antigua de las redes neuronales. Revisar: Hilera y Martínez (2000). *Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones,* 103p-115p para mayor detalle.

La importancia de la red *backpropagation* consiste en su capacidad de autoadaptar los pesos de las neuronas de las capas intermedias para aprender la relación que existe entre un conjunto de patrones dados como ejemplo y sus salidas correspondientes. Para poder aplicar esa misma relación, después del entrenamiento, a nuevos vectores de entrada con ruido o incompletas, dando una salida activa si la nueva entrada es parecida a las presentadas durante el aprendizaje. Esta característica importante, que se exige a los sistemas de aprendizaje, es la capacidad de generalización, entendida como la facilidad de dar salidas satisfactorias a entradas que el sistema no ha visto nunca en su fase de entrenamiento. La red debe encontrar una representación interna que le permita generar las salidas deseadas cuando se le dan las entradas de entrenamiento, y que pueda aplicar, además, a entradas no presentadas durante la etapa de aprendizaje para clasificarlas según las características que comparten con los ejemplos de entrenamiento.

a. Aplicación del algoritmo *Backpropagation*

La aplicación del algoritmo *backpropagation* tiene dos fases, una hacia delante y otra hacia atrás. Durante la primera fase el patrón de entrada es presentado a la red y propagado a través de las capas hasta llegar a la capa de salida. Obtenidos los valores de salida de la red, se inicia la segunda fase, comparándose estos valores con la salida esperada para obtener el error. Se ajustan los pesos de la última capa proporcionalmente al error. Se pasa a la capa anterior con una retropropagación del error, ajustando convenientemente los pesos y continua este proceso hasta llegar a la primera capa. De esta manera se han modificado los pesos de las conexiones de la red para cada ejemplo o patrón de aprendizaje del problema, del que conocíamos su valor de entrada y la salida deseada que debería generar la red ante dicho patrón. A continuación se presentan, los pasos y fórmulas a utilizar para aplicar el algoritmo de entrenamiento:

Paso 1: Inicializar los pesos de la red con valores pequeños aleatorios.

Paso 2: Presentar un patrón de entrada X_p: $x_{p1}, x_{p2}, \ldots x_{pN}$, y especificar la salida deseada que debe generar la red: $d_1, d_2, \ldots, d_M$ (si la red se utiliza como un clasificador, todas las

salidas deseadas serán cero, salvo una, que será la de la clase a la que pertenece el patrón de entrada).

Paso 3: Calcular la salida actual de la red, para ello presentamos las entradas a la red y vamos calculando la salida que presenta cada capa hasta llegar a la capa de salida, esta será la salida de la red $y_1, y_2, \ldots, y_M$. Los pasos son los siguientes:

- Se calculan las entradas netas para las neuronas ocultas procedentes de las neuronas de entrada.

  Para una neurona j oculta:

$$net_{pj}^{h} = \sum_{i=1}^{N} w_{ji}^{h} x_{pi} + \theta_j^h$$

  en donde el índice h se refiere a magnitudes de la capa oculta; el subíndice p, al p-ésimo vector de entrenamiento, y j a la j-ésima neurona oculta. El término θ puede ser opcional, pues actúa como una entrada más.
- Se calculan las salidas de las neuronas ocultas:

$$y_{pj} = f_j^h \left(net_{pj}^h \right)$$

- Se realizan los mismos cálculos para obtener las salidas de las neuronas de salida (capa o : *output*)

$$net_k^o = \sum_{j=1}^{L} w_{kj}^o y_{pj} + \theta_k^o$$
$$y_{pk} = f_k^o \left(net_{pk}^o \right)$$

Paso 4 Calcula los términos de error para todas las neuronas

Si la neurona k es una neurona de la capa de salida, el valor de la delta es:

$$\delta_k^o = \left(d_{pk} - y_{pk} \right) f_k^o \left(net_{pk}^o \right)$$

La función f, como se citó anteriormente, debe cumplir el requisito de ser derivable, lo que implica la imposibilidad de utilizar una función escalón. En general, disponemos de dos formas de función de salida que nos pueden servir: la función lineal de salida ($f_k\left(net_{jk}\right) = net_{jk}$) y la función sigmoidal definida por la expresión:

$$f_k\left(net_{jk}\right) = \frac{1}{1+e^{-net_{jk}}}$$

La selección de la función de salida depende de la forma en que se decida representar los datos de salida: si se desea que las neuronas de salida sean binarias, se utiliza la función sigmoidal, puesto que esta función es casi biestable y, además, derivable. En otros casos es tan aplicable una función como otra.

Para la función lineal, tenemos: $f_k^{o'} = 1$, mientras que la derivada de una función f sigmoidal es:

$$f_k^{o'} = f_k^o\left(1-f_k^o\right) = y_{pk}\left(1-y_{pk}\right)\ f_k^{o'} = f_k^o\left(1-f_k^o\right) = y_{pk}\left(1-y_{pk}\right)$$

por lo que los términos de error para las neuronas de salida quedan:

$$\delta_{pk}^o = \left(d_{pk} - y_{pk}\right)$$

para la salida lineal, y

$$\delta_{pk}^o = \left(d_{pk} - y_{pk}\right) y_{pk}\left(1-y_{pk}\right)$$

para la salida sigmoidal.

Si la neurona j no es de salida, entonces la derivada parcial del error no puede ser evaluada directamente. Por tanto, se obtiene el desarrollo a partir de valores que son conocidos y otros que pueden ser evaluados.

La expresión obtenida en este caso es:

$$\delta_{pj}^h = f_j^h\left(net_{pj}^h\right)\sum_k \delta_{pk}^o w_{kj}^o$$

donde se observa que el error en las capas ocultas depende de todos los términos de error de la capa de salida. De aquí surge el término de propagación hacia atrás. En particular, para la función sigmoidal:

$$\delta_{pj}^{h} = x_{pi}\left(1 - x_{pi}\right)\sum_{k}\delta_{pk}^{o} w_{kj}^{o}$$

donde k se refiere a todas las neuronas de la capa superior a la de la neurona j. Así, el error que se produce en una neurona oculta es proporcional a la suma de los errores conocidos que se producen en las neuronas a las que está conectada la salida de ésta, multiplicado cada uno de ellos por el peso de la conexión. Los umbrales internos de las neuronas se adaptan de forma similar, considerando que están conectados con pesos desde entradas auxiliares de valor constante.

Paso 5 Actualización de los pesos

Para ello, se utiliza el algoritmo recursivo, comenzando por las neuronas de salida y trabajando hacia atrás hasta llegar a la capa de entrada, ajustando los pesos de la forma siguiente:

Para los pesos de las neuronas de la capa de salida:

$$w_{kj}^{o}\left(t+1\right) = w_{kj}^{o}\left(t\right) + \Delta w_{kj}^{o}\left(t+1\right);$$
$$\Delta w_{kj}^{o}\left(t+1\right) = \alpha\delta_{pk}^{o} y_{pj}$$

y para los pesos de las neuronas de la capa oculta:

$$w_{ji}^{h}\left(t+1\right) = w_{ji}^{h}\left(t\right) + \Delta w_{ji}^{h}\left(t+1\right);$$
$$\Delta w_{ji}^{h}\left(t+1\right) = \alpha\delta_{pj}^{h} x_{pi}$$

En ambos casos, para acelerar el proceso de aprendizaje, se puede añadir un término momento de valor: $\beta\left(w_{ji}^{h}\left(t\right) - w_{ji}^{h}\left(t-1\right)\right)$ cuando se trata de una neurona oculta.

Paso 6 El proceso se repite hasta que el término de error

$$E_p = \frac{1}{2}\sum_{k=1}^{M} \delta_{pk}^2$$

resulta aceptablemente pequeño para cada uno de los patrones aprendidos.

3.2.4. VALIDACIÓN DE LOS RESULTADOS

En la presente investigación se emplea la validación funcional para validar los pronósticos. En la validación se utiliza el error cuadrático medio (ECM), la raíz del error cuadrático medio (RECM), el error absoluto medio (EAM), el error absoluto medio normalizado (EAMN), el sesgo (BIAS) y la correlación (COR), definidos por Pielke [45] y Stauffer y Seaman [59]. Estos estadísticos permiten medir la precisión de los resultados, a medida que su valor sea menor, los resultados serán mejores; excepto en el caso de la correlación donde sucede todo lo contrario, cuando este valor es mayor los resultados son mucho mejores.

Por precisión se define como el grado de correspondencia entre pares individuales de valores pronóstico y valores observados. Los valores observados son aquellos datos obtenidos en las estaciones meteorológicas.

Para el cálculo de la precisión se utiliza el error cuadrático medio, que se define como:

$$ECM = \sum_{i=1}^{N} \frac{\left(x_i - x_{iobs}\right)^2}{N}$$

donde

1. x_i es el valor pronóstico para la fila i
2. x_{iobs} es el valor observado para la fila i
3. N es el número de valores analizados

También se suele utilizar la raíz cuadrada del ECM [45], definido como:

$$RECM = \sqrt{\sum_{i=1}^{N} \frac{\left(x_i - x_{iobs}\right)^2}{N}}$$

La RECM calcula la raíz cuadrada de la medida de las diferencias en promedio entre los valores pronóstico y los observados. Esta es una mejor opción, a utilizar en vez del ECM, que da como resultado unidades elevadas al cuadrado. Otro estadístico que describe una información similar es el error absoluto medio [59] se define como:

$$EAM = \sum_{i=1}^{N} \frac{\left|x_i - x_{iobs}\right|}{N}$$

Para tener en cuenta el peso del error respecto al valor de la variable medida, se normalizada el error absoluto, teniendo el error absoluto medio normalizado [59]:

$$EAMN = \sum_{i=1}^{N} \frac{\left|x_i - x_{iobs}\right| / x_{iobs}}{N}$$

El sesgo proporciona información sobre la tendencia del modelo a sobrestimar o subestimar una variable, cuantifica el error sistemático del modelo [45] define BIAS como:

$$BIAS = \sum_{i=1}^{N} \frac{\left(x_i - x_{iobs}\right)}{N}$$

Finalmente la correlación indica la fuerza y la dirección de una relación lineal entre dos variables aleatorias. Se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra: si tenemos dos variables (X e Y) existe correlación si al aumentar los valores de X lo hacen también los de Y, y viceversa. La correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad. En el contexto del análisis de validación, la correlación se va a utilizar como un indicador de que los valores pronósticos tienen el sentido y la fuerza de una relación lineal con los valores observados. Se espera que la correlación entre los valores pronóstico y los valores observados tengan un valor cercano a uno y que sea positiva.

La relación entre dos variables cuantitativas queda representada mediante la línea de mejor ajuste, trazada a partir de la nube de puntos. Los principales componentes elementales de una línea de ajuste y, por lo tanto, de una correlación, son:(1) la fuerza, (2) el sentido y (3) la forma:

1. La fuerza mide el grado en que la línea representa a la nube de puntos: si la nube es estrecha y alargada, se representa por una línea recta, lo que indica que la relación es fuerte; si la nube de puntos tiene una tendencia elíptica o circular, la relación es débil.

2. El sentido mide la variación de los valores de Y con respecto a X: si al crecer los valores de X lo hacen los de Y, la relación es positiva; si al crecer los valores de X disminuyen los de Y, la relación es negativa.

3. La forma establece el tipo de línea que define el mejor ajuste: la línea recta, la curva monotónica o la curva no monotónica.

La correlación se define como (Freedman, *et.al*., [14]):

$$COR = r_{xy} = \frac{Cov_{xy}}{S_x S_y}$$

donde:

$$Cov_{xy} = \sum_{i=1}^{N} \frac{(x_i - \overline{x})(y_i - \overline{y})}{N-1} \quad , \quad S_x = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{N-1}} \quad \text{y} \quad S_y = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \overline{y})^2}{N-1}}$$

LeBlanc [35], utiliza el estimado del error relativo de predicción basado en la muestra de prueba para comparar dos modelos aplicados a los datos:

$$ERP = \frac{\sum_{i=1}^{n_x} \left(y_{xi} - \hat{f}\left(x_{xi}\right) \right)^2}{\sum_{i=1}^{n_x} \left(y_{xi} - \overline{y}_x \right)^2},$$

donde y_{xi} y x_{xi}, son las observaciones de la muestra de prueba, n_x es el numero de las observaciones en la muestra de prueba y $\overline{y}_x$ es la media de las respuestas de la muestra de prueba.

IV. RESULTADOS Y DISCUSIÓN

Esta investigación se inicia con el análisis exploratorio de los datos de precipitación total mensual y temperaturas extremas (mínima y máxima) mensuales promedio para las cuatro estaciones meteorológicas que se estudian en la cuenca del río Mantaro con el fin de conocer su estructura, es decir, su comportamiento temporal, respecto a su media y variancia, además, presencia de valores que puedan considerarse como *outliers*.

Para el desarrollo de la técnica MARS se estratifica el conjunto inicial de todas las variables respuesta. Para el análisis se tendrá 12 subconjuntos por cada conjunto de variables explicativas. Cada subconjunto pertenece a un mes del año. Los valores atípicos extremos se eliminan del análisis pero se mantienen los valores atípicos leves, ya que en las variables climáticas se pueden encontrar estos valores que son verdaderos mas no errores.

En el caso del desarrollo de las RNAB se utilizará toda la serie de cada una de las variables respuesta, exceptuando valores atípicos extremos encontrados en el A.E.D.

Las variables explicativas usadas en el presente estudio tienen un desfase de tres meses, para tener pronósticos a tres meses posteriores. En un estudio anterior de la autora se comprobó que cuando se toman las variables explicativas de uno hasta seis meses de desfase, los mejores resultados (ECM) pertenecen a aquellos estudios con variables explicativas con tres o cuatro meses de anticipación. (Latínez, [33]).

4.1. ANÁLISIS EXPLORATORIO DE LOS DATOS (AED).

4.1.1. ESTACIÓN DE HUAYAO

Se observó que la distribución de la precipitación en Huayao tiene una asimetría hacía la derecha, es decir, la media es mayor que la mediana (ver Cuadro 4.1). Esta asimetría es provocada por algunos valores muy grandes, además, la mayoría de los valores se ubica en la parte inferior de la distribución, es decir, en la zona izquierda. Al mismo tiempo, se observó una desviación estándar igual a 51.0 mm. El histograma y el diagrama de cajas se presentan en los Anexos 4 y 5, respectivamente. El resumen descriptivo mensual se encuentra en el Anexo 6. En el diagrama de cajas por meses (Anexo 7) se distingue un alza en las medianas de los meses de enero, febrero, marzo y diciembre con respecto a los demás, y una disminución considerable en los meses de junio y julio, esto se debe al cambio de estación que conlleva a variaciones estacionales de precipitación.

Asimismo, la distribución de la temperatura mínima tiene una asimetría hacia la izquierda, lo que indica más concentración de datos en la parte superior de la distribución. Entonces, la media es menor que la mediana (ver Cuadro 4.1) esta distorsión es provocada por valores muy pequeños. Asimismo, se observó una desviación estándar de 2.5 °C. El histograma, el diagrama de cajas y el resumen descriptivo por meses se encuentran en los Anexos 8, 9 y 10, respectivamente. Se observó que las medianas (Anexo 11) en los meses de mayo hasta agosto disminuyen en comparación a los demás meses. Nuevamente, estas diferencias en temperatura corresponden a las variaciones estacionales de esta variable.

Luego, la distribución de la temperatura máxima es simétrica, quiere decir que hay igual cantidad de datos a la derecha que a la izquierda de la media (19.4 °C), y que la media y la mediana son iguales o muy cercanas (ver Cuadro 4.1). También se observó que la desviación estándar es 1.2 °C, entonces, la temperatura máxima tiene menos variabilidad que la temperatura mínima. El histograma, el diagrama de cajas y el resumen descriptivo por meses se encuentran en los Anexos 12, 13 y 14, respectivamente. En el diagrama de cajas por meses (Anexo 15) se observó que no hay marcadas diferencias en cuanto a la media pero sí un aumento de temperatura leve a partir de agosto.

A continuación se presentan las estadísticas descriptivas de la precipitación, la temperatura mínima y la temperatura máxima.

Cuadro 4.1: Estadísticas descriptivas para la estación de Huayao

	Mínimo	Máximo	Mediana	Media	Error Estándar de la Media	Desviación Estándar	Asimetría
Precipitación (mm)	0.00	239.70	54.80	62.402	2.3004	51.0244	0.773
Temp. Mínima (°C)	-2.40	8.70	5.10	4.396	0.1039	2.5196	-0.661
Temp. Máxima (°C)	15.80	22.70	19.40	19.368	0.0495	1.1996	0.009

FUENTE: Elaboración propia

Se analizó la precipitación por meses con un diagrama de cajas (Anexo 7), se encontró varios valores fuera de los bigotes. Se eliminaron aquellas observaciones que se consideran valores atípicos extremos. En junio se eliminó una observación que se encuentra fuera del intervalo [0.0 mm, 40.9 mm], que pertenece al año 1990 con un valor de 62.9 mm. Asimismo, en agosto la observación del año 1971 igual a 79.1 mm, se eliminó ya que no se encontraba en el intervalo [0.0 mm, 70.2 mm].

Luego, se analizó temperatura mínima por meses a los datos (Anexo 11). Se observó que existen valores muy diferentes de las medianas de temperatura entre los meses de febrero y julio con valores iguales a 6.8 °C y 0.3 °C, respectivamente. Sin embargo, no se encontraron valores atípicos extremos, debido a esto no se elimina ninguna observación a pesar de que hay 9 valores atípicos leves.

Finalmente, se analizó la temperatura máxima y se encontró que existen diferencias en las medianas de cada mes (Anexo 15), en este caso las diferencias no son grandes como es el caso de la precipitación y la temperatura mínima. Pero, sí se observa una tendencia positiva, es decir, aumento de la temperatura máxima desde enero a diciembre.

4.1.2. ESTACIÓN DE JAUJA

Primero, se analizó la precipitación encontrándose varios valores ausentes. El resumen de estos casos se encuentra en el Anexo 16, no se hizo imputación de datos. También se observó que la distribución es asimétrica hacia la derecha, es decir, que la

media es mayor que la mediana (ver Cuadro 4.2) y existen valores muy grandes que producen esta distorsión. La dispersión de los datos, se obtiene de la desviación estándar calculada que es 54.4 mm. El histograma, el diagrama de cajas y el resumen descriptivo por meses se encuentran en los Anexos 17, 18 y 19, respectivamente. Las gráficas del diagrama de cajas por meses (Anexo 20) muestran que en los primeros tres meses del año hay mayor precipitación que en el resto del año, siendo los meses con menos e inclusive ninguna precipitación junio y julio.

En los datos de temperatura mínima se presentaron meses en los que hubo valores ausentes (ver Anexo 21), no se imputaron dicho valores por lo que no se tomaron en cuenta para el análisis. Asimismo, la distribución de la temperatura mínima es asimétrica a la izquierda, es decir, que la media es menor que la mediana por la presencia de valores muy pequeños. Los datos presentan una desviación estándar igual a 2.4 °C. El histograma, el diagrama de cajas y el resumen descriptivo por meses se encuentran en los anexos 22, 23 y 24, respectivamente. En el diagrama de cajas por meses (Anexo 25) se observó que en los meses de mayo hasta agosto hay una disminución de temperatura con respecto a los demás meses de por lo menos 2 °C. Esta variabilidad se debe a las variaciones estacionales propias de la zona.

Asimismo, en la temperatura máxima hubo presencia de valores ausentes (ver Anexo 26) los cuales no se imputaron ni se tomaron en cuenta para el análisis. Luego se encontró que la distribución es simétrica, indicando que hay similar cantidad de datos a la derecha e izquierda de la media (19.3 °C). La desviación estándar es 1.2 °C, entonces, la temperatura máxima tiene menor variabilidad que la temperatura mínima. El histograma, diagrama de cajas y resumen descriptivo por meses se encuentran en los Anexos 27, 28 y 29, respectivamente. En el diagrama de cajas por meses (Anexo 30) se observó que las medianas para cada mes varían en un grado centígrado aproximadamente a lo largo del año.

A continuación se presentan las estadísticas descriptivas de la precipitación, la temperatura mínima y la temperatura máxima.

Cuadro 4.2: Estadísticas descriptivas para la estación de Jauja

	Mínimo	Máximo	Mediana	Media	Error Estándar de la Media	Desviación Estándar	Asimetría
Precipitación (mm)	0.00	286.70	52.30	60.600	2.6333	54.4775	0.827
Temp. Mínima (°C)	-3.20	9.00	5.09	4.378	0.1187	2.4240	-0.803
Temp. Máxima (°C)	15.60	23.50	19.30	19.320	0.6010	1.2499	-0.016

FUENTE: Elaboración propia

En los datos de la precipitación de Jauja se encontraron valores atípicos leves y extremos, sólo se eliminaron los valores atípicos extremos de cada mes, la representación gráfica se encuentra en el Anexo 20. En el mes de junio los valores que no se consideran para el análisis son los correspondientes a los años 1974 (20.6 mm), 1989 (18.3 mm), 1990 (44.8 mm) y 1992 (19.4 mm). Debido a que sus valores caen fuera del intervalo [0.0 mm, 17.4 mm] que representa los limites de los valores para ser considerados valores atípicos extremos. En el caso de la precipitación del mes de octubre se tiene como valores atípicos extremos a los años 1973 (120.3 mm), 1980 (131.6 mm) y 1998 (122.0 mm), estos valores caen fuera del intervalo [0.0 mm, 119.6 mm] y, por tanto, son eliminados. Ver Anexo 20.

Luego, al analizar la temperatura mínima en el diagrama de cajas por meses (Anexo 25), se muestra que las medianas son muy distintas de mes a mes. Los valores atípicos extremos presentes en esta data son dos: febrero (2.06 °C) y marzo (2.0 °C) de 1971. El intervalo de valores atípicos extremos para el mes de febrero es [2.1 °C, 10.8 °C], y para el mes de marzo es [2.4 °C, 10.3 °C]. Los valores atípicos leves no se eliminan, empero hay 6 en los datos.

Finalmente, se analizó la temperatura máxima de Jauja con el diagrama de cajas por meses (Anexo 30). En el estudio de los límites para los valores atípicos extremos no se encontró ningún valor que salga de los límites calculados, pero sí la presencia de 7 valores atípicos leves que no se removerán del estudio. También se observó en el Anexo 30 que la variabilidad en el mes de febrero es mayor que en los meses que se consideran homogéneos a él, que son enero y marzo. El mes de mayo es el que presenta menor variabilidad.

4.1.3. ESTACIÓN DE SANTA ANA

En la estación de Santa Ana se tienen las observaciones de la precipitación, temperatura mínima y máxima, en un período de tiempo muy corto, ya que se tienen datos medidos desde 1992 hasta 2004 (precipitación) y desde 1992 hasta 2002 (temperatura mínima y máxima).

Es debido a la poca cantidad de datos que no se podrían estimar pronósticos confiables para esta estación. Debido a lo expuesto anteriormente, ésta estación queda fuera del estudio.

4.1.4. ESTACIÓN DE VIQUES

Dentro de los datos de precipitación de Viques existió la presencia de datos ausentes (Anexo 31), las cuales no se imputaron por ende no fueron utilizadas en el estudio. La distribución de la precipitación es asimétrica a la derecha, es decir, la media se encuentra por encima de la mediana esto se debe a la presencia de valores muy grandes que jalan a la media hacia la derecha. Y la desviación estándar tiene un valor de 65.9 mm, esto muestra que la variabilidad es bastante alta en esta serie. El histograma, el diagrama de cajas y el resumen descriptivo por meses se encuentran en los Anexos 32, 33 y 34, respectivamente. Se observó en el diagrama de cajas por meses (Anexo 35) que las medias de los tres primeros meses del año son mucho más altas que en el resto del año, pero en comparación con las otras estaciones del estudio en Viques llueve menos.

A continuación se presentan las estadísticas descriptivas de la precipitación, la temperatura mínima y la temperatura máxima.

Cuadro 4.3: Estadísticas descriptivas para la estación de Viques

	Mínimo	Máximo	Mediana	Media	Error Estándar de la Media	Desviación Estándar	Asimetría
Precipitación (mm)	0.00	496.00	13.70	40.722	3.2898	65.8778	3.098

FUENTE: Elaboración propia

Se analizó la precipitación de Viques y se encontró que en los meses de enero y febrero valores atípicos extremos correspondientes al año 1973, son 473.0 mm y 496.0 mm, respectivamente, que fueron eliminados. Estos valores están fuera de los límites para valores atípicos extremos, a pesar de que ese año es considerado un año Niño, y se esperaría que la precipitación fuese mayor al promedio, pero valores tan altos pueden ser producto de un error sistemático. Ya que al revisar este año en otras estaciones, los valores encontrados en promedio no sobrepasan los 157.0 mm de precipitación. Los intervalos calculados para que las observaciones de enero y febrero sean consideradas valores atípicos extremos son [-272.1 mm; 425.1 mm] y [-321.8 mm; 495.1 mm], respectivamente. En mayo, se eliminaron tres valores pertenecientes a los años 1964, 1990 y 1991; cuyos valores son 42.0 mm, 110.2 mm, 23.0 mm, respectivamente. Éstos valores se encuentran fuera del intervalo calculado para mayo que es [0.0 mm; 18.6 mm]. En junio se elimina dos valores de pertenecen a los años 1990 y 1992, con valores de 136.7 mm y 14.4 mm, respectivamente. Ya que el intervalo límite para valores atípicos extremos es [0.0 mm; 13.3 mm]. En julio, se elimina una sola observación a pesar de que se calculan que cuatro son valores atípicos extremos, para el intervalo [0.0 mm; 3.2 mm]. Debido a que al revisar los valores para el mes de julio en las otras estaciones el máximo valor para la precipitación esta alrededor de 18.6 mm. Por esto, sólo se elimina el valor 96.1 mm perteneciente al año 1992, que excede por mucho al valor máximo esperado, y no el valor 21.1 mm del año 1997. En agosto, se eliminan tres observaciones que pertenecen a los años 1990, 1992 y 2000; con valores iguales a 31.0 mm, 35.0 mm y 35.1 mm, respectivamente. Éstos valores sobrepasaron el límite [0.0 mm; 17.2 mm]. En noviembre se elimina el valor 234.3 mm del año 1990, por estar fuera del intervalo [0.0 mm; 174.075 mm]. Finalmente, en diciembre se elimina también un valor, 342.6 mm del año 1978, que se encuentra fuera del intervalo [0.0 mm; 257.5 mm].

4.1.5. RESUMEN DEL AED

Se observó que las estaciones de Huayao y Jauja son bastante similares en cuanto a precipitación y temperaturas extremas. Al analizar por meses, se observó un comportamiento similar de éstas variables. En cambio Viques se diferencia de las dos estaciones antes mencionadas, porque la precipitación promedio es mucho menor y tiene

mayor variabilidad, por lo que se podría inferir que Viques representa un clima diferente al de las otras estaciones.

Las estaciones de Jauja y Viques presentaron valores ausentes, además de que ambas estaciones son las que presentan la mayor cantidad de valores atípicos extremos. Siendo Huayao la más completa y con menos valores atípicos extremos. Santa Ana por ser una estación con diez años de datos no fue elegible para entrar en el análisis por la falta de datos.

Se pudo apreciar que la precipitación sigue una distribución asimétrica a la derecha en las tres estaciones La temperatura mínima en las Huayao y Jauja presentó una distribución asimétrica a la izquierda y la temperatura máxima, en las mismas estaciones, tiene una distribución simétrica.

4.2. APLICACIÓN DEL MODELO MARS

Se tienen tres series de tiempo que corresponden a precipitación total mensual, temperatura mínima promedio mensual y temperatura máxima promedio mensual. Éstas variables, fueron particionadas en 12 grupos cada una; dichos grupos equivalen a los 12 meses del año.

Por ejemplo, en precipitación total mensual de Huayao, se tiene precipitación de todos los eneros, precipitación de todos los febreros, precipitación de todos los marzos, y así sucesivamente hasta llegar a diciembre.

El modelo MARS para esta aplicación queda definido como:

$$Y_{ij} = \alpha_{ij} + \sum_{m_{ij}=1}^{M_{ij}} \beta_{m_{ij}} \prod_{k=1}^{K} \left[S_{km} \left(x_{v(k,m)} - t_{km} \right) \right]_{+} , \; i = 1,2,3 , \; j = 1,2,3$$

donde

1. Y_{ij} es la j-ésima variable respuesta de la i-ésima estación meteorológica.
2. α_{ij} es el término constante del modelo.
3. $\beta_{m_{ij}}$ es el coeficiente de la m-ésima función base de la j-ésima variable respuesta en la i-ésima estación meteorológica.
4. $x_{v(k,m)}$ es la v-ésima variable explicativa que se encuentra en la m-ésima función base en el k-ésimo nodo de la j-ésima variable respuesta en la i-ésima estación meteorológica.
5. t_{km} es el valor del k-ésimo nodo en la m-ésima función base de la j-ésima variable respuesta en la i-ésima estación meteorológica.
6. S_{km} es el valor ± 1
7. La sumatoria representa, la suma de las M_{ij} funciones base calculadas para la j-ésima variable respuesta de la i-ésima estación meteorológica.

Luego teniendo las series sin valores atípicos extremos, se calcularon los modelos MARS para cada grupo. Estos modelos se calcularon hasta con dos y tres interacciones de las variables explicativas, y, además, con un desfase de los valores de las variables explicativas de tres meses. Es decir, que al generar los coeficientes del modelo, éstos tendrían funciones base asociadas que estarían conformadas por una sola variable o la combinación de dos o tres variables, según sea el caso. Cuando se calcula un modelo hasta con dos interacciones, se llama un modelo MARS de grado 2. Éste modelo, además de contener funciones base de una sola variable contiene, también, funciones base que son la combinación de dos variables. Asimismo, cuando se calcula un modelo de hasta con tres interacciones, se llama un modelo MARS de grado 3. Éste modelo, tiene funciones base de una sola variables, funciones base que son la combinación de dos variables y, además, funciones base que son la combinación de tres variables. Los resultados se analizaron con el software R versión 2.8.1, utilizando la librería mda.

La selección del mejor modelo se basa en el mejor conjunto de índices, en este análisis se utiliza índices como: GCV (*generalized cross-validation*: validación cruzada

generalizada), RSQ (*R-Squared of the model*: coeficiente de determinación del modelo), ECM (*mean squared error*: error cuadrático medio) y COR (*correlation*: correlación). El mejor modelo, es aquél que tenga un GCV y un ECM menor que el otro además de un RSQ y un COR mayor que el otro.

4.2.1. ANÁLISIS DE LOS DATOS CON MARS

ESTACIÓN DE HUAYAO

Precipitación

Los modelos de precipitación que fueron seleccionados en su mayoría son de grado 2, a excepción de los modelos para los meses de octubre y noviembre que son de grado 3. Para revisar los índices de cada modelo de precipitación, ver Anexo 36.

El error cuadrático medio (ECM) de los valores observados y los valores estimados es 421.394 mm^2, la raíz cuadrada del ECM (RECM) es 20.528 mm, el error absoluto medio (AMNEAM) es 13.889 mm, el error normalizado medio (EAMN) es 93.683 y la correlación entre las mismas series es de 0.916 (Cuadro 4.4). En general los errores son pequeños y la asociación lineal entre el observado y el estimado es alta.

Cuadro 4.4: Evaluación de los modelos de precipitación de Huayao utilizando MARS

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	COR
421.394	20.528	13.889	93.683	0.916

FUENTE: Elaboración propia

En el Cuadro 4.5 se encuentran las variables que intervienen en cada uno de los modelos. Esta selección de variables es determinada por el algoritmo MARS. El modelo que presenta la menor cantidad de variables es noviembre con sólo dos variables en el modelo y son EA y PNA.

Cuadro 4.5: Variables que intervienen en los modelos MARS de precipitación de Huayao

Mes	Variables que intervienen en el modelo
Enero	PDO, CAR, EA
Febrero	EAWR, PDO, SOI, PNA, SCA
Marzo	WP, N4, NAO, TSA, PDO
Abril	N12, TNA, D, SCA, PDO
Mayo	N3, NAO, SOI, EAWR
Junio	N4, SCA, TNA, D
Julio	NAO, EA, PNA
Agosto	NAO, N4, EA, N12, N3
Setiembre	N4, D, TSA
Octubre	PDO, SCA, NAO
Noviembre	EA, PNA
Diciembre	N3, D, PNA, TSA

FUENTE: Elaboración propia

Temperatura Mínima

Los modelos de temperatura mínima que se seleccionaron son en su mayoría del grado 2, con excepción de los meses de enero, julio, agosto y setiembre que son de grado 3, ya que presentan mejores indicadores. Ver Anexo 37.

El ECM asociado a las estimaciones en conjunto es 0.380 $°C^2$, el RECM es 0.616 °C, el EAM es 0.469 °C, el EAMN es 2.305 y la correlación es 0.970 (Cuadro 4.6). En general el error observado es pequeño y existe una asociación alta de las series.

Cuadro 4.6: Evaluación de los modelos de temperatura mínima de Huayao de utilizando MARS

ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	COR
0.380	0.616	0.469	2.305	0.970

FUENTE: Elaboración propia

En la Cuadro 4.7 se indica que variables intervienen en cada modelo de temperatura mínima que han sido seleccionadas por el algoritmo del modelo MARS. El mes que tiene menos variables en su modelo es mayo, con las variables N3 y CAR. Y el mes que tiene mayor cantidad de variables es octubre, con la presencia de 8 de las 15 variables utilizadas en el análisis que son SOI, N34, PDO, EA, N12, WP, D, N4

Cuadro 4.7. Variables que intervienen en los modelos MARS de temperatura mínima de Huayao

Mes	Variables que intervienen en el modelo
Enero	NAO, PDO, PNA, TSA, N3
Febrero	PDO, N12, TNA, EA, CAR
Marzo	TSA, WP, PNA, N12, CAR
Abril	CAR, TSA, TNA, WP
Mayo	N3, CAR
Junio	NAO, TNA, N34, CAR
Julio	N34, TNA, WP, EA, N4, SOI
Agosto	SOI, N34, PDO, EA, N12, WP, D, N4
Setiembre	CAR, WP, PNA, D, TSA
Octubre	N34, PDO, N3, N4, TNA, TSA, PNA, CAR
Noviembre	N34, PNA, N3, N4, WP, TNA
Diciembre	N34, PDO, PNA, TSA, CAR, SOI

FUENTE: Elaboración propia

Temperatura Máxima

Los modelos de temperatura máxima que se han seleccionado son en su mayoría del grado 2, salvo marzo, mayo, junio y agosto donde los modelos de grado 3 resultaron seleccionados (ver Anexo 38). En el caso del mes de octubre no se pudo trazar un modelo, esto se debe que no se puede trazar alguna relación funcional entre la variable respuesta y alguna de las variables explicativas, por lo que quedó como una constante 20.30 °C.

El ECM asociado a las estimaciones totales es 0.437 $°C^2$, el RECM es 0.661 °C, el EAM es 0.506 °C, el EAMN es 0.026 y la correlación es 0.834. Es decir, que existe poco error de estimación y las dos series están asociadas positivamente. (Cuadro 4.8)

Cuadro 4.8: Evaluación de los modelos de temperatura máxima de Huayao utilizando MARS

ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	COR
0.437	0.661	0.506	0.026	0.834

FUENTE: Elaboración propia

En la Cuadro 4.9, se encuentran las variables que intervienen en cada modelo. En el mes de octubre el algoritmo MARS no encuentra alguna variable apropiada por lo que

queda como una constante igual a 20.30mm. En el mes de enero sólo una variable esta presente la cual es N3.

Cuadro 4.9. Variables que intervienen en los modelos MARS de temperatura máxima de Huayao

Mes	Variables que intervienen en el modelo
Enero	N3
Febrero	N3, N34, PNA, EA
Marzo	N3, N34, N4, TNA, PDO
Abril	N4, N34, EA
Mayo	N12, N34, NAO, PNA, CAR
Junio	N4, PDO, N12, T, TNA, WP, CAR
Julio	N34, PDO, SOI, N12, NAO, TNA
Agosto	TSA, PDO, CAR, NAO
Setiembre	PNA, T, TNA, SOI
Octubre	-
Noviembre	WP, SOI, N12, N4, NAO, WP, TSA
Diciembre	N12, TSA, PNA

FUENTE: Elaboración propia

ESTACIÓN DE JAUJA

Precipitación

Los modelos de precipitación en Jauja seleccionados son en su mayoría del grado 2, con excepción de los modelos para los meses de junio y julio que son de grado 3 (ver Anexo 39). No se pudo calcular un modelo adecuado para los meses de enero, noviembre y diciembre, estos quedaron como constantes 125.29 mm, 72.06 mm y 99.16 mm; respectivamente.

El ECM asociado a las estimaciones es 598.988 mm^2, el RECM es 24.474 mm, el EAM es 16.074 mm, el EAMN es 188.595 y la correlación es 0.893 (Cuadro 4.10). Esto indica que hay cierto grado de error, debe tomarse en cuenta que la precipitación tiene grandes magnitudes, por lo que los errores pueden ser más amplios que lo esperado. Existe un alto grado de asociación positiva entre los valores observados y las estimaciones.

Cuadro 4.10: Evaluación de los modelos de precipitación de Jauja utilizando MARS

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	COR
598.988	24.474	16.074	188.595	0.893

FUENTE: Elaboración propia

En la Cuadro 4.11, se muestran las variables que han sido seleccionadas por el algoritmo del modelo MARS para la estimación de los modelos de cada mes. Los meses de enero, noviembre y diciembre muestran un guión ya que para esos meses no se pudo encontrar un modelo adecuado en donde intervenga al menos una variable explicativa, dejándose como constantes 125.29 mm, 72.06 mm y 99.16 mm respectivamente. Luego, en el mes de setiembre sólo se incluyen dos variables, N3 y WP. El mes con mayor cantidad de variables en su modelo es el mes de junio con 7 variables que son: PNA, N3, N34, PDO, PNA, EAWR, TSA.

Cuadro 4.11. Variables que intervienen en los modelos MARS de precipitación de Jauja

Mes	Variables que intervienen en el modelo
Enero	-
Febrero	TSA, EA, CAR, SOI
Marzo	PDO, WP, SCA
Abril	N12, N34, SCA
Mayo	TNA, EA, TSA
Junio	PNA, N3, N34, PDO, PNA, EAWR, TSA
Julio	NAO, PNA, EAWR, CAR, TSA
Agosto	SOI, N3, EA, PNA, EAWR, CAR
Setiembre	N3, WP
Octubre	NAO, PDO
Noviembre	-
Diciembre	-

FUENTE: Elaboración propia

Temperatura Mínima

La mayoría de los modelos seleccionados para la temperatura mínima en Jauja son de grado 2 a excepción de los modelos para los meses de agosto y setiembre, que son de grado 3. Ver Anexo 40.

El ECM asociado a las estimaciones es 0.915 °C^2, el RECM es 0.957 °C, el EAM es 0.734 °C, el EAMN es 0.183 y la correlación es 0.919. El error es pequeño y existe una alta asociación de las variables (Cuadro 4.12).

Cuadro 4.12: Evaluación de los modelos de temperatura mínima de Jauja utilizando MARS

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	COR
0.915	0.957	0.734	0.183	0.919

FUENTE: Elaboración propia

En la Cuadro 4.13 se encuentran las variables que intervienen en los modelos. El lector puede verificar que en los meses de febrero y abril la variable que interviene en el modelo es N12, siendo estos meses los que tienen la menor cantidad de variables. El mes con la mayor cantidad de variables en su modelo es agosto con 6 variables que son: PDO, N4, N12, TNA, T, CAR.

Cuadro 4.13. Variables que intervienen en los modelos MARS de temperatura mínima de Jauja

Mes	Variables que intervienen en el modelo
Enero	N4, TSA, SOI, PDO, D
Febrero	N12
Marzo	N4, N34
Abril	N12
Mayo	N3, NAO
Junio	PDO, NAO
Julio	SOI, TNA
Agosto	PDO, N4, N12, TNA, T, CAR
Setiembre	CAR, WP, PNA, PDO, D
Octubre	N4, PDO, EA, T
Noviembre	PNA, PDO
Diciembre	N34, WP, TNA, EA, TSA

FUENTE: Elaboración propia

Temperatura Máxima

La mayoría de los modelos para la temperatura máxima en Jauja son de grado 2, a excepción de los modelos para los meses de abril y octubre que son de grado 3 (ver Anexo 41).

El ECM asociado a las estimaciones es 0.588 °C^2, el RECM es 0.767 °C, el EAM es 0.570 °C, el EAMN es 0.03 y la correlación es 0.790. Se tiene un error bajo y una correlación media alta. (Cuadro 4.14)

Cuadro 4.14: Evaluación de modelos de temperatura máxima de Jauja utilizando MARS

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	COR
0.588	0.767	0.570	0.030	0.790

FUENTE: Elaboración propia

En la Cuadro 4.15 se muestran las variables que intervienen en los modelos. El mes que tiene la menor cantidad de variables en su modelo es julio, con la presencia de la variable TNA. Y el mes con la mayor cantidad de variables es abril, con 6 variables en su modelo que son: N4, NAO, SOI, PNA, TNA, TSA.

Cuadro 4.15. Variables que intervienen en los modelos MARS de temperatura máxima de Jauja

Mes	Variables que intervienen en el modelo
Enero	N3, TSA
Febrero	N3, NAO, TSA
Marzo	N34, NAO, D, PNA, SOI
Abril	N4, NAO, SOI, PNA, TNA, TSA
Mayo	N34, TSA
Junio	PNA, TSA, N3, T, TNA
Julio	TNA
Agosto	TNA, SOI, N4, EA
Setiembre	TNA, N12, PNA
Octubre	N12, N4, WP, SOI
Noviembre	PNA, TSA
Diciembre	N34, TSA

FUENTE: Elaboración propia

ESTACIÓN DE VIQUES

Los modelos seleccionados para la precipitación en Viques, en su mayoría son de grado 2 a excepción de agosto y octubre, que obtuvieron mejores resultados con modelos de grado 3. Ver Anexo 42.

El ECM entre valores observados y estimados es 1419.999 mm^2, el RECM es 37.683 mm, el EAM es 21.346 mm, el EAMN es 155.851 y la correlación es 0.740. Se muestra que hay un grado elevado de error, y se tiene una asociación media alta entre las mismas series. (Cuadro 4.16)

Cuadro 4.16: Evaluación de modelos de precipitación de Viques utilizando MARS

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	COR
1419.999	37.683	21.346	155.851	0.740

FUENTE: Elaboración propia

En la Cuadro 4.17 se muestran las variables que están presentes en los modelos. Los meses de enero y marzo son constantes iguales a 80.07 mm y 77.04 mm, respectivamente. El mes donde sólo una variable está presente es noviembre con la presencia de CAR.

Cuadro 4.17. Variables que intervienen en los modelos MARS de precipitación de Viques

Mes	Variables que intervienen en el modelo
Enero	-
Febrero	SCA, EAWR
Marzo	-
Abril	N12, TSA, EA, SOI
Mayo	N34, NAO
Junio	TNA, TSA, N3, NAO
Julio	N4, WP, SCA
Agosto	N4, N34, PDO, PNA, EA
Setiembre	D, PDO, PNA
Octubre	CAR, EAWR, N4, SCA, SOI
Noviembre	CAR
Diciembre	PDO, PNA, CAR, TNA

FUENTE: Elaboración propia

4.2.2. VALIDACIÓN DEL MODELO MARS

ESTACIÓN DE HUAYAO

Precipitación

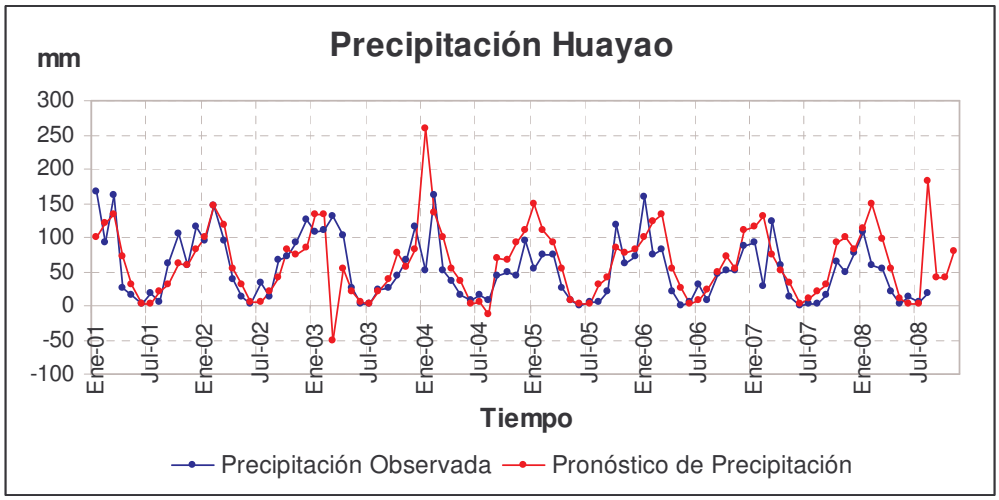
El ECM asociado a los pronósticos es 2069.60 mm^2, el RECM es 45.493 mm, el EAM es 28.87 mm, el EAMN es 22.645 y la correlación es 0.579. También presenta un sesgo promedio de 10.031 mm (Cuadro 4.18). Esta presente un alto grado de error en los pronósticos debido a dos picos negativo iguales a -51.21 mm en marzo de 2003 y otro de -13.61 mm en agosto de 2004, además de dos picos positivos iguales a 258.16 mm en enero de 2004 y otro de 191.31 mm en agosto de 2008 (ver Figura 4.1). Estos valores pueden producirse por cambios inusuales en las variables explicativas que intervienen en cada modelo.

Cuadro 4.18: Evaluación del pronóstico de precipitación de Huayao utilizando MARS

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
2069.595	45.493	28.870	22.645	10.031	0.579

FUENTE: Elaboración propia

Figura 4.1. Precipitación de Huayao. Valores observados y los valores pronóstico con MARS



FUENTE: Elaboración propia

Temperatura Mínima

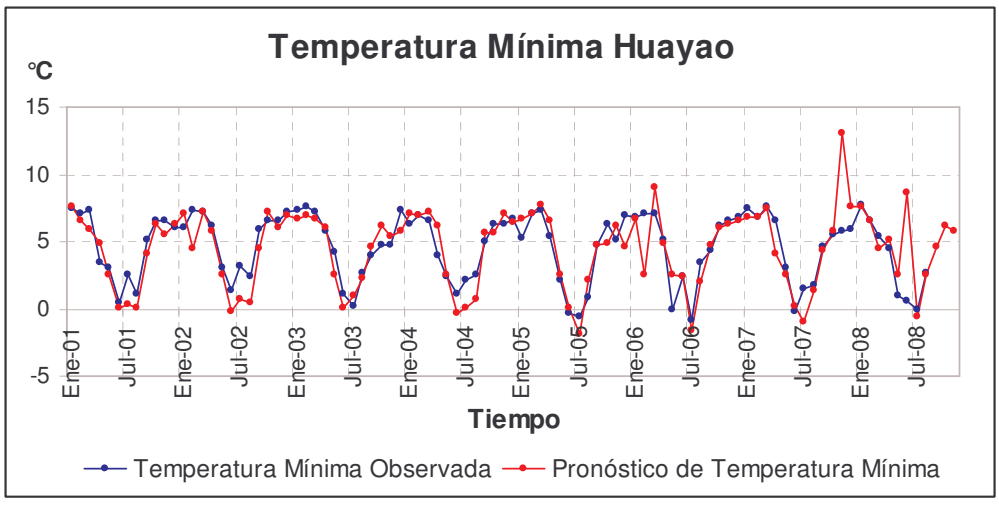
El ECM observado en los modelos de temperatura mínima es 2.773 °C^2, el RECM es 1.665 °C, el EAM es 1.063 °C, el EAMN es 0.069 y la correlación es 0.816. Presenta un sesgo promedio de -0.126 °C (Cuadro 4.19). La Figura 4.2 muestra el gráfico de las series de temperatura mínima observada y los valores pronóstico, en el gráfico se puede observar que las series son muy similares en forma, como indica la correlación antes mencionada. En el último año la serie pronosticada se tienen dos picos uno de 13.05 °C en noviembre de 2007 y 8.68 °C en junio de 2008 (Figura 4.2). Estos picos pueden ser causados por algún valor inusual de las variables explicativas.

Cuadro 4.19: Evaluación de pronósticos de temperatura mínima de Huayao utilizando MARS

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
2.773	1.665	1.063	0.069	-0.126	0.816

FUENTE: Elaboración propia

Figura 4.2. Temperatura mínima de Huayao. Valores observados y valores pronóstico con MARS



FUENTE: Elaboración propia

Temperatura Máxima

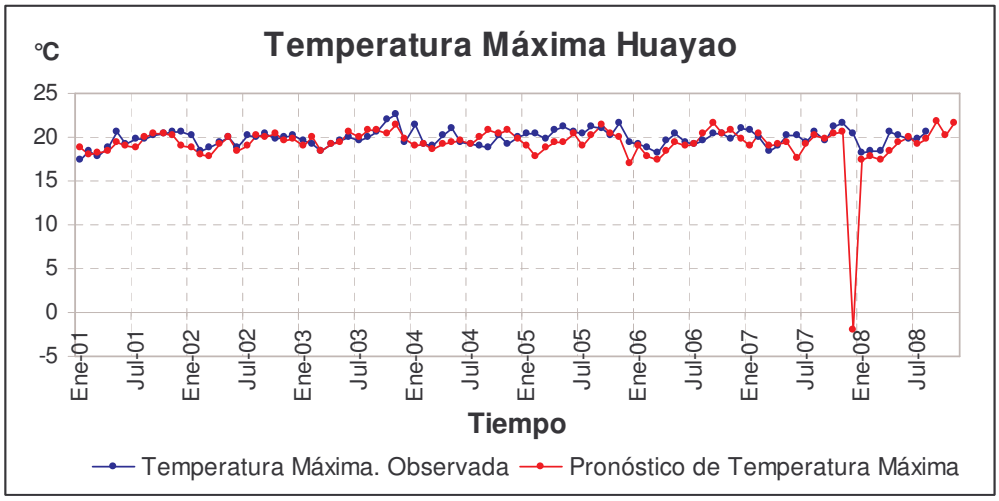
El ECM asociado a las predicciones de temperatura máxima es 6.494 °C^2, el RECM es 2.548 °C, el EAM es 1.044 °C, el EAMN es 0.052 y su correlación es 0.163. El sesgo promedio que se obtiene es -0.711 °C (Cuadro 4.20). Se observa que el ECM es relativamente bajo, pero la correlación es bastante pequeña, esto se debe al valor inusual calculado de –1.96 °C en diciembre de 2007 (Figura 4.3). Ya que al eliminar ese valor inusual, los valores del ECM, RECM, EAM, EAMN y la correlación cambian a 1.071 °C^2, 1.035 °C, 0.81 °C, 0.04 y 0.538 respectivamente. El sesgo disminuye a -0.473 °C. Alguno o algunos de los valores de las variables explicativas deben presentar valores inusuales que influyen en el resultado. El valor eliminado no será removido del estudio, sólo se calculo para demostrar la diferencia que existe al no tener un calculo inusual. Estos cálculos inusuales nuevamente podrían deberse a cambio en los valores de las variables explicativas.

Cuadro 4.20: Evaluación de pronósticos de temperatura máxima de Huayao utilizando MARS

	ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
Todos los pronósticos	6.494	2.548	1.044	0.052	-0.711	0.163
Sin el valor inusual	1.071	1.035	0.810	0.040	-0.473	0.538

FUENTE: Elaboración propia

Figura 4.3: Temp. Máxima Huayao. Valores observados y valores pronóstico con MARS



FUENTE: Elaboración propia

ESTACIÓN DE JAUJA

Precipitación

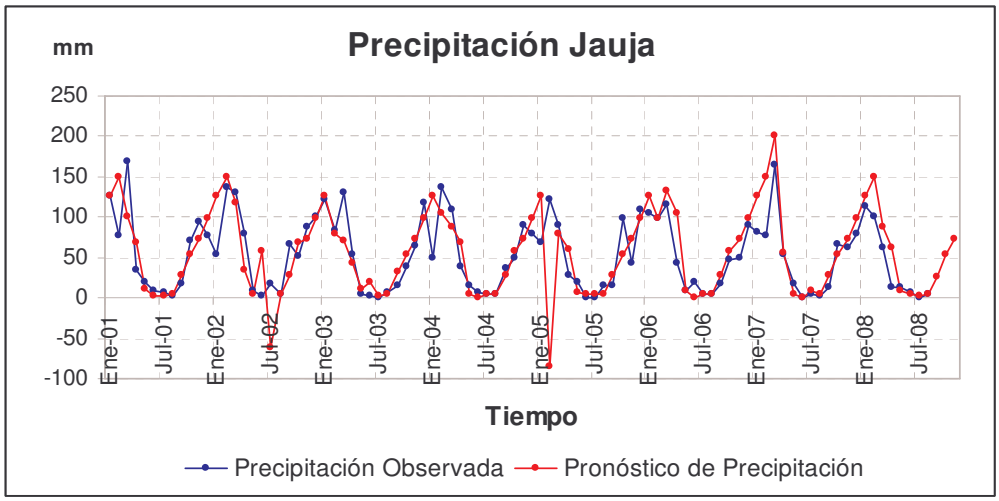
El ECM es 1277.648 mm2, el RECM es 35.744 mm, el EAM es 21.784 mm, el AMN es 64.646 y la correlación es 0.741. El sesgo promedio es 3.004 mm (Cuadro 4.21). El error es grande, este error se debe a los dos picos negativos –61.64 mm y –84.51 mm que corresponden a julio de 2002 y febrero de 2005 respectivamente. Estos picos influyen también en la correlación. Ver Figura 4.4.

Cuadro 4.21: Evaluación de pronósticos de precipitación de Jauja utilizando MARS

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
1277.648	35.744	21.784	64.646	3.004	0.741

FUENTE: Elaboración propia

Figura 4.4. Precipitación de Jauja. Valores observados y valores pronóstico con MARS



FUENTE: Elaboración propia

Temperatura Mínima

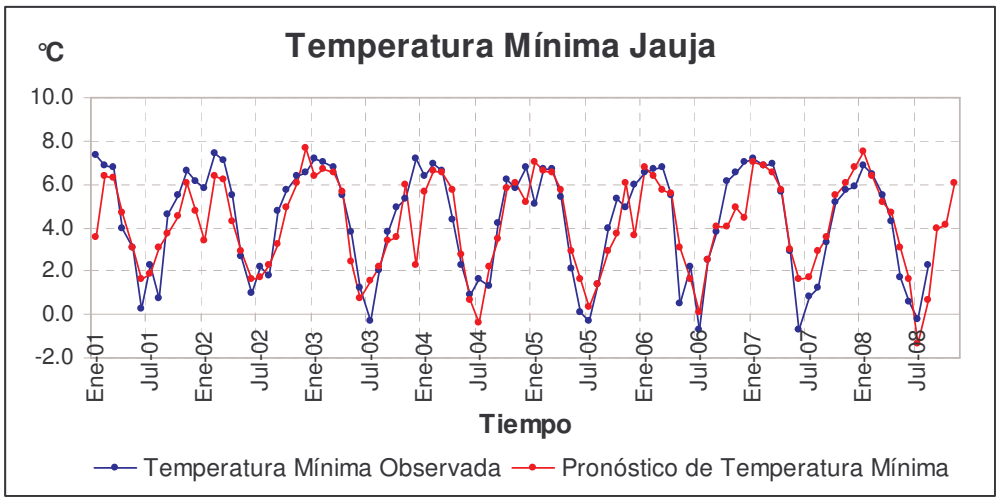
El ECM es 1.561 °C^2, el RECM es 1.249 °C, el EAM es 0.918 °C, el EAMN es 0.339 y la correlación es 0.861. El sesgo promedio es -0.186 °C (Cuadro 4.22). El error es bajo y la correlación alta, dado que no hay ningún pico inusual. En la Figura 4.5 se muestra la gráfica de las series, y la serie pronóstico es muy similar a la serie observada. Esto concuerda con la correlación alta y positiva que se había calculado.

Cuadro 4.22: Evaluación de pronósticos de temperatura mínima de Jauja utilizando MARS

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
1.561	1.249	0.918	0.339	-0.186	0.861

FUENTE: Elaboración propia

Figura 4.5. Temperatura mínima de Jauja. Valores observados y valores pronóstico con MARS



FUENTE: Elaboración propia

Temperatura Máxima

El ECM es 1.739 °C^2, el RECM es 1.319 °C, el EAM es 1.093, el EAMN es 0.06 y la correlación es 0.464. El sesgo promedio es 0.774 (Cuadro 4.23). El error presente en la serie pronostico es relativamente bajo, del mismo modo que la correlación es baja. Esta correlación baja se debe a que por espacios de tiempo la serie pronóstico sobreestima a la

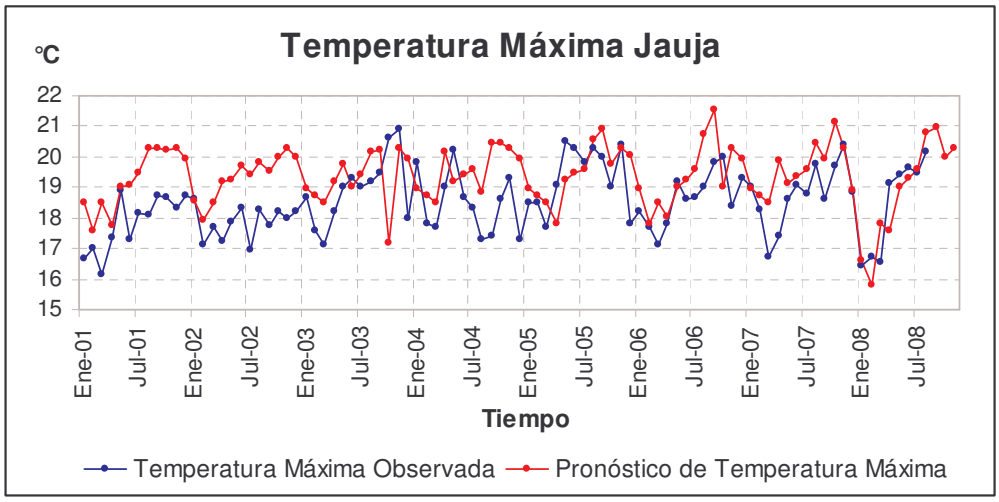
observada. En la Figura 4.6, se muestra la gráfica de las series observada y pronóstico, se observa que la serie pronóstico casi en la mayor parte del tiempo esta sobre la serie observada. Con excepciones como en enero de 2003 que se estima 17.16 °C y el valor observado es 20.6 °C, más de tres grados de diferencia, en este caso el pronóstico es inferior al observado.

Cuadro 4.23: Evaluación de pronósticos de temperatura máxima de Jauja utilizando MARS

ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
1.739	1.319	1.093	0.060	0.774	0.464

FUENTE: Elaboración propia

Figura 4.6. Temperatura máxima de Jauja. Valores observados y valores pronóstico con MARS



FUENTE: Elaboración propia

ESTACIÓN DE VIQUES

El ECM es 1248.179 mm^2, el RECM es 35.33 mm, el EAM es 24.167 mm, el EAMN es 376.931 y la correlación es 0.685. El sesgo promedio es -10.105 mm (Cuadro 4.24). El error es relativamente bajo y la correlación media alta. No existen picos inusuales. En la Figura 4.7 se muestra la gráfica de las series observada y pronostico, que son bastante similares. En algunos puntos la serie pronóstico sobrestima a la observada

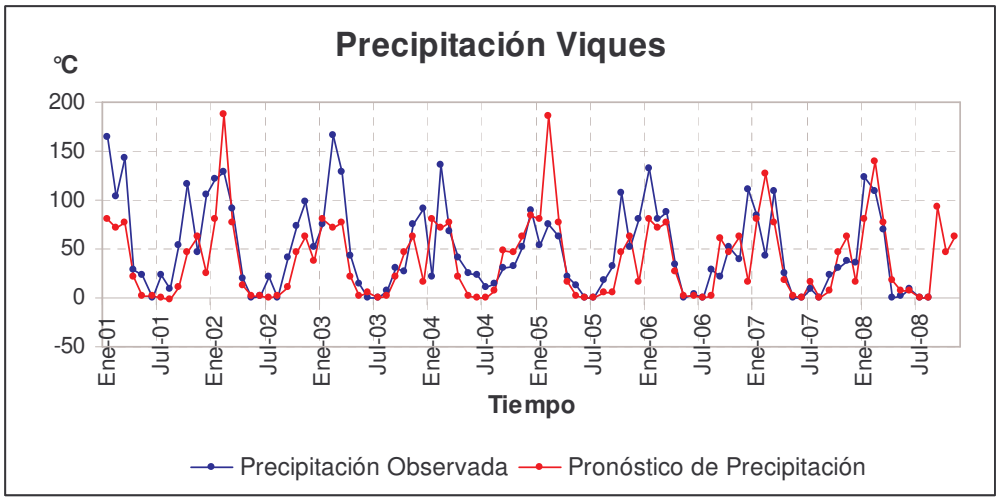
como en el caso del periodo enero – marzo de 2002 y 2005, sin embargo, entre agosto de 2002 – marzo de 2003 y julio de 2005 – marzo de 2006 la serie pronóstico subestima a la observada.

Cuadro 4.24: Evaluación de pronósticos de precipitación de Viques utilizando MARS

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
1248.179	35.330	24.167	376.931	-10.105	0.685

FUENTE: Elaboración propia

Figura 4.7. Precipitación de Viques. Valores observados y valores pronóstico con MARS



FUENTE: Elaboración propia

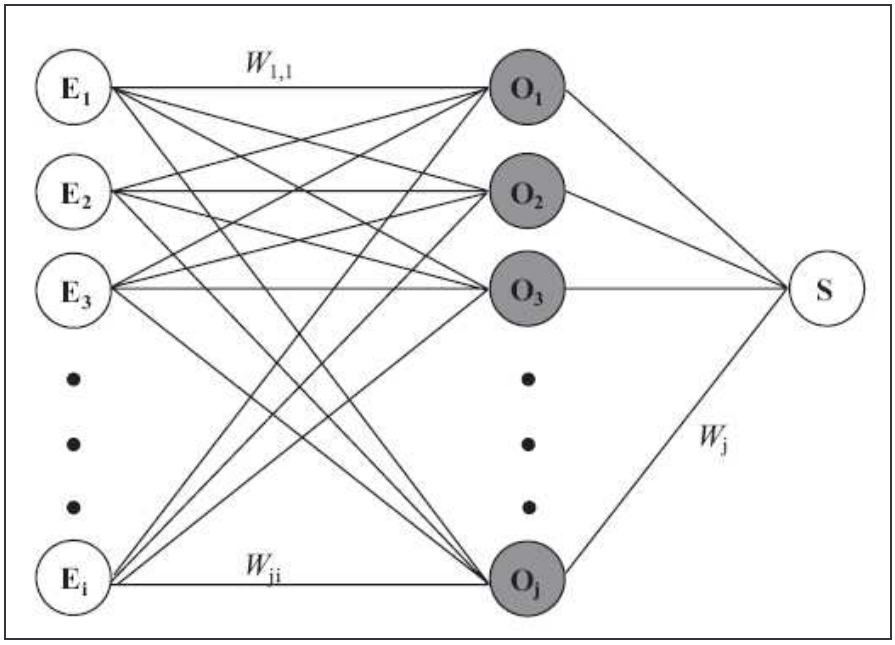
4.3. APLICACIÓN DE LAS REDES NEURONALES ARTIFICIALES BACKPROPAGATION

Para aplicar las redes neuronales, se tomaron en cuenta todos los valores de las series, a excepción de aquellos valores que resultaron ser valores atípicos extremos dentro del AED. Debido a que para aplicar una red neuronal se necesita una gran cantidad de datos, el conjunto final no se clasifica por meses y se calcula una sola red por variable respuesta.

En todas las redes se utilizó una capa con el número de neuronas igual al número de variables explicativas en estudio. En el caso de la precipitación el número de neuronas ocultas que se utilizó fue 17, esto se debió a que la cantidad de variables utilizadas para el análisis de las precipitaciones fue 17, es decir, el numero de neuronas fue igual al numero de variables. En la temperatura mínima y la temperatura máxima se utilizaron 15 puesto que las variables asociadas a la temperatura son 15. El análisis se realizó utilizando el software WEKA 3.4.

A continuación se presenta el esquema general de la RNAB utilizada en el análisis de los datos (Figura 4.8). La red neuronal propuesta en el presente trabajo consta de tres capas de neuronas. Esta compuesta por una capa de entrada cuyas neuronas son Ei y una capa de neuronas ocultas Oj y una capa de neuronas de salida, S. Durante la fase de entrenamiento de la red, se determinan iterativamente los pesos de las conexiones, Wji que conectan a las neuronas de entrada con las de la capa oculta y Wj para la transferencia entre la capa oculta y la de salida. En este caso i,j = 17 ó 15, 17 para precipitación y 15 para la temperatura mínima y la temperatura máxima.

Figura 4.8: Esquema general de conexiones de las redes neuronales utilizadas en el análisis de precipitación y temperatura de la Cuenca del río Mantaro.



FUENTE: Elaboración propia

4.3.1. ANÁLISIS DE LOS DATOS CON RNAB

ESTACIÓN DE HUAYAO

Precipitación

El ECM es 1601.263 mm^2, el RECM es 10.016 mm, el EAM es 33.555 mm, el EAMN es 931.583 y la correlación es 0.867. El sesgo promedio es 29.259 mm (Cuadro 4.25). El error entre los valores observados y los valores estimados es relativamente bajo con una correlación alta. Hay momentos en los que se sobreestima al valor observado en más de 119 mm como es el caso de setiembre de 1972 y setiembre de 1987.

Cuadro 4.25: Evaluación de la estimación de precipitación de Huayao utilizando RNAB

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
1601.263	10.016	33.555	931.583	29.259	0.867

FUENTE: Elaboración propia

Temperatura Mínima

El ECM es 1.213 °C2, el RECM es 1.101 °C, el EAM es 0.853 °C, el EAMN es 5.257 y la correlación es 0.912. El sesgo promedio es 0.367 °C (Cuadro 4.26). El error es bajo y existe una correlación alta entre las series. Existen algunos momentos en los que se sobreestima por más de 4.0 °C al observado como es el caso de marzo de 1998 y agosto de 1954 y 1958.

Cuadro 4.26: Evaluación de la estimación de temperatura mínima de Huayao utilizando RNAB

ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
1.213	1.101	0.853	5.257	0.367	0.912

FUENTE: Elaboración propia

Temperatura Máxima

El ECM es 0.627 °C^2, el RECM es 0.792 °C, el EAM es 0.633 °C, el EAMN es 0.033 y la correlación es 0.794. El sesgo promedio es 0.304 °C (Cuadro 4.27). El error es bajo y la correlación es media alta entre los valores observados y los valores estimados. Hay algunos momentos en los que la diferencia entre el observado y el estimado es mayor a $|\pm 2|$ °C, como es el caso de enero de 1970 y noviembre de 1993 que sobreestiman al observado y en el caso de noviembre de 2000 que subestima al observado.

Cuadro 4.27: Evaluación de la estimación de temperatura máxima de Huayao utilizando RNAB

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
0.627	0.792	0.633	0.033	0.304	0.794

FUENTE: Elaboración propia

ESTACIÓN DE JAUJA

Precipitación

El ECM es 693.227 mm^2, el RECM es 26.329 mm, el EAM es 20.516 mm, el EAMN es 1322.013 y la correlación es 0.884. El sesgo promedio es 6.704 mm (Cuadro 4.28). El error entre los valores observados y los valores estimados es bajo, y la correlación es alta. En el caso de la sobrestimación se tiene que marzo de 1960 y 1970; y febrero de 1968 y 1990 sobrepasan al observado en vas de 62 mm. En diciembre de 1961 y 1981, octubre de 1967 y enero de 1978 se subestima al observado en más de 69mm.

Cuadro 4.28: Evaluación de la estimación de precipitación de Jauja utilizando RNAB

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
693.227	26.329	20.516	1322.013	6.704	0.884

FUENTE: Elaboración propia

Temperatura Mínima

El ECM es 1.426 °C^2, el RECM es 1.194 °C, el EAM es 0.924 °C, el EAMN es 0.280 y la correlación es 0.908. El sesgo promedio es 0.624 °C (Cuadro 4.29). El error es bajo y la correlación es alta. Sin embargo, hay momentos en los que se sobreestima al observado en más de 3.5 °C, como por ejemplo junio de 1960, abril de 1963, enero de 1971, mayo de 1971 y diciembre de 1981. De los cuales enero de 1971 es el que presenta la mayor diferencia (4.5 °C), y diciembre de 1981 (3.9 °C).

Cuadro 4.29: Evaluación de la estimación de temperatura mínima de Jauja utilizando RNAB

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
1.426	1.194	0.924	0.280	0.624	0.908

FUENTE: Elaboración propia

Temperatura Máxima

EL ECM es 0.825 °C2, el RECM es 0.908 °C, el EAM es 0.727 °C, el EAMN es 0.038 y la correlación es 0.804. El sesgo promedio es 0.524 °C (Cuadro 4.30). Los momentos en que se sobrestima en más de 2.5 °C son: noviembre de 1961, enero de 1972, marzo de 1986 y octubre de 1999.

Cuadro 4.30: Evaluación de la estimación de temperatura máxima de Jauja utilizando RNAB

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
0.825	0.908	0.727	0.038	0.524	0.804

FUENTE: Elaboración propia

ESTACIÓN DE VIQUES

El ECM es 4438.011 mm^2, el RECM es 66.618 mm, el EAM es 57.560 mm, el EAMN es 8144.645 y la correlación es 0.809. El sesgo promedio es -57.088 mm (Cuadro 4.31). El error presente entre los valores observados y los valores estimados es bastante

alto, sin embargo, la correlación es alta, esto se debe a que la forma de la serie estimada es muy similar a la serie observada. El valor que presenta la mayor diferencia es marzo de 1979, en este momento se subestima al observado por 276.4 mm.

Cuadro 4.31: Evaluación de la estimación de precipitación de Viques utilizando RNAB

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
4438.011	66.618	57.560	8144.645	-57.088	0.809

FUENTE: Elaboración propia

4.3.2. VALIDACIÓN DE LAS RNAB

ESTACIÓN DE HUAYAO

Precipitación

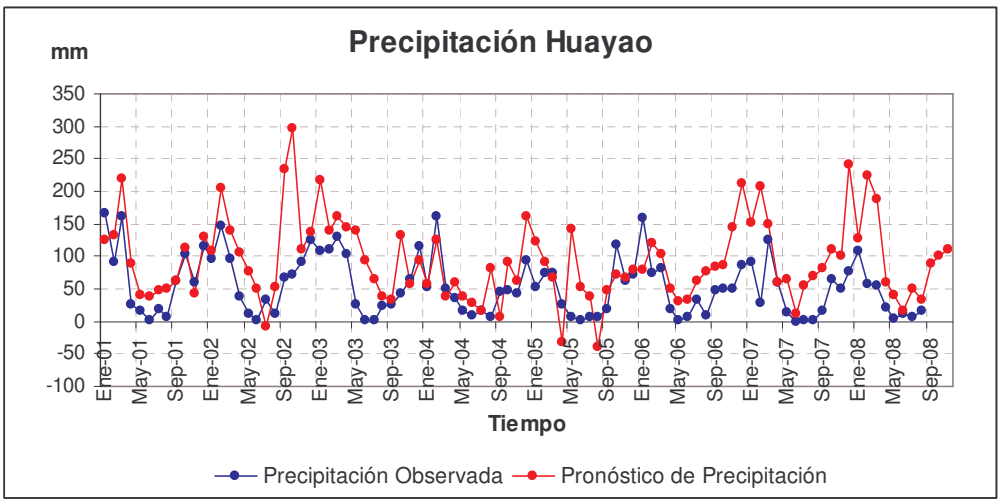
El ECM es 4280.598 mm^2 y la correlación es 0.575 (Cuadro 4.32). El error es bastante grande. El BIAS muestra una sobreestimación de 39.289 mm. El EAMN es alto. Se puede ver en la Figura 4.9 que existe un error bastante grande.

Cuadro 4.32: Evaluación de pronósticos de precipitación de Huayao utilizando RNAB

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
4280.598	65.426	49.127	114.836	39.289	0.575

FUENTE: Elaboración propia

Figura 4.9. Precipitación de Huayao. Valores observados y valores pronóstico con RNAB



FUENTE: Elaboración propia

Temperatura Mínima

El ECM es 4.169 °C^2 y la correlación es 0.732. Según el RECM el error es de 2.042 °C, es una variación pequeña. Si se observa el BIAS se puede ver que en promedio se esta sobrestimando 0.8 °C. La correlación es alta. (Cuadro 4.33)

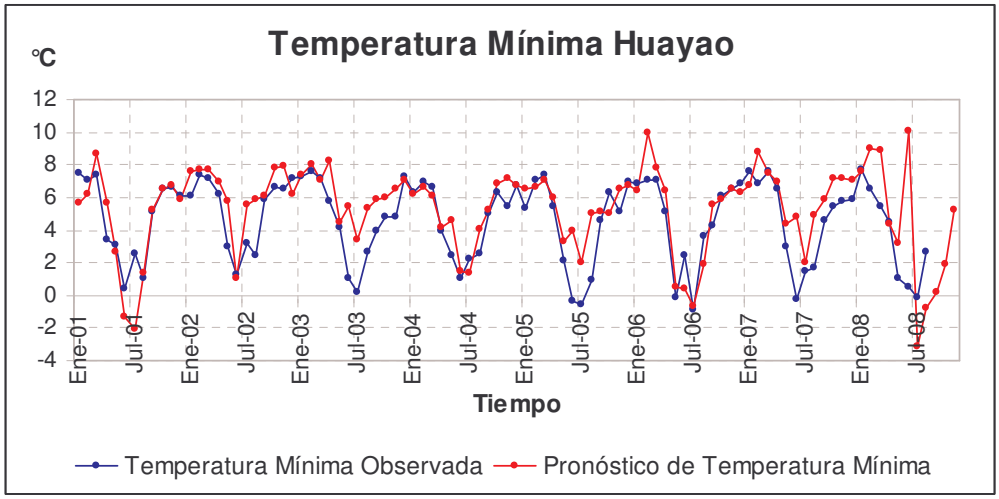
En la Figura 4.10 se puede observar el gráfico de la series pronóstico y la serie observada. La serie pronóstico sigue la forma de la serie observada, y hay presencia de algunos picos en los últimos meses.

Cuadro 4.33: Evaluación de pronósticos de temperatura mínima de Huayao utilizando RNAB

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
4.169	2.042	1.416	-0.131	0.808	0.732

FUENTE: Elaboración propia

Figura 4.10. Temperatura mínima de Huayao. Valores observados y valores pronóstico con RNAB



FUENTE: Elaboración propia

Temperatura Máxima

El ECM es 1.208 $°C^2$ y la correlación es 0.408. El error es pequeño según el RECM, EAM y el EAMN. El BIAS indica que en promedio se esta subestimando 0.169 °C. La correlación es baja. (Cuadro 4.34)

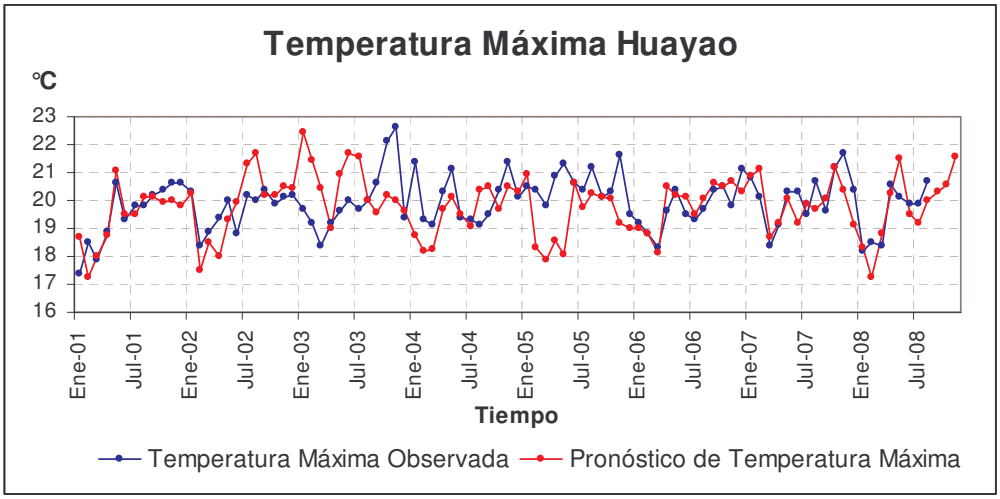
La Figura 4.11 muestra la gráfica de la serie observada versus la serie pronóstico. Se puede percibir que en algunos momentos como en el año 2004 el pronóstico tiende a subestimar y en el año 2002 hay una tendencia a la sobreestimación.

Cuadro 4.34: Evaluación de pronósticos de temperatura máxima de Huayao utilizando RNAB

ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
1.208	1.099	0.814	0.041	-0.169	0.408

FUENTE: Elaboración propia

Figura 4.11. Temperatura máxima de Huayao. Valores observados y valores pronóstico con RNAB



FUENTE: Elaboración propia

ESTACIÓN DE JAUJA

Precipitación

El ECM es 2417.902 mm^2 es bastante alto, indicando un alto error. En cuanto al RECM y EAM son aceptables ya que muestran un error general entre 38.5 mm y 49.2 mm. El BIAS muestra que se esta sobreestimando en promedio en 13.178 mm. La correlación es alta con 0.61. (Cuadro 4.35)

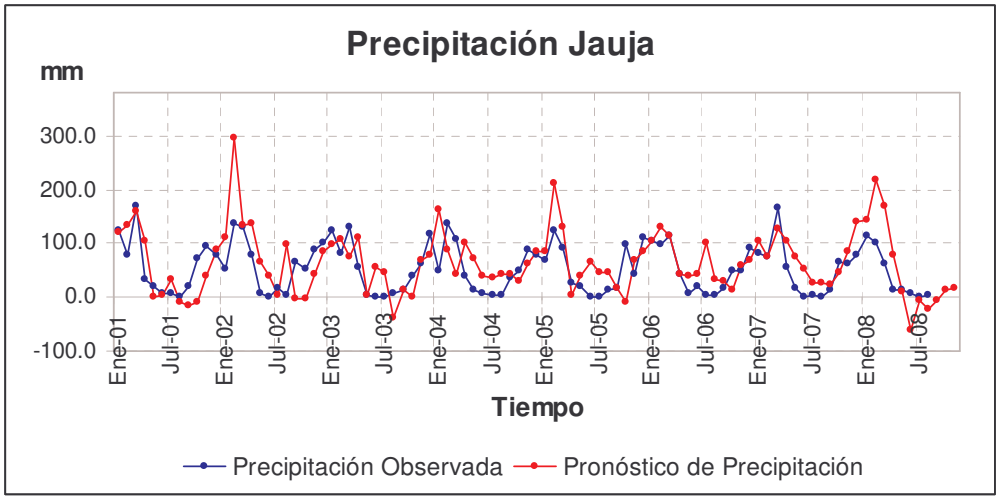
En la Figura 4.12 se muestra la gráfica de la serie observada y la serie pronóstico. Se puede ver claramente la presencia de errores.

Cuadro 4.35: Evaluación de pronósticos de precipitación de Jauja utilizando RNAB

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
2417.902	49.172	38.524	1299.085	13,178	0.610

FUENTE: Elaboración propia

Figura 4.12. Precipitación de Jauja. Valores observados y valores pronóstico con RNAB



FUENTE: Elaboración propia

Temperatura Mínima

El ECM es 7.038 °C^2 es alto, pero observando el RECM (2.653 °C) y el AMN (2.074 °C) se puede observar alrededor de 2 a 2.5 °C de error. Además, el BIAS muestra que en promedio se esta sobreestimando 0.78 °C. La correlación es baja, 0.559. (Cuadro 4.36)

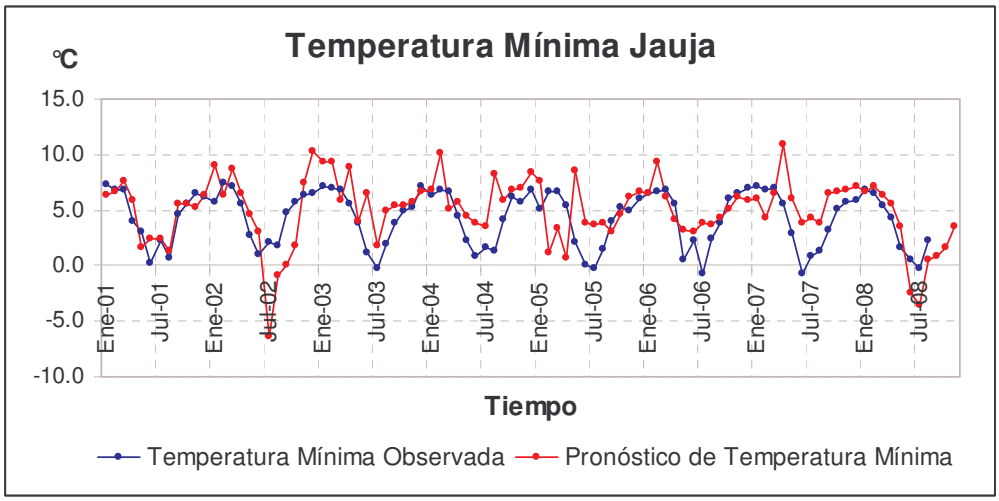
En la Figura 4.13 se observa la gráfica de ambas series, la serie observada y la serie pronóstico. Hay algunos puntos en donde se observa claramente errores altos presentes.

Cuadro 4.36: Evaluación de pronósticos de temperatura mínima de Jauja utilizando RNAB

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
7.038	2.653	2.074	0.708	0.776	0.559

FUENTE: Elaboración propia

Figura 4.13. Temperatura mínima de Jauja. Valores observados y valores pronóstico con RNAB



FUENTE: Elaboración propia

Temperatura Máxima

El ECM es 3.25 °C^2 es bajo, lo que indica que hay pocos errores. El RECM y el AMN muestran que hay un error de entre 1.4 a 2 °C. El BIAS muestra que en promedio hay una sobreestimación de 0.86 °C. Y la correlación en muy baja. (Cuadro 4.37)

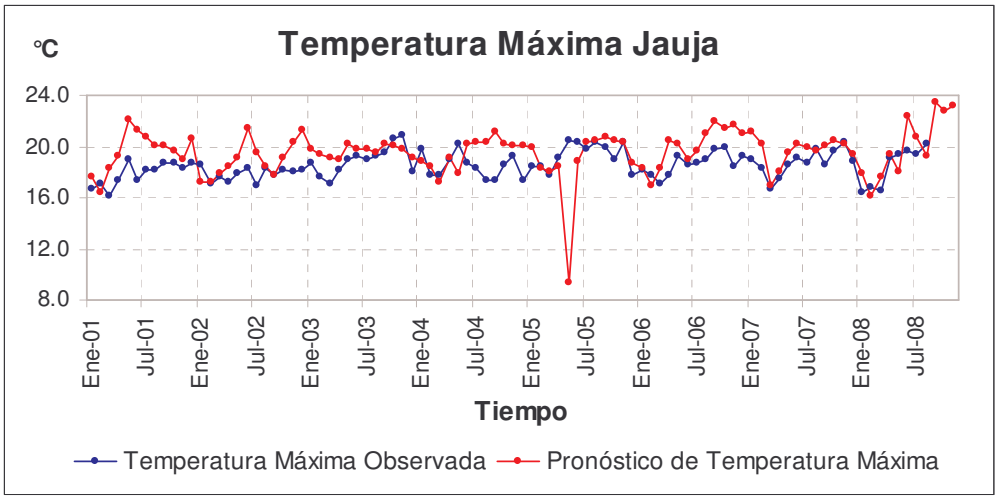
En la Figura 4.14 se muestra la gráfica de la series observada y la series pronóstico se observa que existe un pico en el año 2005 que corresponde a mayo, en donde se calcula un valor muy bajo para la temperatura máxima, que es 9.3 °C .

Cuadro 4.37: Evaluación de pronósticos de temperatura máxima de Jauja utilizando RNAB

ECM (°C^2)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR
3.827	1.956	1.398	0.076	0.859	0.245

FUENTE: Elaboración propia

Figura 4.14. Temperatura máxima de Jauja. Valores observados y valores pronóstico con RNAB



FUENTE: Elaboración propia

ESTACIÓN DE VIQUES

El ECM es bastante grande, tiene un valor de 5911.159 mm2, lo que indica que existen muchos errores en el pronóstico y no es aconsejable utilizar este tipo de red neuronal. El RECM y el AMN muestran que hay errores entre 64.4 a 76.9 mm. El BIAS proporciona un error promedio que indica que se subestima la serie en 48.4 mm. La correlación es baja. (Cuadro 4.38)

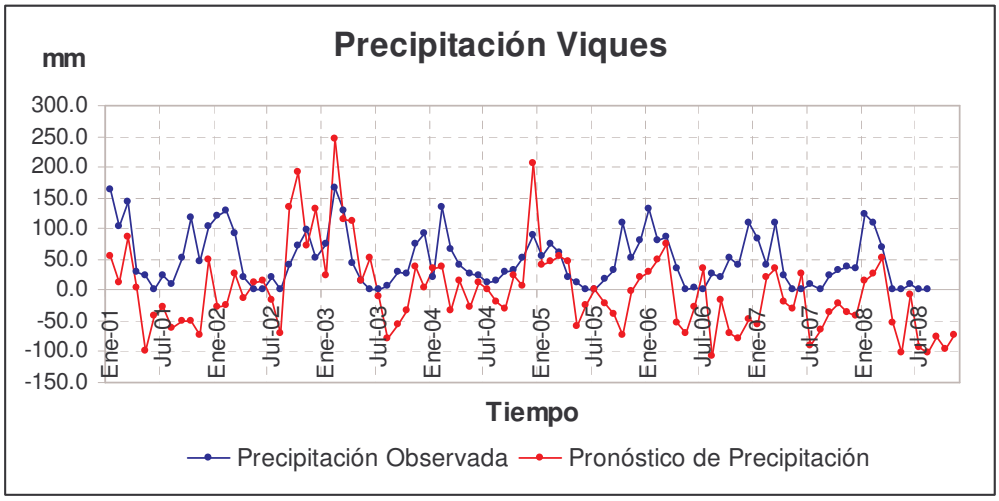
La Figura 4.15 muestra la gráfica de ambas series. Se puede observar con claridad que la serie pronóstico subestima a la serie observada, se ve un desfase, como ya se había notado con el BIAS.

Cuadro 4.38: Evaluación de pronósticos de precipitación de Viques utilizando RNAB

ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR
5911.159	76.884	64.405	7451.633	-48.421	0.468

FUENTE: Elaboración propia

Figura 4.15. Precipitación de Viques. Valores observados y valores pronóstico con RNAB



FUENTE: Elaboración propia

4.4. COMPARACIÓN ENTRE MARS Y RNAB

4.4.1. ESTACIÓN DE HUAYAO

PRECIPITACIÓN

Al comparar el ECM, RECM, AMN, EAMN, BIAS y la correlación de ambas técnicas, resulta que para precipitación en la estación de Huayao, la técnica con menor error y altamente correlacionada es MARS (ver Cuadro 4.39). Por lo que se escogería este modelo para próximos pronósticos. Para un detalle de los pronósticos hasta noviembre de 2008 ver Anexo 25.

Cuadro 4.39: Comparación de la precipitación entre MARS y RNAB en la estación de Huayao

TECNICA	ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR	Error Relativo de Predicción
MARS	2069.60	45.49	28.87	22.65	-10.03	0.579	1.04
RNAB	4280.60	65.43	49.13	114.84	39.29	0.575	2.15

FUENTE: Elaboración propia

TEMPERATURA MÍNIMA

En temperatura mínima se tiene que el ECM, RECM y AMN cuando se utiliza el MARS es menor que al usar las RNAB. Sin embargo, la correlación es alta para las estimaciones de ambas técnicas (Ver Cuadro 4.40). Entonces se puede afirmar que, en este caso, con el modelo MARS se obtiene menor error que con las RNAB. Ver pronósticos hasta noviembre de 2008 en el Anexo 26.

Cuadro 4.40: Comparación de la temperatura mínima entre MARS y RNAB en la estación de Huayao

TECNICA	ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR	Error Relativo de Predicción
MARS	2.77	1.67	1.06	0.07	0.13	0.816	0.45
RNAB	4.17	2.04	1.42	-0.13	0.81	0.732	0.67

FUENTE: Elaboración propia

TEMPERATURA MÁXIMA

En temperatura máxima, el lector puede observar que el ECM, RECM, AMN y el EAMN son mayores cuando se utiliza la técnica MARS que al utilizar las RNAB. En este caso se debería a esa observación anormal que se observó en la series pronóstico de temperatura máxima en Huayao con el modelo MARS. (Cuadro 4.41)

Aunque el valor de la correlación para las RNAB en bajo, se selecciona este modelo para los próximos pronósticos. Ver los pronósticos hasta noviembre de 2008 en el Anexo 27.

Cuadro 4.41: Comparación de la temperatura máxima entre MARS y RNAB en la estación de Huayao

TECNICA	ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR	Error Relativo de Predicción
MARS	6.95	2.55	1.04	0.05	0.71	0.163	7.34
RNAB	1.21	1.10	0.81	0.04	-0.17	0.408	1.41

FUENTE: Elaboración propia

4.4.2. ESTACIÓN DE JAUJA

PRECIPITACIÓN

El ECM, RECM, AMN, EAMN y BIAS son menores al utilizar el modelo MARS que al utilizar las RNAB. Además, la correlación del MARS en mayor que la correlación con RNAB. (Cuadro 4.42)

Los pronósticos con el MARS presentan menos error que con la RNAB. Se escogería este modelo para próximos pronósticos. Ver pronósticos hasta noviembre de 2008 en el Anexo 28.

Cuadro 4.42: Comparación de la precipitación entre MARS y RNAB en la estación de Jauja

TECNICA	ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR	Error Relativo de Predicción
MARS	1277.66	35.74	21.78	64.65	-3.00	0.741	0.62
RNAB	2417.90	49.17	38.52	1299.1	13.18	0.610	1.18

FUENTE: Elaboración propia

TEMPERATURA MÍNIMA

El ECM, RECM, AMN, EAMN son menores cuando se utiliza el modelo MARS que al utilizar las RNAB. La correlación es mucho más alta con el modelo MARS que con RNAB. (Cuadro 4.43)

Los pronósticos para temperatura mínima son mejores con el MARS. Para ver los pronósticos hasta noviembre de 2008 ver el Anexo 29.

Cuadro 4.43: Comparación de la temperatura mínima entre MARS y RNAB en la estación de Jauja

TECNICA	ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR	Error Relativo de Predicción
MARS	1.56	1.25	0.92	0.34	0.19	0.861	0.27
RNAB	7.04	2.65	2.07	0.71	0.78	0.559	1.20

FUENTE: Elaboración propia

TEMPERATURA MÁXIMA

El ECM, RECM, AMN, EAMN y BIAS son menores utilizando el modelo MARS que al utilizar las RNAB. La correlación es mayor con el MARS que con RNAB, pero a la vez esta correlación es baja. (Cuadro 4.44)

Los pronósticos para temperatura máxima de Jauja se calcularán a partir del modelo MARS, para los próximos meses. El lector puede observar los pronósticos hasta noviembre de 2008 en el Anexo 30.

Cuadro 4.44: Comparación de la temperatura máxima entre MARS y RNAB en la estación de Jauja

TECNICA	ECM ($°C^2$)	RECM (°C)	EAM (°C)	EAMN	BIAS (°C)	COR	Error Relativo de Predicción
MARS	1.74	1.32	1.09	0.06	-0.77	0.464	1.51
RNAB	3.83	1.96	1.40	0.08	0.86	0.245	3.32

FUENTE: Elaboración propia

4.4.3. ESTACIÓN DE VIQUES

El ECM, RECM, AMN, EAMN y BIAS son mucho menores con el MARS que con las RNAB. Se puede ver que en este caso la RNAB presenta un alto grado de error. La correlación al utilizar el MARS en mayor que con la RNAB. (Cuadro 4.45)

Los pronósticos se harán a partir de ahora con el modelo MARS, ya que éste presenta mejores estadísticos que las RNAB. Los pronósticos hasta noviembre de 2008 el lector los encontrará en el Anexo 31.

Cuadro 4.45: Comparación de la precipitación entre MARS y RNAB en la estación de Viques

TECNICA	ECM (mm^2)	RECM (mm)	EAM (mm)	EAMN	BIAS (mm)	COR	Error Relativo de Predicción
MARS	1248.179	35.33	24.17	376.93	-10.11	0.685	0.66
RNAB	5911.16	76.88	64.41	7451.6	-48.42	0.468	3.00

FUENTE: Elaboración propia

V. CONCLUSIONES

5.1. Fue necesaria la aplicación de métodos exploratorios para verificar la calidad de los datos. Se encontró que la variabilidad de la precipitación y temperaturas extremas de las estaciones Huayao y Jauja son bastante similares. Calculándose las desviaciones estándar de precipitación, 51.02 mm y 54.48 mm; con una media 62.4 mm y 60.6 mm respectivamente. Con respecto a las temperaturas las dos estaciones también son bastante similares. En el caso de la temperatura mínima, ésta tiene una media igual a 4.4 °C en Huayao y 4.38 °C en Jauja, con desviaciones estándar iguales a 2.52 °C y 2.42 °C, respectivamente. En el caso de temperatura máxima, la media en Huayao es 19.37 °C y la media en Jauja es 19.32 °C, con desviaciones estándar iguales a 1.2 °C y 1.25 °C, respectivamente. En conclusión, Huayao y Jauja presentan condiciones similares en cuanto a precipitación y temperaturas extremas.

5.2. Viques presentó mayor cantidad de valores extremos, mostrando que esta serie tiene bastantes problemas de calidad, puede estar presente gran cantidad de errores sistemáticos. La variabilidad de la precipitación en Viques es mayor que en las otras dos estaciones, con una desviación estándar de 65.88 mm y una media de 40.72 mm se piensa que Viques representa un clima diferente al de las otras dos estaciones analizadas.

5.3. En general, los pronósticos de cada una de las variables respuesta aplicando el modelo MARS son mejores que aplicando las RNAB, a excepción de la serie de temperatura máxima de Huayao en donde se aprecian mejores estadísticos utilizando las RNAB que el MARS. Sin embargo, este resultado se debe a que un valor de la serie predicha por el modelo MARS es un valor negativo fuera del rango esperado (se esperaría un valor entre 15 °C y 20 °C), y es por este valor que el ECM, la RECM, el EAM, el EAMN y el BIAS aumentan y la correlación disminuye. Al eliminar este valor dudoso (Cuadro 4.20) se observa una gran diferencia en los

estadísticos, y mejoran el pronóstico. Inclusive, sin este valor, al comparar la serie predicha con el MARS y la predicha con las RNAB, la serie estimada con el modelo MARS es mucho mejor que la estimada con las RNAB. Este valor inusual puede deberse a valores atípicos en las variables explicativas que el MARS no esta en la capacidad de asimilar.

5.4. Los pronósticos de precipitación utilizando el MARS en Huayao presentan errores entre 0.47 mm/mes y 206.9 mm/mes. Los pronósticos de temperatura mínima utilizando el MARS en Huayao presentan errores entre 0.1 °C/mes y 8.2 °C/mes. Los pronósticos de temperatura máxima utilizando las RNAB en Huayao presentan errores entre 0.003 °C/mes y 3.2 °C/mes. Los EAM para cada una de las variables son los siguientes: (a) precipitación es 28.9 mm, (b) temperatura mínima es 1.06 °C y (c) temperatura máxima es 0.81 °C, estos valores al compararse con las anomalías absolutas promedio que son: (a) precipitación es 21.4 mm, (b) temperatura mínima es 0.7 °C y (c) temperatura máxima es 0.7 °C. Los EAM y las anomalías absolutas promedio son similares, por lo que se dice que los EAM son aceptables.

5.5. Los pronósticos de precipitación utilizando el MARS en Jauja presentan errores entre 0.08 mm/mes y 206.7 mm/mes. Los pronósticos de temperatura mínima utilizando el MARS en Jauja presentan errores entre 0.01 °C/mes y 4.9 °C/mes. Los pronósticos de temperatura máxima utilizando el MARS en Jauja presentan errores entre 0.01 °C/mes y 3.4 °C/mes. Los EAM para cada una de las variables son los siguientes: (a) precipitación es 21.8 mm, (b) temperatura mínima es 0.9 °C y (c) temperatura máxima es 1.1 °C, se compararon con los valores de las anomalías absolutas promedio que son: (a) precipitación es 15.6 mm, (b) temperatura mínima es 0.6 °C y (c) temperatura máxima es 0.9 °C, estos valores son similares, por lo que los EAM son aceptables.

5.6. Los pronósticos de precipitación utilizando el MARS en Viques presentan errores entre 0.05 mm/mes y 111.7 mm/mes. El EAM es 24.2 mm, este se comparó con la anomalía absoluta promedio que es 19.5 mm, estos valores son similares varían en 5 mm aproximadamente, entonces se dice que son aceptables.

5.7. En la estación de Huayao en los meses de Septiembre y Octubre se espera que la precipitación sea menor que el promedio, 40.9 mm y 40.5 mm, respectivamente y en Noviembre la precipitación será mayor al promedio, 79.7 mm (precipitación promedio Septiembre es 44.9 mm, Octubre es 68.6 mm y Noviembre es 69.6 mm). En el caso de la temperatura mínima, Septiembre se prevé que sea igual al promedio (4.6 °C) y Octubre se espera que sea más cálido que el promedio, 6.2 °C (promedio Octubre 5.7 °C), y que Noviembre sea más frío, 5.8 °C (promedio Noviembre, 5.9 °C). Y en el caso de la temperatura máxima, se espera que en general los tres meses, Septiembre, Octubre y Noviembre sean más cálidos con temperatura de 20.3 °C, 20.6 °C y 21.6 °C, respectivamente (temperatura promedio 20 °C, 20.3 °C y 20.5 °C).

5.8. En la estación de Jauja se espera que la precipitación sea menor o igual al promedio para los tres meses, Septiembre, Octubre y Noviembre, 24.9 mm, 54 mm y 72.1 mm (promedios mensuales 30.1 mm, 56.9 mm y 72.1 mm). El pronóstico de temperatura mínima se espera que para los dos primeros meses sea más frío que el promedio, 4 °C y 4.1 °C (promedios respectivos, 4.2 °C y 5.4 °C), en cambio par el mes de Noviembre se espera un aumento de temperatura respecto al promedio, 6 °C, siendo el promedio 5.6 °C. En el caso de la temperatura máxima, se espera que la temperatura sea mayor o igual al promedio, 21 °C, 20 °C y 20.3 °C, respectivamente (promedios 19.9 °C, 20 °C y 20.1 °C).

5.9. En el caso de la precipitación de Viques, se espera que las precipitaciones observadas para los siguientes meses de Septiembre, Octubre y Noviembre sean mayores que el promedio, 92.5 °C, 47 °C y 62.2 °C, respectivamente (promedios 19.7 °C, 41.4 °C y 42.6 °C).

5.10. Finalmente, existe presencia de valores negativos en el pronóstico de las precipitaciones cuando se utiliza ambos métodos, pero no existe precipitación negativa en la naturaleza, por lo que al parecer ambas modelos tienen problemas con esta variable.

VI. RECOMENDACIONES

El estudio se enriquecería si se agregaran a las variables explicativas datos de las variables relacionadas a la zona como: humedad del suelo, humedad atmosférica, radiación solar, entre otras. Además, en próximos estudios se podría agregar como variable explicativa a la temperatura promedio mensual para analizar la precipitación.

Se podría mejorar los pronósticos haciendo una revisión de calidad de las variables explicativas, ya que obtuvieron mediante información secundaria, asumiéndose que no tienen valores atípicos.

Se debería hacer el análisis con otro tipo de red neuronal, para comprobar si es que el MARS siempre tiene mejores pronósticos en comparación con las redes neuronales artificiales.

Para próximos estudios podría aplicarse técnicas de imputación de datos, para obtener la mayor cantidad de información de las variables respuesta estudiadas.

Se debe utilizar mayor cantidad de estaciones, para enriquecer las pronósticos locales en la cuenca. Además de utilizar interpolación espacial para cubrir las zonas en donde no existe estaciones.

VII. BIBLIOGRAFÍA

[1] Acuña, E. *Regresión no paramétrica*, Capítulo 9 [25 de setiembre de 2008]. Disponible en < http://sigma.univalle.edu.co/index_archivos/Modelos/regresion%20no%20parametrica.pdf>

[2] Barrientos, A.F., Olaya, J. y González, V.M. 2007. *Un modelo spline para el pronóstico de la demanda de energía eléctrica*. Revista Colombiana de Estadística, Volumen 30 No. 2. pp. 187-202. Diciembre 2007

[3] Breiman, L., Friedman, J., Ilshen, R.A., y Stone, C.J. 1984. *Classification and Regression Trees*. California, USA: Wadsworth, Inc.

[4] Chatfield, C. Collins, A.J., 1980. *Introduction to multivariate analysis*. Chapman and Hall.Londres.

[5] Colaboradores de Wikipedia. *Multicolinealidad* [en línea]. Wikipedia, La enciclopedia libre, 2008 [20 de octubre de 2008: 12 de junio del 2008]. Disponible en <http://es.wikipedia.org/w/index.php?title=Multicolinealidad&oldid=18100106>.

[6] Craven and Wahba, G. 1979. *Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the methods of generalized cross-validation*. *Numerische Mathematik*, 31: pp. 377--403.

[7] De Boor, C. 1978. *A Practical Guide to Splines*. Springer.

[8] Dee, D.P. 1995. *A pragmatic approach to model validation*. Pp. 1-13, in: D.R. Lynch and A.M. Davies (eds). *Quantitative skill assessment of coastal ocean models*. Washington, DC: AGU.

[9] Doucet, P., & Sloep, P.B. 1992. *Mathematical modeling in the life sciences*. Chichester, Inglaterra: Ellis Horwood Limited.

[10] Fahrmeir, L., Tute, G., Hennevogl, W. 2001. *Multivariate Statistical Modeling Based on Generalized Liner Models*. Springer. 517 pp.

[11] Flores, E. 2004. *Validación de los pronósticos de Temperatura y Precipitación con el modelo operacional de Mesoescala MM5 para la costa norte del Perú.* Compendio de trabajos de investigación CNDG. Instituto Geofísico del Perú. V.5. pp 83 – 94.

[12] Fox, J. 2000.*Nonparametric Simple Regression: Smoothing Scatterplots*, SAGE. 83pp.

[13] Fox, J. 2000. *Multiple and Generalized Nonparametric Regression*, SAGE. 83 pp.

[14] Freedman, D., Corduras, A. & Cuffi, T. 1993. *Estadística,* 2da Edición. Traducción y edición: Bosch, A. Barcelona.

[15] Friedman, J.H. 1977. "*A Recursive Partitioning Decision Rule for Nonparametric Classifications*"; IEEE. Transactions on Computers, pp. 404-509.

[16] Friedman, J.H. 1988. *Fitting functions to noisy data in high dimensions. In computer Science and Statistics*: Proceeding of the 20th Symposium (E. Wegman, D. Gantzz, and J. Miller, eds.). Amer. Statist. Assoc., Washington, D.C., pp. 13-43.

[17] Friedman, J.H. 1991. *Multivariate Adaptive Regression Splines. The Annals of Statistics*, Vol. 19, No. 1 (Mar., 1991), pp. 1-67

[18] Friedman, J.H. & Silverman, B.W. 1989. *Flexible parsimonious smoothing and additive modeling*. Technometrics, vol. 31, pp. 3-39.

[19] Gujarati, D.N. 1999. *Econometría*. Segunda parte. Estados Unidos

[20] Hair, J.F., Anderson, R.E., Tatham, R.L., Blanck. W.C. 1999. *Análisis multivariante*. Quinta edición. Prentice may. Iberia. Madrid, España. 832 pp.

[21] Härdle W. 1990. *Applied nonparametric Regression*, Cambridge University Press.

[22] Hastie, T., Tibshirani, R. 1990. *Generalized Additive Models*. CRS Press. 335 pp.

[23] Hastie, T., Tibshirani, R., Friedman, J.H. 2001. *The Elements of Statistical learning: Data Mining, Inference and Prediction.* Springer. 533 pp.

[24] Hecht-Nielsen, R. 1988. *Neurocomputing: Picking the Human Brain.* IEEE Spectrum, 25, pp. 36-41, marzo de 1988. Reimpreso en el texto *Artificial Neural Networks: Theorical Concepts* (Vemuri, V. *ed*.), pp. 13-18, IEEE Computer Society Press Technology Series.

[25] Hilera y Martínez. 2000. *Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones*, Alfaomega, Santafé de Bogotá.. 391 pp

[26] Hiromi, K. 2002. *Variabilidad de la precipitación pluviométrica en Santa Catarina*. XII Congreso Brasilero de Meteorología. Brasil.

[27] Instituto Geofísico del Perú. 2005. *Atlas Climático de precipitación y temperatura del aire en la Cuenca del Río Mantaro, Volumen I*. CONAM. 107 pp.

[28] Instituto Geofísico del Perú. 2005. *Diagnóstico de la Cuenca del Mantaro bajo la visión del cambio climático, Volumen II*. CONAM. 90 pp.

[29] Instituto Geofísico del Perú. 2005. *Vulnerabilidad Actual y Futura ante el Cambio Climático y Medidas de Adaptación en la Cuenca del Río Mantaro, Volumen III*. CONAM. 104 pp.

[30] Ihaka R. & Gentleman R. 1996. *R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics* 5: pp. 299–314.

[31] ISA-UMH*, Introducción a las Redes Neuronales*.- ISA-UMH © T-98-012V1.0, < http://www.isa.umh.es/doct/rnia/Redes-1.PDF>, [Consulta 8 de enero de 2009].

[32] Kohonen, T. 1988. *An Introduction to Neural Computing.* Neural Networks, Vol.1, pp. 3-16.

[33] Latínez, K. 2008. *Pronóstico de Precipitación y Temperaturas Extremas del Aire con Meses de Anticipación Usando el Modelo MARS*. IGP – Instituto Geofísico del Perú. 13 pp.

[34] Law, A.M. & Kelton W.D. 2000. *Simulation Modeling and analysis*. 3rd edition, McGraw Hill, Nueva York.

[35] LeBlanc, M. 1995. *An Adaptive Expansion Method for Regression*. Statistica Sinica 5, marzo de 1995. pp 737 – 748. [en línea] STATISTICAL SCIENCE < http://www3.stat.sinica.edu.tw/statistica/oldpdf/A5n222.pdf> [noviembre de 2008]

[36] Mardia, K.V., Kent, J.T. and Bibby, J.M. 1979. Multivariate Analysis. *Probability and mathematical statistics*. Academic Press, London, UK,

[37] Marsh, L. y Cormier, D.R. 2001. *Spline Regression Models*. SAGE, p 69.

[38] McCulloch, W.S. y Pitts, W.A. 1943. *A Logical Calculus of the Ideas Immanent in Nervous Activity.* Boulletin of Mathematics and biophysics, 5, pp. 115-133.

[39] McCullagh, P., & Nelder, J.A. 1989. *Generalized linear models*. 2nd ed. Londres: Chapman & Hall.

[40] Medeiros, C. 2002. *Series Temporais: Estimaçâo no dominio de Freqüência*. Departamento de Economía Pontificia Universidad Católica do Rio de Janeiro PUC, Rio – Brasil.

[41] Menacho Casimiro, E.E. 1993. *Pronóstico de Temperatura y precipitación y elaboración de calendarios agrícolas para Huayao – Huancayo*, UNALM, Lima.

[42] Menacho Casimiro, E.E. 2007. *Pronóstico de precipitación para el departamento de Puno*, UNALM, Lima.

[43] Morettin, P. 1985. *Series Temporais*. Sao Paolo, Brasil

[44] Núñez, V., Tussell, A. 2005. *Regresión y Análisis de Variancia*.

[45] Pielke, R.A., 1984: *Mesoscale Meteorological Modeling*. Orlando, Academic Press, 611 pp.

[46] R Development Core Team 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[47] Reinsch, C.H. 1967. *Smoothing by Spline Functions*. *Numerische Mathematik,* 10: pp. 177–183

[48] Render, B., Stair, R.M., Hanna, M.E., 2006. "*Quantitative analysis of management"*. Pearson.

[49] Romero, R., Zúnica, L.R., 2005. "*Métodos Estadísticos en Ingeniería"*. Universidad Politécnica de Valencia, Departamento de Estadística e Informática.

[50] Rumelhart, D. E., Hinton, G. E. y Williams, R. J. 1986. *Learning representations by back-propagation errors.* Nature, 323, pp. 533-536. Reimpreso en el texto *Parallel Distributed Processing: Explorations in the microstructure of cognition,* Vol.1, Foundations. (Rumelhart, D. y McClelland, J. ed.), pp. 318-362, MIT Press, 1986. Reimpreso en el texto *Neurocomputing* (Anderson, J. y Rosenfeld, E. Ed.), MIT Press, 1988.

[51] Ruppert, D. And Carroll, R.J. 2000. *Spatially-adaptive penalties for spline fitting,* Austral & New Zealand J. Statist. 42, pp 205 – 223.

[52] S original by Trevor Hastie & Robert Tibshirani. 2006. R port by Friedrich Leisch, Kurt Hornik and Brian D. Ripley. mda: Mixture and flexible discriminant analysis. R package version 0.3-2.

[53] Salvador Figueras, M y Gargallo, P. 2003:. *Análisis Exploratorio de Datos*, [en línea] 5campus.com, Estadística <http://www.5campus.com/leccion/aed> [junio 2008]

[54] Schumaker, L.L. 1980. *Spline Functions: Basic Theory*, Wiley–Interscience, 553 pp. (Reprinted by Krieger, Malabar, Fla., 1993)

[55] Schumaker, L.L. 1993. *On Shape Preserving Quadratic Spline Interpolation*. SIAM J. Numer. Anal. 20: pp 854-864

[56] Serinaldi, F. 2005. *Multivariate linear parametric models applied to daily rainfall time series*. Department of Hydraulics, Transportation and Highways. University of Rome "La Sapienza", Roma. Italia.

[57] Silva, C., Alvarado, S., Montaño, R. y Pérez, P. 2003. Modelamiento de la contaminación atmosférica por partículas: Comparación de cuatro procedimientos en Santiago, Chile. MIOMETRICA XIII, pp 113 - 127

[58] Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. CRC Press. 175 pp.

[59] Stauffer, D. R., and N. L. Seaman, 1990: *Use of four-dimensional data assimilation in a limited-area mesoscale model*. Part I: Experiments with synoptic-scale data. Mon. Wea. Rev.*,* **118,** pp. 1250–1277

[60] Thisted, R.A. 1988. *Elements of Statistical Computing: Numerical Computation*. CRC Press. 448 pp.

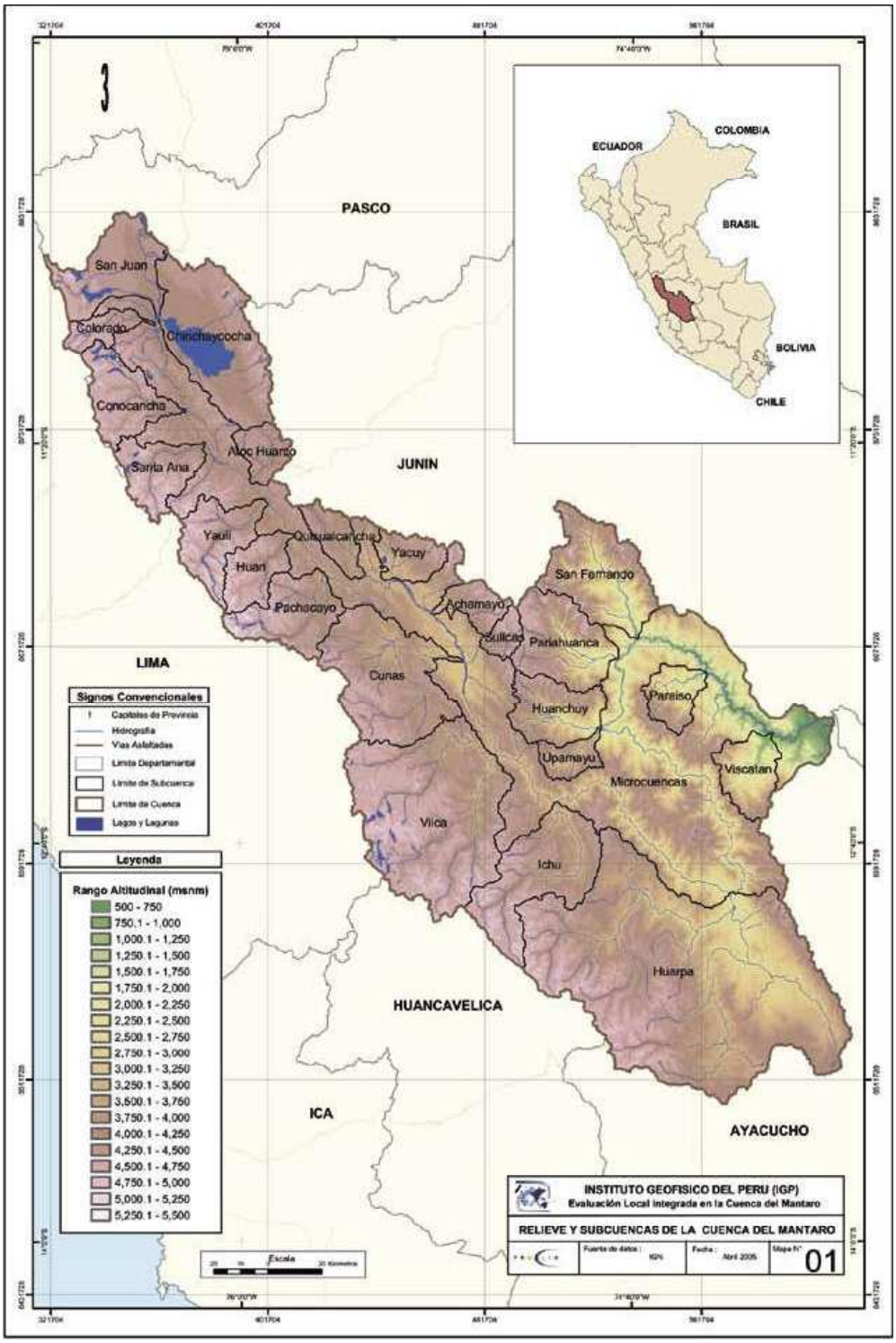
[61] Wahba, G. (1990), *Spline Models for Observational Data,* CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59. Philadelphia: SIAM.

[62] Witten I.H. and Frank E. 2005. "*Data Mining: Practical machine learning tools and techniques*", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[63] Zapata Robles, U. 2004. *Análisis de Regresión Semiparamétrico: caso lineal*. UNALM, Lima.

VIII. ANEXOS

Anexo 1: Relieve y subcuencas de la cuenca del río Mantaro

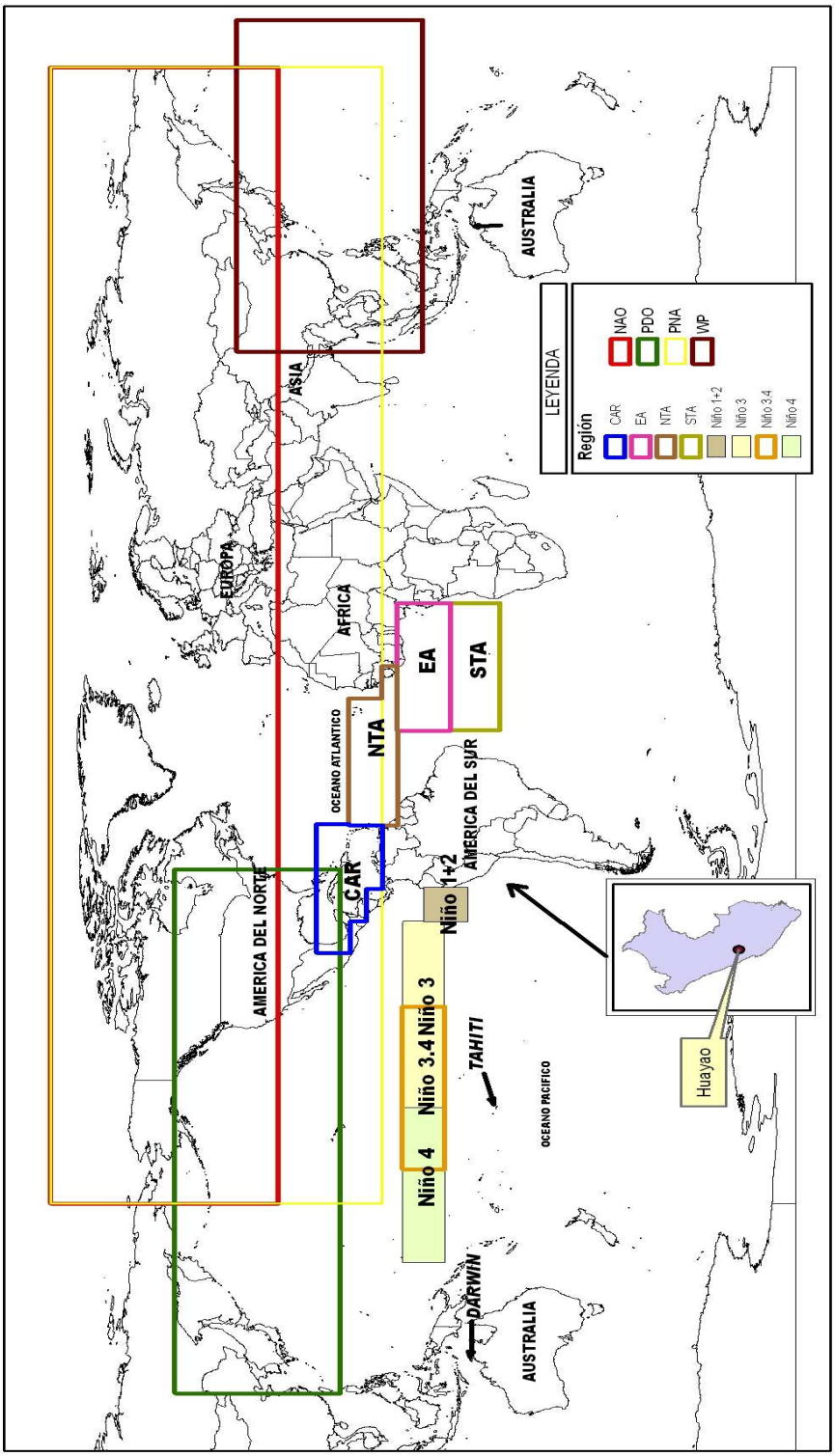


FUENTE: Instituto Geofísico del Perú [27]

Anexo 2: Variables globales. nombre, abreviatura y uso			
Variable	**Abreviatura**	**Se utiliza en el modelado de**	
		pp	**tm y tx**
Oscilación Decadal del Pacífico	PDO	Sí	Sí
Oscilación del Atlántico Norte	NAO	Sí	Sí
Patrón del Atlántico Este	EA	Sí	Sí
Índice del Pacífico Oeste	WP	Sí	Sí
Índice del Pacífico Norteamericano	PNA	Sí	Sí
Patrón del Atlántico Este / Oeste de Rusia	EA/WR	Sí	No
Patrón de Escandinavia	SCA	Sí	No
Índice de TSM en el Extremo Oriental del Pacífico (Región Niño 1+2)	N12	Sí	Sí
Índice de TSM Oriental del Pacífico (Región Niño 3)	N3	Sí	Sí
Índice de TSM Occidental del Pacífico (Región Niño 4)	N4	Sí	Sí
Índice de TSM en el Pacífico Central Oriental (Región Niño 3.4)	N34	Sí	Sí
Presión a nivel del mar en Darwin	D	Sí	Sí
Presión a nivel del mar en Tahití	T	Sí	Sí
Índice de TSM en el Caribe	CAR	Sí	Sí
Índice del Atlántico Tropical Norte	TNA	Sí	Sí
Índice del Atlántico Tropical Sur	TSA	Sí	Sí
Índice de Oscilación del Sur	SOI	Sí	Sí

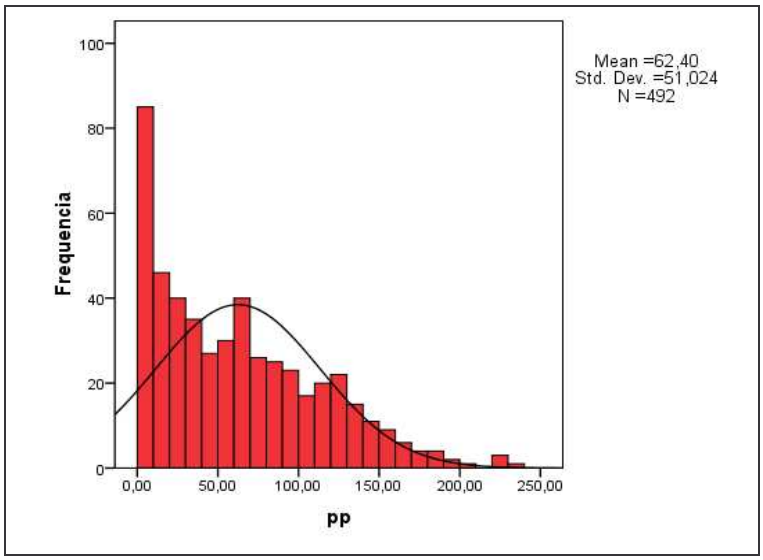
FUENTE: Elaboración propia

Anexo 3. Ubicación geográfica de los índices utilizados en el análisis de los datos



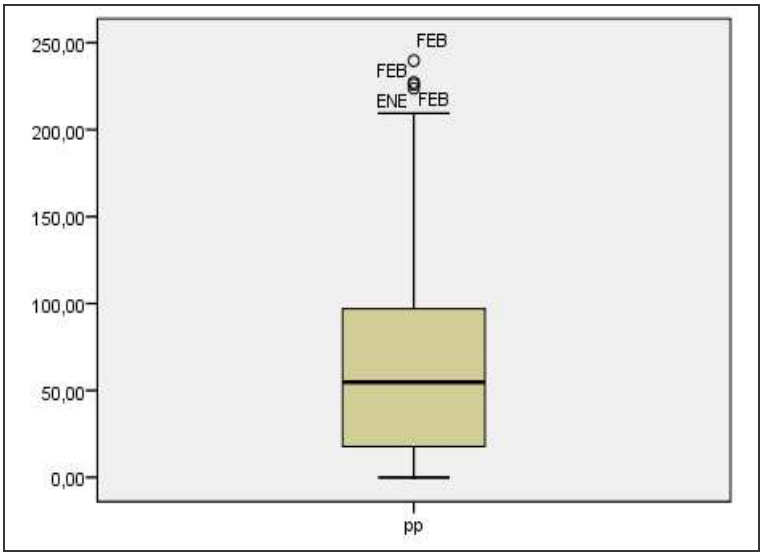
Fuente: Adaptado de *Change Master Directory*: //gcmd.nasa.gov

Anexo 4. Histograma de precipitación de Huayao



FUENTE: Elaboración propia

Anexo 5. Diagrama de cajas de precipitación de Huayao



FUENTE: Elaboración propia

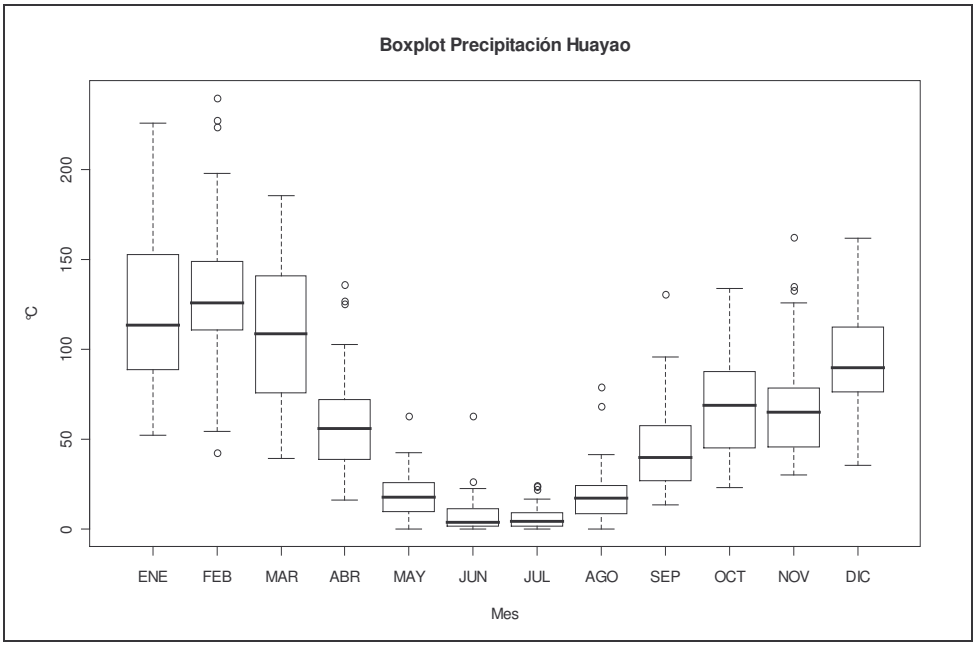
Anexo 6: Resumen de precipitación de Huayao por meses sin eliminar valores atípicos extremos

Resumen Descriptivo Precipitación Huayao

	N	Mínimo	Máximo	Media	Std. Error of Mean	Mediana	Std. Deviation	Curtosis	Asimetría
ene	41	51,90	226,00	119,4707	6,67996	113,3000	42,77262	-,125	,635
feb	41	42,50	239,70	130,5537	7,01829	125,9000	44,93898	,373	,459
mar	41	39,10	185,40	110,5634	5,77159	108,7000	36,95619	-,798	,124
abr	41	16,10	135,70	59,4683	4,51960	55,8000	28,93955	,584	,958
may	41	,00	62,80	19,4293	2,12862	17,8000	13,62984	1,130	,923
jun	41	,00	62,90	7,5634	1,76564	3,6000	11,30559	13,929	3,260
jul	41	,00	23,90	6,1902	,98157	4,3000	6,28509	1,884	1,426
ago	41	,00	79,10	19,7195	2,53530	17,2000	16,23386	4,793	1,848
sep	41	13,20	130,70	44,9390	3,99142	39,9000	25,55754	1,973	1,273
oct	41	22,90	133,60	68,5659	4,54657	68,5000	29,11224	-,459	,452
nov	41	29,80	162,30	69,5512	4,79526	64,8000	30,70465	1,216	1,204
dic	41	35,10	161,50	92,8122	4,75470	89,9000	30,44491	-,436	,130

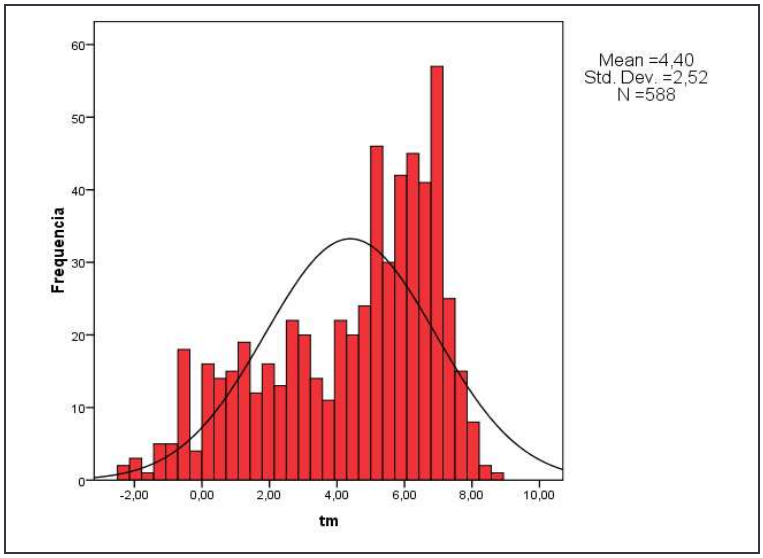
FUENTE: Elaboración propia

Anexo 7. Diagrama de cajas de precipitación de Huayao por meses



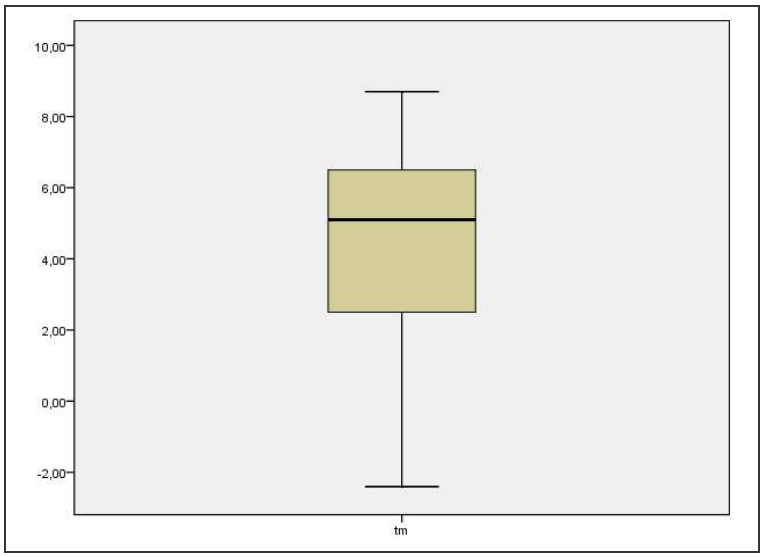
FUENTE: Elaboración propia

Anexo 8. Histograma de temperatura mínima de Huayao



FUENTE: Elaboración propia

Anexo 9. Diagrama de cajas de temperatura mínima de Huayao



FUENTE: Elaboración propia

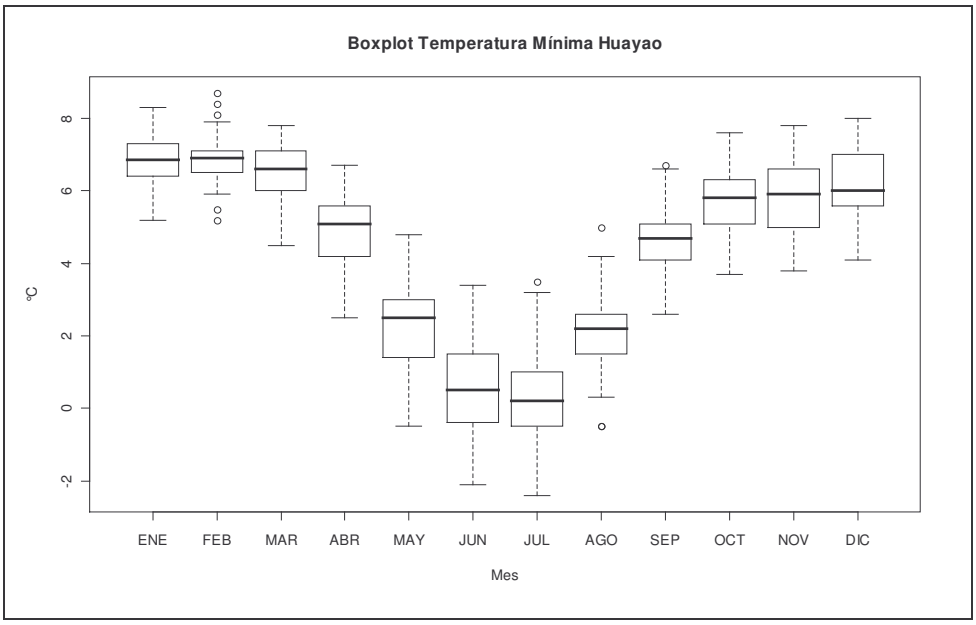
Anexo 10: Resumen de temperatura mínima de Huayao por meses

Resumen Descriptivo Temp. Mínima Huayao

	N	Mínimo	Máximo	Media	Std. Error of Mean	Mediana	Std. Deviation	Curtosis	Asimetría
ene	49	5,20	8,30	6,8245	,10439	6,8000	,73073	-,030	-,076
feb	49	5,20	8,70	6,8857	,09588	6,9000	,67113	,998	,216
mar	49	4,50	7,80	6,4755	,11095	6,6000	,77662	-,130	-,467
abr	49	2,50	6,70	4,8980	,13824	5,1000	,96771	-,136	-,436
may	49	-,50	4,80	2,3510	,16665	2,5000	1,16657	,042	-,194
jun	49	-2,10	3,40	,5816	,18882	,5000	1,32172	-,693	,181
jul	49	-2,40	3,50	,3286	,18447	,2000	1,29132	,293	,174
ago	49	-,50	5,00	2,0837	,15815	2,2000	1,10705	,523	-,041
sep	49	2,60	6,70	4,6000	,13718	4,7000	,96025	-,176	,089
oct	49	3,70	7,60	5,7286	,12966	5,8000	,90761	-,581	,105
nov	49	3,80	7,80	5,8551	,13976	5,9000	,97832	-,767	-,120
dic	49	4,10	8,00	6,1408	,12613	6,0000	,88292	-,312	-,235

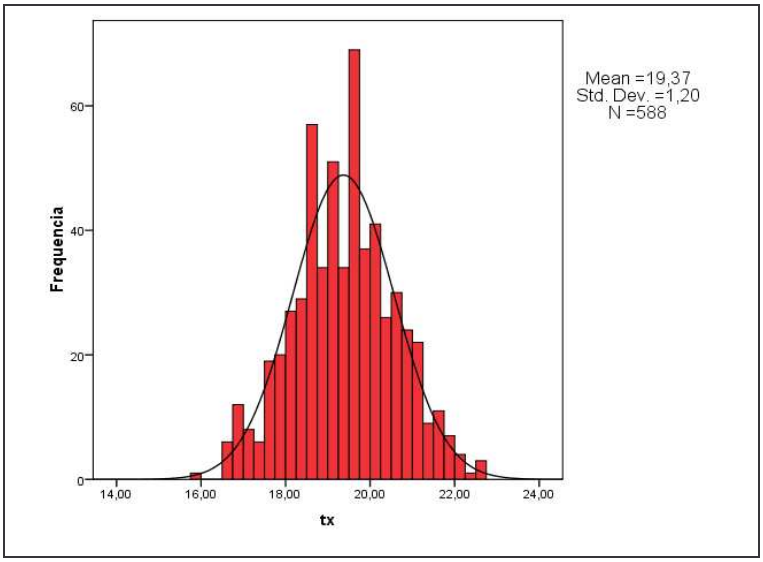
FUENTE: Elaboración propia

Anexo 11. Diagrama de cajas de temperatura mínima de Huayao por meses



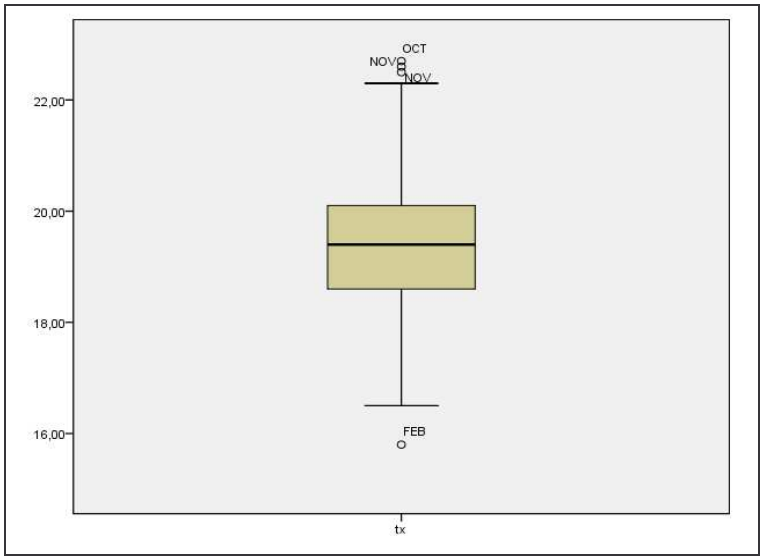
FUENTE: Elaboración propia

Anexo 12. Histograma de temperatura máxima de Huayao



FUENTE: Elaboración propia

Anexo 13. Diagrama de cajas de temperatura máxima de Huayao



FUENTE: Elaboración propia

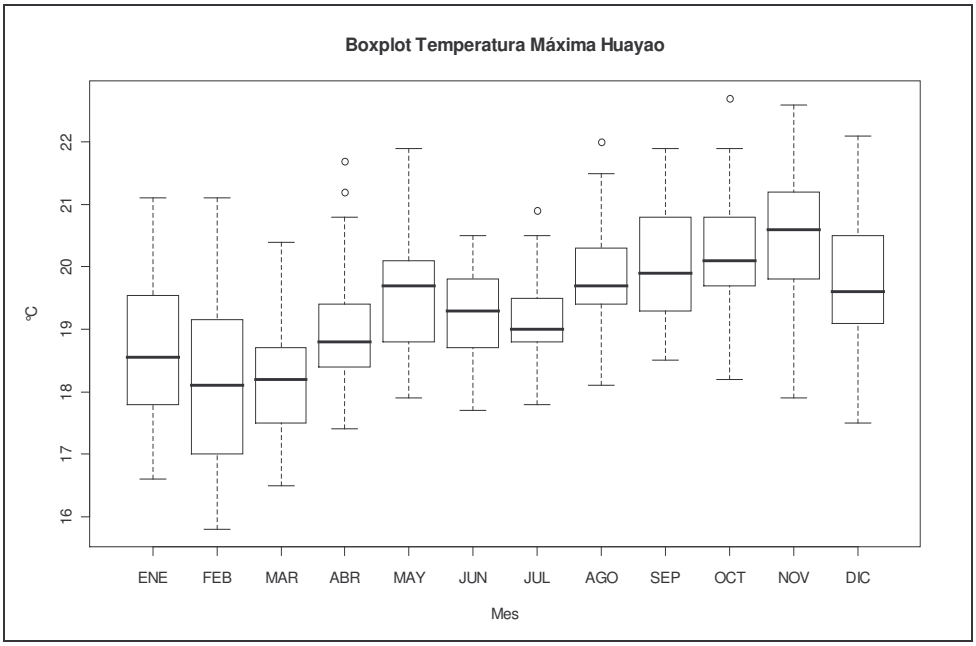
Anexo 14: Resumen de temperatura máxima de Huayao por meses

Resumen Descriptivo Temp. Máxima Huayao

	N	Mínimo	Máximo	Media	Std. Error of Mean	Mediana	Std. Deviation	Curtosis	Asimetría
ene	49	16,60	21,10	18,6796	,16017	18,5000	1,12119	-,769	,409
feb	49	15,80	21,10	18,1592	,17800	18,1000	1,24598	-,464	,504
mar	49	16,50	20,40	18,2143	,13988	18,2000	,97916	-,427	,351
abr	49	17,40	21,70	19,0531	,13351	18,8000	,93455	,645	,991
may	49	17,90	21,90	19,5429	,13442	19,7000	,94097	,051	,315
jun	49	17,70	20,50	19,2449	,09606	19,3000	,67239	-,872	,041
jul	49	17,80	20,90	19,1510	,09332	19,0000	,65324	,387	,418
ago	49	18,10	22,00	19,8265	,11398	19,7000	,79786	,392	,501
sep	49	18,50	21,90	19,9878	,12806	19,9000	,89644	-,882	,263
oct	49	18,20	22,70	20,3020	,13545	20,1000	,94813	-,114	,173
nov	49	17,90	22,60	20,4694	,15447	20,6000	1,08132	-,093	-,290
dic	49	17,50	22,10	19,7816	,15928	19,6000	1,11499	-,369	,138

FUENTE: Elaboración propia

Anexo 15. Diagrama de cajas de temperatura máxima de Huayao por meses



FUENTE: Elaboración propia

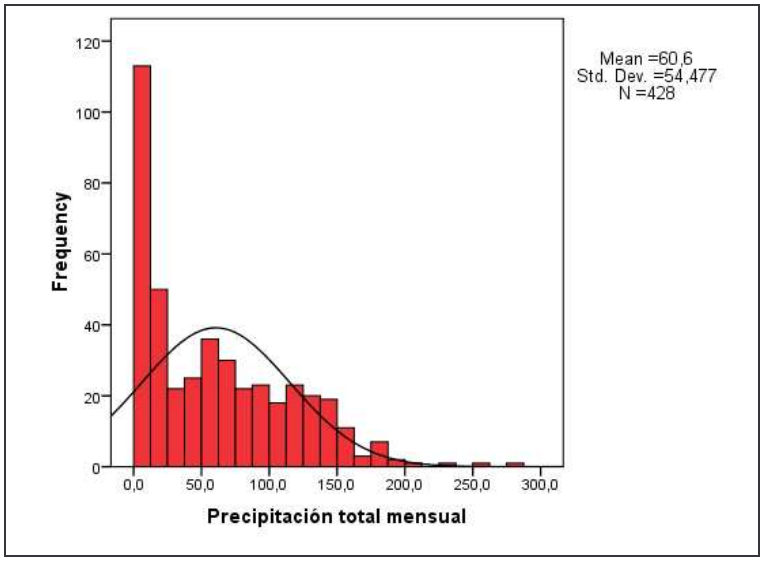
Anexo 16: Resumen de casos de precipitación de Jauja por meses

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
ene	37	90,2%	4	9,8%	41	100,0%
feb	37	90,2%	4	9,8%	41	100,0%
mar	37	90,2%	4	9,8%	41	100,0%
abr	37	90,2%	4	9,8%	41	100,0%
may	33	80,5%	8	19,5%	41	100,0%
jun	35	85,4%	6	14,6%	41	100,0%
jul	32	78,0%	9	22,0%	41	100,0%
ago	33	80,5%	8	19,5%	41	100,0%
sep	36	87,8%	5	12,2%	41	100,0%
oct	37	90,2%	4	9,8%	41	100,0%
nov	37	90,2%	4	9,8%	41	100,0%
dic	37	90,2%	4	9,8%	41	100,0%

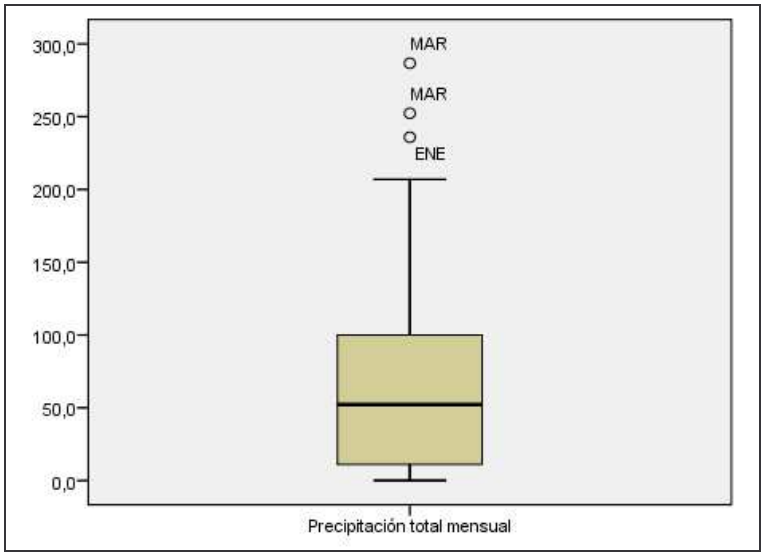
FUENTE: Elaboración propia

Anexo 17. Histograma de precipitación de Jauja



FUENTE: Elaboración propia

Anexo 18. Diagrama de cajas de precipitación de Jauja



FUENTE: Elaboración propia

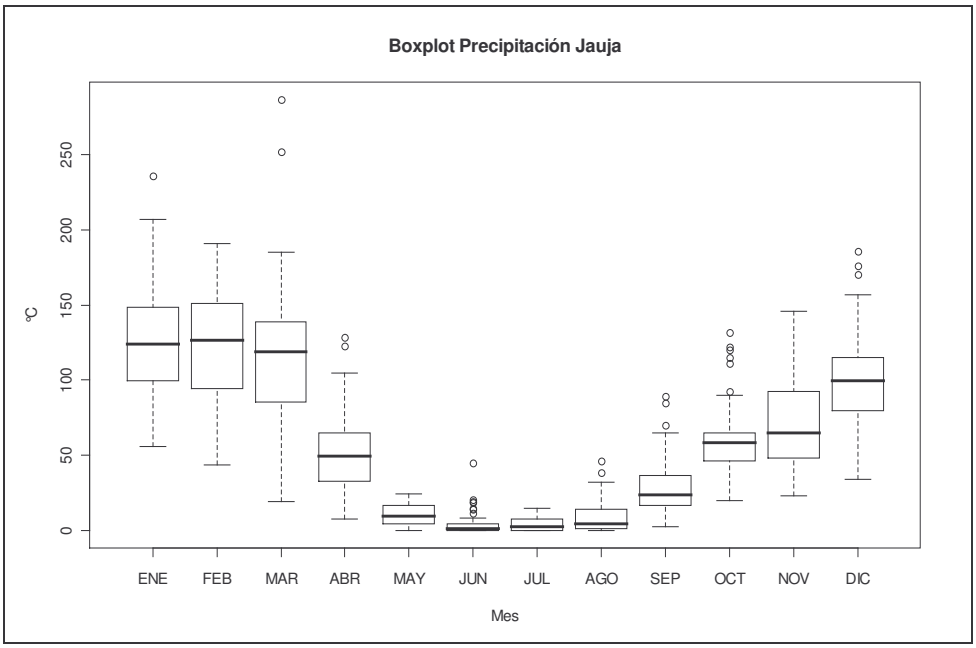
Anexo 19: Resumen de precipitación de Jauja por meses

Resumen Descriptivo Precipitación Jauja

	N	Mínimo	Máximo	Media	Std. Error of Mean	Mediana	Std. Deviation	Curtosis	Asimetría
ene	37	56,10	235,90	125,2946	6,94418	123,8000	42,23982	,143	,470
feb	37	43,80	190,90	119,5892	6,35505	126,4000	38,65628	-,655	-,278
mar	37	19,30	286,70	116,8649	8,67251	118,6000	52,75279	2,660	1,112
abr	37	7,80	128,80	50,5919	4,74343	49,2000	28,85317	1,005	,853
may	33	,00	24,00	9,8364	1,26369	9,6000	7,25934	-1,088	,203
jun	35	,00	44,80	5,2971	1,55303	1,2000	9,18786	9,444	2,793
jul	32	,00	14,80	4,3281	,84110	2,1000	4,75796	-,435	,926
ago	33	,00	46,20	9,2212	2,01209	4,6000	11,55857	3,029	1,801
sep	36	2,20	89,20	30,0722	3,53346	23,9500	21,20079	1,384	1,276
oct	37	19,70	131,60	62,4216	4,62573	58,5000	28,13720	,532	,970
nov	37	22,80	146,00	72,0622	5,45737	64,6000	33,19589	-,419	,692
dic	37	34,20	185,70	99,1568	5,94775	99,3000	36,17877	,306	,520

FUENTE: Elaboración propia

Anexo 20. Diagrama de cajas de precipitación de Jauja por meses



FUENTE: Elaboración propia

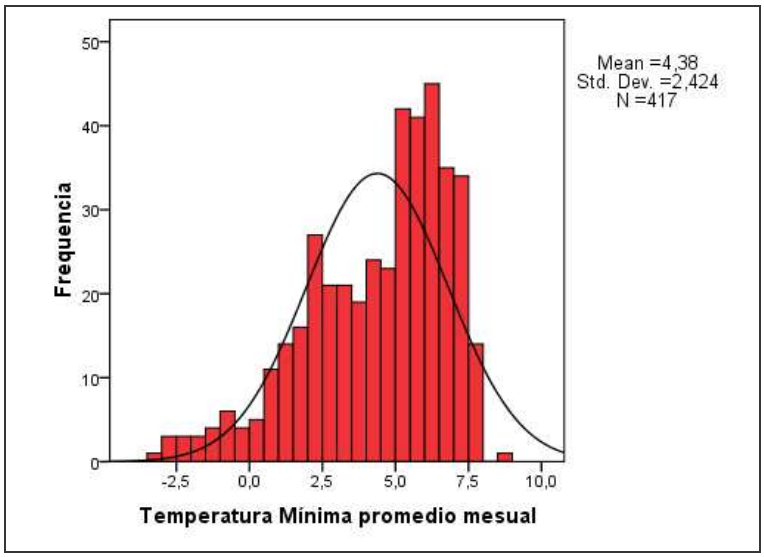
Anexo 21: Resumen de casos de Jauja por meses

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
ene	35	85,4%	6	14,6%	41	100,0%
feb	35	85,4%	6	14,6%	41	100,0%
mar	35	85,4%	6	14,6%	41	100,0%
abr	35	85,4%	6	14,6%	41	100,0%
may	34	82,9%	7	17,1%	41	100,0%
jun	36	87,8%	5	12,2%	41	100,0%
jul	34	82,9%	7	17,1%	41	100,0%
ago	35	85,4%	6	14,6%	41	100,0%
sep	34	82,9%	7	17,1%	41	100,0%
oct	35	85,4%	6	14,6%	41	100,0%
nov	35	85,4%	6	14,6%	41	100,0%
dic	34	82,9%	7	17,1%	41	100,0%

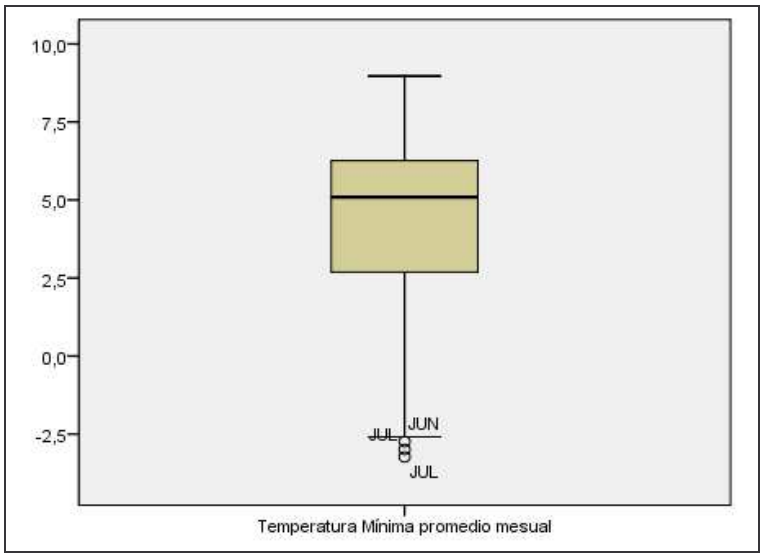
FUENTE: Elaboración propia

Anexo 22. Histograma de temperatura mínima de Jauja



FUENTE: Elaboración propia

Anexo 23. Diagrama de cajas de temperatura mínima de Jauja



FUENTE: Elaboración propia

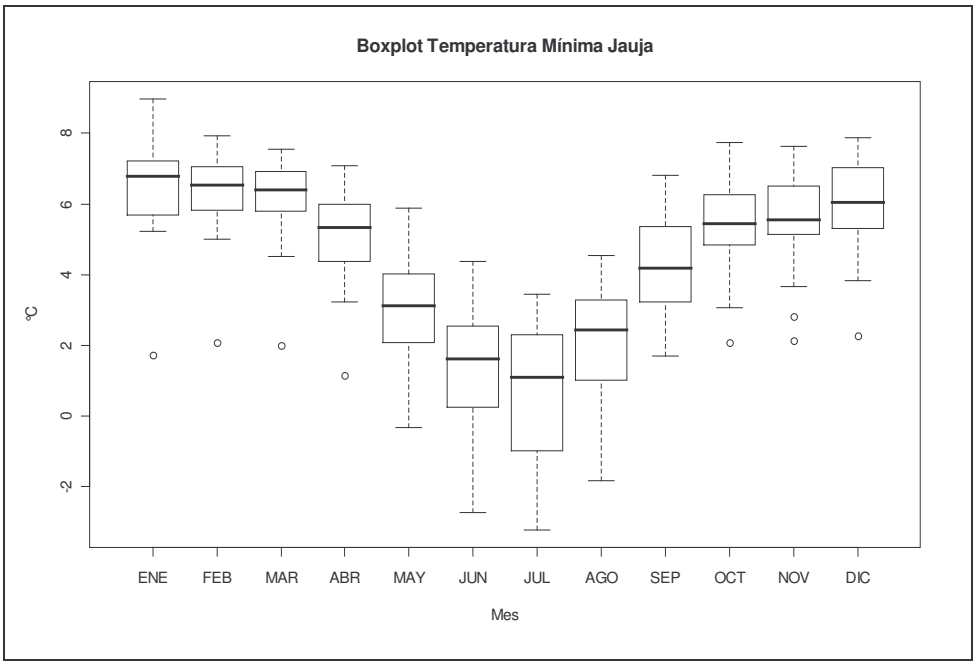
Anexo 24: Resumen de temperatura mínima de Jauja por meses

Resumen Descriptivo de Temperatura Mínima Jauja

	N	Mínimo	Máximo	Media	Std. Error of Mean	Mediana	Std. Deviation	Curtosis	Asimetría
ene	35	1,73	8,97	6,4437	,20402	6,7900	1,20701	5,908	-1,555
feb	35	2,06	7,92	6,4040	,18528	6,5400	1,09616	6,093	-1,786
mar	35	2,00	7,55	6,2586	,18068	6,4000	1,06891	6,313	-1,949
abr	35	1,13	7,08	5,1371	,19778	5,3300	1,17011	2,614	-1,139
may	34	-,34	5,89	2,9385	,23622	3,1200	1,37740	-,057	-,339
jun	36	-2,74	4,38	1,3072	,28715	1,6050	1,72289	-,231	-,565
jul	34	-3,23	3,45	,6509	,34166	1,1000	1,99220	-1,004	-,441
ago	35	-1,83	4,54	2,1337	,25898	2,4300	1,53216	,158	-,649
sep	34	1,70	6,82	4,2229	,23546	4,1850	1,37296	-,958	-,056
oct	35	2,08	7,75	5,4057	,20750	5,4500	1,22760	,616	-,645
nov	35	2,13	7,62	5,6146	,21150	5,5500	1,25123	,844	-,746
dic	34	2,26	7,87	6,0024	,20688	6,0500	1,20633	1,407	-,905

FUENTE: Elaboración propia

Anexo 25. Diagrama de cajas de temperatura mínima de Jauja por meses



FUENTE: Elaboración propia

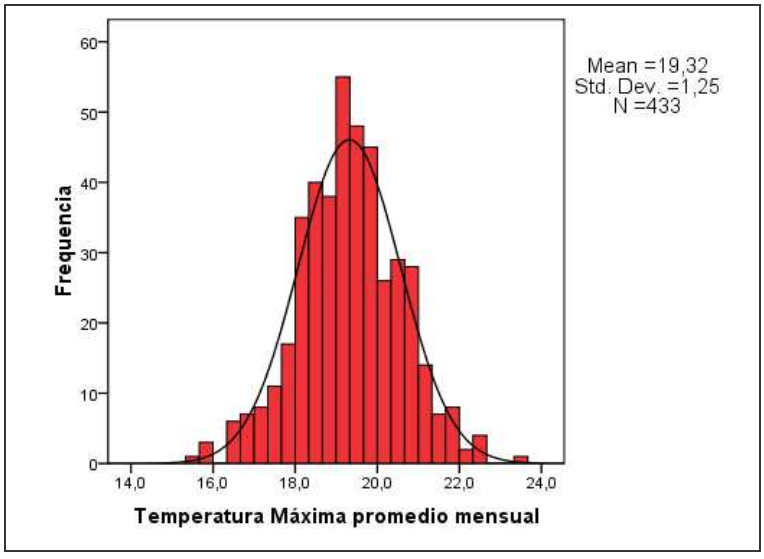
Anexo 26: Resumen de casos de temperatura máxima de Jauja por meses

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
ene	36	87,8%	5	12,2%	41	100,0%
feb	36	87,8%	5	12,2%	41	100,0%
mar	36	87,8%	5	12,2%	41	100,0%
abr	36	87,8%	5	12,2%	41	100,0%
may	35	85,4%	6	14,6%	41	100,0%
jun	37	90,2%	4	9,8%	41	100,0%
jul	35	85,4%	6	14,6%	41	100,0%
ago	36	87,8%	5	12,2%	41	100,0%
sep	36	87,8%	5	12,2%	41	100,0%
oct	37	90,2%	4	9,8%	41	100,0%
nov	37	90,2%	4	9,8%	41	100,0%
dic	36	87,8%	5	12,2%	41	100,0%

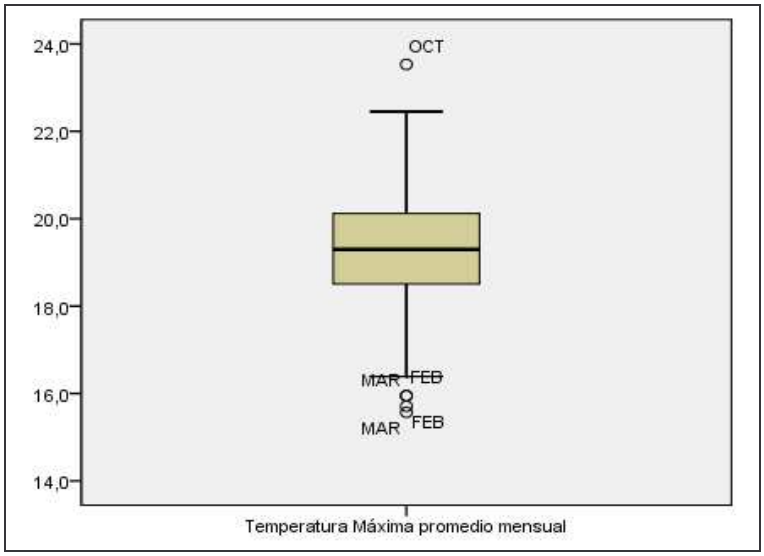
FUENTE: Elaboración propia

Anexo 27. Histograma de temperatura máxima de Jauja



FUENTE: Elaboración propia

Anexo 28. Diagrama de caja de temperatura máxima de Jauja



FUENTE: Elaboración propia

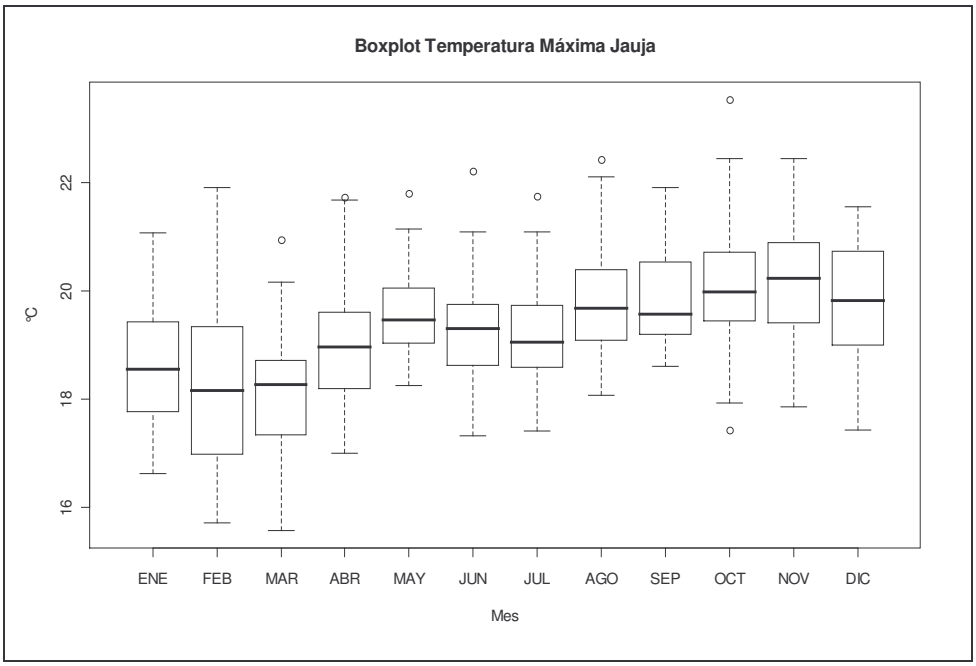
Anexo 29: Resumen de temperatura máxima de Jauja por meses

Resumen Descriptivo Temperatura Máxima Jauja por meses

	N	Mínimo	Máximo	Media	Std. Error of Mean	Mediana	Std. Deviation	Curtosis	Asimetría
ene	36	16,62	21,07	18,6564	,18790	18,5500	1,12739	-,554	,198
feb	36	15,71	21,91	18,2275	,22801	18,1600	1,36805	,020	,233
mar	36	15,57	20,94	18,1969	,20173	18,2600	1,21041	-,102	,140
abr	36	17,00	21,73	19,0497	,18303	18,9550	1,09818	,573	,846
may	35	18,25	21,79	19,5794	,14685	19,4600	,86879	-,107	,555
jun	37	17,32	22,21	19,2881	,15704	19,3000	,95523	1,411	,607
jul	35	17,40	21,74	19,2269	,16399	19,0400	,97015	,435	,830
ago	36	18,07	22,42	19,7506	,17220	19,6650	1,03319	,506	,701
sep	36	18,60	21,90	19,9036	,15382	19,5650	,92289	-,556	,727
oct	37	17,43	23,53	20,0162	,19884	19,9800	1,20950	1,258	,446
nov	37	17,85	22,44	20,1422	,17862	20,2300	1,08649	-,182	-,022
dic	36	17,42	21,54	19,7672	,17814	19,8100	1,06885	-,640	-,391

FUENTE: Elaboración propia

Anexo 30. Diagrama de caja de temperatura máxima de Jauja por meses



FUENTE: Elaboración propia

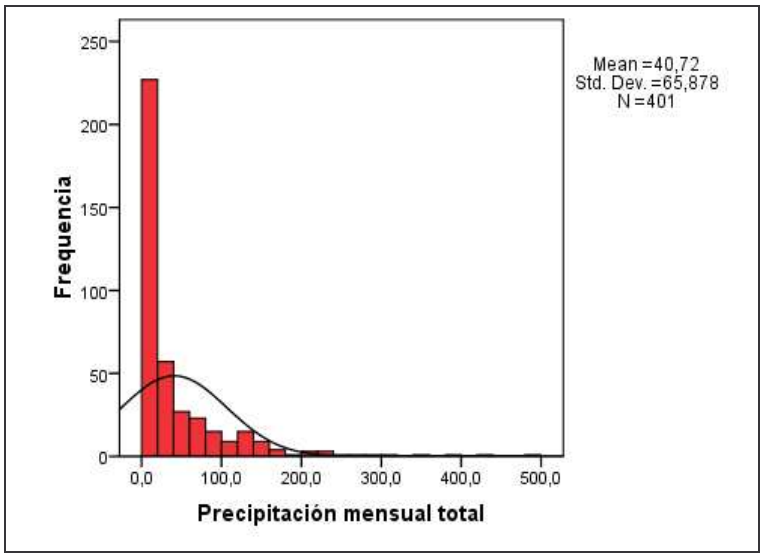
Anexo 31: Resumen de casos de precipitación de Viques

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
ene	33	86,8%	5	13,2%	38	100,0%
feb	33	86,8%	5	13,2%	38	100,0%
mar	34	89,5%	4	10,5%	38	100,0%
abr	34	89,5%	4	10,5%	38	100,0%
may	34	89,5%	4	10,5%	38	100,0%
jun	34	89,5%	4	10,5%	38	100,0%
jul	33	86,8%	5	13,2%	38	100,0%
ago	33	86,8%	5	13,2%	38	100,0%
sep	34	89,5%	4	10,5%	38	100,0%
oct	34	89,5%	4	10,5%	38	100,0%
nov	34	89,5%	4	10,5%	38	100,0%
dic	31	81,6%	7	18,4%	38	100,0%

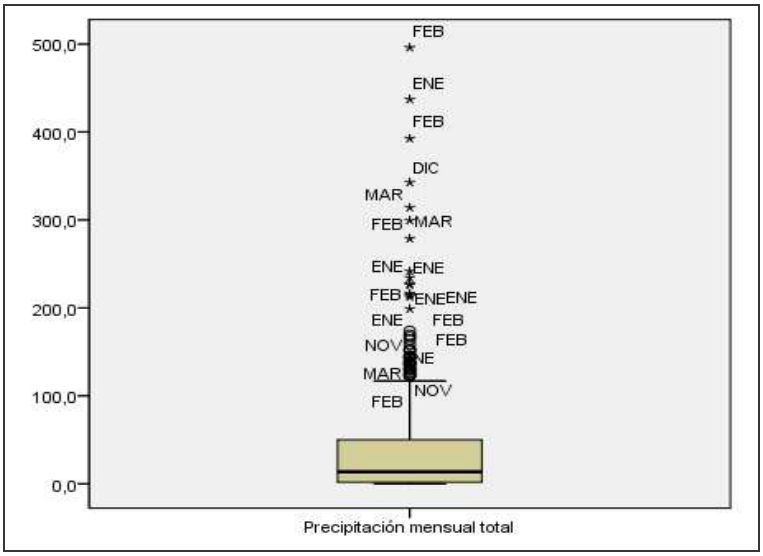
FUENTE: Elaboración propia

Anexo 32. Histograma de precipitación de Viques



FUENTE: Elaboración propia

Anexo 33. Diagrama de caja de precipitación de Viques



FUENTE: Elaboración propia

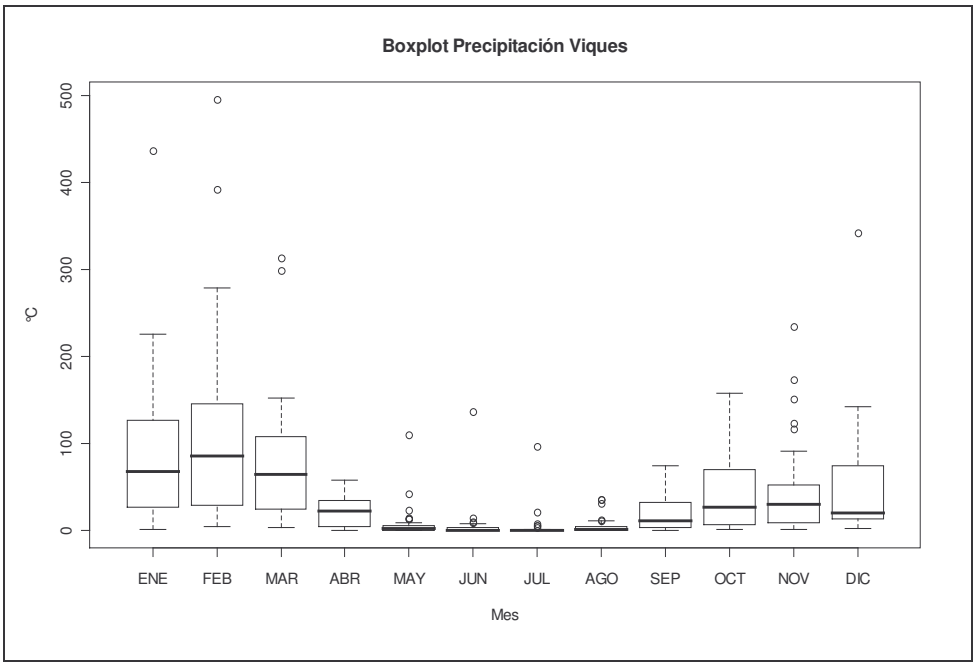
Anexo 34: Resumen descriptivo de precipitación de Viques por meses

Resumen Descriptivo Precipitación Viques

	N	Mínimo	Máximo	Media	Std. Error of Mean	Mediana	Std. Deviation	Curtosis	Asimetría
ene	33	1,00	437,00	90,8848	16,00589	67,4000	91,94686	5,013	1,895
feb	33	3,90	496,00	115,2576	19,81232	85,9000	113,81310	3,255	1,686
mar	34	3,30	314,00	77,0382	12,56969	63,8000	73,29328	4,101	1,873
abr	34	,00	57,90	21,3735	2,78813	22,3500	16,25748	-,792	,335
may	34	,00	110,20	7,7176	3,42377	1,7000	19,96385	22,292	4,502
jun	34	,00	136,70	6,1559	4,00028	,0000	23,32545	32,375	5,634
jul	33	,00	96,10	4,3333	2,95087	,0000	16,95145	29,117	5,296
ago	33	,00	35,10	5,4000	1,70815	1,0000	9,81259	4,826	2,368
sep	34	,10	74,00	19,7088	3,60972	10,6000	21,04809	,882	1,300
oct	34	,40	158,30	41,3853	7,17126	26,6000	41,81525	,724	1,153
nov	34	1,10	234,30	48,2618	9,38063	29,6500	54,69800	3,423	1,850
dic	31	1,90	342,60	53,8677	12,34328	19,9000	68,72449	9,771	2,735

FUENTE: Elaboración propia

Anexo 35. Diagrama de cajas de precipitación de Viques por meses



FUENTE: Elaboración propia

Anexo 36: Precipitación de Huayao utilizando MARS*

MODELO	GCV	RSQ	ECM	COR
Enero	1515.68	0.381	1104.52	0.617
	1515.68	0.381	1104.52	0.617
Febrero	1519.84	0.515	954.98	0.718
	1547.36	0.428	1127.61	0.654
Marzo	1188.03	0.599	534.47	0.774
	1188.03	0.599	534.47	0.774
Abril	585.49	0.678	263.70	0.823
	609.41	0.664	274.16	0.815
Mayo	162.00	0.521	86.74	0.722
	162.00	0.521	86.74	0.722
Junio	22.89	0.712	14.20	0.844
	22.89	0.712	14.20	0.844
Julio	30.03	0.649	13.51	0.806
	30.03	0.649	13.51	0.806
Agosto	104.30	0.683	54.82	0.827
	104.30	0.683	54.82	0.827
Setiembre	492.81	0.514	309.65	0.717
	492.81	0.514	309.65	0.717
Octubre	804.82	0.388	505.70	0.623
	776.73	0.410	488.06	0.640
Noviembre	966.34	0.000	919.78	0.235
	836.47	0.337	609.56	0.581
Diciembre	828.76	0.332	603.95	0.576
	828.76	0.332	603.95	0.576

*El valor superior corresponde a modelos de grado dos y el valor inferior a modelos de grado tres.

FUENTE: Elaboración propia

Anexo 37: Temperatura mínima de Huayao utilizando MARS*

MODELO	GCV	RSQ	ECM	COR
Enero	0.431	0.437	0.295	0.661
	0.319	0.680	0.168	0.824
Febrero	0.395	0.708	0.129	0.841
	0.337	0.704	0.130	0.839
Marzo	0.584	0.677	0.191	0.823
	0.584	0.677	0.191	0.823
Abril	0.604	0.655	0.317	0.809
	0.604	0.655	0.317	0.809
Mayo	1.327	0.142	1.144	0.376
	1.348	0.128	1.162	0.358
Junio	1.276	0.662	0.579	0.814
	1.276	0.662	0.579	0.814
Julio	1.355	0.624	0.615	0.790
	1.020	0.758	0.395	0.871
Agosto	1.064	0.318	0.819	0.563
	1.027	0.812	0.226	0.901
Setiembre	0.831	0.207	0.716	0.455
	0.719	0.522	0.432	0.722
Octubre	0.795	0.678	0.260	0.824
	0.793	0.554	0.360	0.744
Noviembre	0.824	0.601	0.374	0.775
	0.824	0.601	0.374	0.775
Diciembre	0.759	0.549	0.344	0.741
	0.759	0.549	0.344	0.741

*El valor superior corresponde a modelos de grado dos y el valor inferior a modelos de grado tres.

FUENTE: Elaboración propia

Anexo 38: Temperatura máxima de Huayao utilizando MARS*

MODELO	GCV	RSQ	ECM	COR
Enero	1.031	0.278	0.889	0.527
	1.031	0.278	0.889	0.527
Febrero	0.772	0.609	0.595	0.780
	0.827	0.531	0.713	0.729
Marzo	0.598	0.510	0.460	0.714
	0.667	0.627	0.350	0.792
Abril	0.362	0.711	0.248	0.843
	0.362	0.711	0.248	0.843
Mayo	0.549	0.512	0.423	0.716
	0.544	0.623	0.327	0.789
Junio	0.356	0.381	0.274	0.617
	0.272	0.800	0.089	0.894
Julio	0.345	0.777	0.093	0.881
	0.335	0.518	0.202	0.719
Agosto	0.484	0.470	0.331	0.685
	0.450	0.566	0.271	0.752
Setiembre	0.482	0.632	0.290	0.795
	0.482	0.632	0.290	0.795
Octubre	0.918	0.000	0.881	0.000
	0.918	0.000	0.881	0.000
Noviembre	1.085	0.570	0.492	0.755
	1.174	0.300	0.802	0.548
Diciembre	1.062	0.405	0.725	0.636
	1.062	0.405	0.725	0.636

*El valor superior corresponde a modelos de grado dos y el valor inferior a modelos de grado tres.

FUENTE: Elaboración propia

Anexo 39: Precipitación de Jauja utilizando MARS*

MODELO	GCV	RSQ	ECM	COR
Enero	1833.76	0.000	1735.98	0.273
	1833.76	0.000	1735.98	0.273
Febrero	1039.28	0.647	513.19	0.804
	1039.28	0.647	513.19	0.804
Marzo	2417.84	0.559	1193.91	0.748
	1962.40	0.491	1377.55	0.701
Abril	633.28	0.536	375.73	0.732
	633.28	0.536	375.73	0.732
Mayo	47.68	0.375	31.92	0.613
	47.68	0.375	31.92	0.613
Junio	11.56	0.696	4.81	0.834
	9.43	0.861	2.21	0.928
Julio	21.26	0.360	14.04	0.600
	19.45	0.618	8.38	0.786
Agosto	58.76	0.880	15.60	0.938
	58.02	0.881	15.40	0.939
Setiembre	444.34	0.171	362.14	0.414
	444.34	0.171	362.14	0.414
Octubre	396.60	0.421	268.98	0.649
	396.60	0.421	268.98	0.649
Noviembre	1132.58	0.000	1072.18	0.000
	1132.58	0.000	1072.18	0.000
Diciembre	1345.26	0.000	1273.53	0.000
	1345.26	0.000	1273.53	0.000

*El valor superior corresponde a modelos de grado dos y el valor inferior a modelos de grado tres.

FUENTE: Elaboración propia

Anexo 40: Temperatura mínima de Jauja utilizando MARS*

MODELO	GCV	RSQ	ECM	COR
Enero	0.913	0.757	0.345	0.870
	0.913	0.757	0.345	0.870
Febrero	0.667	0.148	0.537	0.385
	0.667	0.148	0.537	0.385
Marzo	0.535	0.389	0.363	0.624
	0.535	0.389	0.363	0.624
Abril	0.948	0.423	0.768	0.650
	0.948	0.423	0.768	0.650
Mayo	1.875	0.309	1.272	0.556
	1.875	0.309	1.272	0.556
Junio	2.954	0.166	2.407	0.407
	2.954	0.166	2.407	0.407
Julio	2.504	0.559	1.698	0.748
	2.538	0.553	1.721	0.744
Agosto	2.138	0.463	1.225	0.680
	1.799	0.629	0.846	0.793
Setiembre	1.690	0.373	1.146	0.611
	1.585	0.604	0.725	0.777
Octubre	1.248	0.599	0.587	0.774
	1.248	0.599	0.587	0.774
Noviembre	1.178	0.468	0.809	0.684
	1.203	0.359	0.974	0.599
Diciembre	1.262	0.591	0.578	0.769
	1.262	0.591	0.578	0.769

*El valor superior corresponde a modelos de grado dos y el valor inferior a modelos de grado tres.

FUENTE: Elaboración propia

Anexo 41: Temperatura máxima de Jauja utilizando MARS*

MODELO	GCV	RSQ	ECM	COR
Enero	1.014	0.430	0.704	0.656
	1.014	0.430	0.704	0.656
Febrero	1.052	0.598	0.731	0.774
	1.052	0.598	0.731	0.774
Marzo	1.204	0.493	0.703	0.702
	1.387	0.304	0.963	0.552
Abril	0.684	0.660	0.399	0.812
	0.670	0.824	0.207	0.908
Mayo	0.512	0.521	0.351	0.722
	0.512	0.521	0.351	0.722
Junio	0.472	0.867	0.118	0.931
	0.472	0.867	0.118	0.931
Julio	0.849	0.248	0.687	0.498
	0.849	0.248	0.687	0.498
Agosto	0.944	0.561	0.455	0.749
	0.983	0.228	0.801	0.477
Setiembre	0.554	0.610	0.323	0.781
	0.554	0.610	0.323	0.781
Octubre	1.088	0.623	0.537	0.789
	1.072	0.696	0.432	0.834
Noviembre	1.166	0.168	0.955	0.410
	1.166	0.168	0.955	0.410
Diciembre	0.958	0.401	0.665	0.633
	0.958	0.401	0.665	0.633

*El valor superior corresponde a modelos de grado dos y el valor inferior a modelos de grado tres.

FUENTE: Elaboración propia

Anexo 42: Precipitación de Viques utilizando MARS*

MODELO	GCV	RSQ	ECM	COR
Enero	4894.77	0.000	4593.622	-0.073
	4894.77	0.000	4593.622	-0.073
Febrero	5768.60	0.540	3808.18	0.735
	5768.60	0.540	3808.18	0.735
Marzo	5534.69	0.000	5213.91	-0.771
	5534.69	0.000	5213.91	-0.771
Abril	206.38	0.708	75.03	0.841
	206.38	0.708	75.03	0.841
Mayo	10.86	0.465	8.55	0.682
	10.86	0.465	8.55	0.682
Junio	3.83	0.874	0.96	0.935
	4.52	0.851	1.13	0.923
Julio	1.66	0.944	0.90	0.972
	1.66	0.944	0.90	0.972
Agosto	11.51	0.579	5.91	0.761
	11.07	0.684	4.44	0.827
Setiembre	294.14	0.615	165.45	0.784
	294.14	0.615	165.45	0.784
Octubre	1801.50	0.000	1697.09	0.000
	1432.15	0.525	805.59	0.725
Noviembre	1635.72	0.316	1307.15	0.562
	1635.72	0.316	1307.15	0.562
Diciembre	1333.61	0.539	853.51	0.734
	1609.34	0.322	1255.73	0.567

*El valor superior corresponde a modelos de grado dos y el valor inferior a modelos de grado tres.

FUENTE: Elaboración propia

Anexo 43: Pronósticos de precipitación de Huayao para el 2008 usando MARS

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov
Pronóstico	112.1	149.2	97.1	54.6	9.2	2.0	2.8	181.3	40.9	40.5	79.70
Observado	107.6	58.0	55.0	19.9	3.8	11.8	6.4	17.2	-	-	-

FUENTE: Elaboración propia

Anexo 44: Pronósticos de temperatura mínima de Huayao para el 2008 usando MARS

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov
Pronóstico	7.54	6.59	4.44	5.16	2.50	8.68	-0.60	2.59	4.62	6.16	5.81
Observado	7.7	6.5	5.4	4.5	1.0	0.5	-0.1	2.7	-	-	-

FUENTE: Elaboración propia

Anexo 45: Pronósticos de temperatura máxima de Huayao para el 2008 usando RNAB

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov
Pronóstico	18.3	17.2	18.8	20.2	21.5	19.5	19.2	20.0	20.3	20.6	21.6
Observado	18.2	18.5	18.4	20.6	20.2	19.9	19.9	20.7	-	-	-

FUENTE: Elaboración propia

Anexo 46: Pronósticos de precipitación de Jauja para el 2008 usando MARS

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov
Pronóstico	125.3	149.8	87.8	61.4	8.4	5.5	2.3	4.5	24.9	54.0	72.06
Observado	112.7	100.1	62.8	12.2	13.3	6.4	0.6	4.2	-	-	-

FUENTE: Elaboración propia

Anexo 47: Pronósticos de temperatura mínima de Jauja para el 2008 usando MARS

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov
Pronóstico	7.5	6.3	5.2	4.7	3.0	1.7	-1.3	0.7	4.0	4.1	6.03
Observado	6.8	6.5	5.5	4.3	1.7	0.5	-0.2	2.3	-	-	-

FUENTE: Elaboración propia

Anexo 48: Pronósticos de temperatura máxima de Jauja para el 2008 usando MARS

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov
Pronóstico	16.6	15.8	17.8	17.6	19.0	19.3	19.6	20.8	21.0	20.0	20.26
Observado	16.5	16.7	16.6	19.1	19.4	19.7	19.5	20.2	-	-	-

FUENTE: Elaboración propia

Anexo 49: Pronósticos de precipitación de Viques para el 2008 usando MARS

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov
Pronóstico	80.1	138.9	77.0	18.6	6.8	6.9	0.2	0.2	92.5	47.0	62.23
Observado	123.8	109.2	69.1	0.0	2.1	9.6	0.0	0.0	-	-	-

FUENTE: Elaboración propia