ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

PREDICCIÓN DE LA PRECIPITACIÓN A PARTIR DE VARIABLES METEOROLÓGICAS UTILIZANDO MODELOS DE REGRESIÓN FUNCIONAL

PROYECTO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

MENCIÓN EN ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

DANILO LEANDRO LOZA QUISPILLO

danilo.loza@epn.edu.ec

ÁNGEL OMAR LLAMBO DELGADO

angel.llambo@epn.edu.ec

DIRECTOR: Ph.D. MIGUEL FLORES SÁNCHEZ

miguel.flores@epn.edu.ec

Quito, Enero 2022

DECLARACIÓN

Nosotros, Danilo Leandro Loza Quispillo y Ángel Omar Llambo Delgado, declaramos que el trabajo aquí descrito es de nuestra autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que hemos consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional, puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normativa institucional vigente.

DANILO LEANDRO LOZA QUISPILLO

ÁNGEL OMAR LLAMBO DELGADO

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Danilo Leandro Loza Quispillo y Ángel Omar Llambo Delgado, bajo mi supervisión.

Ph.D. MIGUEL FLORES SÁNCHEZ
DIRECTOR DEL PROYECTO

AGRADECIMIENTO

Deseo expresar gratitud a mis padres David y Mayra por el apoyo que me han brindado durante toda la vida y a quienes debo todo desarrollo profesional que he obtenido. A mi hermano Gustavo a quien siempre apoyaré y le desearé lo mejor. A ellos que han sido el pilar fundamental para alcanzar mis objetivos.

Agradecer a Dios y la Virgen que siempre me han escuchado y han cuidado de mí.

A mis amigos más allegados con quienes siempre he contado y he compartido todas las experiencias maravillosas de esta vida. A ellos les agradezco por haberme brindado su lealtad durante tantos años.

De la manera más cordial también agradezco a mi amigo, compañero y colega Ángel ya que sin su apoyo este proyecto no hubiese sido posible.

Finalmente quiero dar mi más sincero agradecimiento al Ph.D Miguel Flores quien ha dirigido esta tesis, y cuyo conocimiento, instrucción y contribución permitió el desarrollo de este trabajo.

Danilo Leandro Loza Quispillo

AGRADECIMIENTO

A Dios, a la música y a la vida, que siempre me han consentido y han cuidado tanto de mi.

A mi familia, en especial a mis padres, Efrain y Betty, que me apoyaron indispensablemente en esta aventura y me regalaron los valores éticos que solo se aprenden en casa. Que con su ejemplo, trabajo, fortaleza y humildad se convirtieron en los mejores profesores que jamás tuve. A mis hermanos: Mauro, Paulo y Sammy por su imprescindible apoyo y compañía en mi vida.

A mis amigos que desinteresadamente me obsequiaron su amistad. Por todo el tiempo y los grandiosos momentos compartidos; mi aprecio, gratitud y respeto siempre para ustedes.

A mi amigo Danilo, quien me acompañó en las aulas y en el presente trabajo, que sin su ayuda no hubiese sido posible concluirlo.

Quiero agradecer de manera especial a nuestro tutor de tesis, Ph.D Miguel Flores, por la asesoría y por su tiempo brindado para el desarrollo de este proyecto.

También agradezco de sobremanera a todos mis profesores que tuve a lo largo de mis diferentes etapas como estudiante. Este logro guarda eternamente sus huellas.

Finalmente, mi gratitud con la Escuela Politécnica Nacional por haberme brindado un lugar en sus prestigiosas aulas.

Ángel Omar Llambo Delgado

DEDICATORIA

“Somos viajeros en un viaje cósmico,
polvo de estrellas que gira
y baila en los remolinos del infinito.
La vida es eterna.
Nos hemos detenido un momento a encontrarnos,
a conocernos, a amarnos, a compartir.
Este es un momento precioso.
Un pequeño paréntesis en la eternidad.”
Paulo Coelho

Cuánto más exactamente cierta posición se percibe menos preciso será el momento, es la incertidumbre que existe en el universo. Y por suerte en este caso, todo aquel caos perverso se solventa con un intento.
A mi familia especialmente a mi madre y a mi padre a quienes admiro tanto.

Danilo Leandro Loza Quispillo

DEDICATORIA

No dejes de creer que las palabras,
la risa y la poesía sí pueden cambiar el mundo.
No dejes nunca de soñar,
porque sólo a través de sus sueños
puede ser libre el hombre.
No permitas que la vida
te pase por encima sin que la vivas.
No abandones tus ansias de hacer de tu vida
algo extraordinario.

Walt Whitman

A mis padres, por darme las alas y enseñarme a volar; por su apoyo incondicional y el amor que me dan. A mis hermanos y a mi amigo Leo que forman parte de mi ser.

Este trabajo es para ustedes. Recuerden, siempre los llevo conmigo.

Ángel Omar Llambo Delgado

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

En el presente trabajo se emplean los Modelos de Regresión Lineal Funcional donde la variable de respuesta pertenece al campo escalar ($\mathbb{R}$), mientras que las covariables explicativas tienen una estructura funcional y se encuentran en el espacio cuadrado integrable ($\mathcal{L}^2$).

Previamente se abordan las técnicas de análisis exploratorio y detección de valores atípicos, en el marco del Análisis de Datos Funcionales. Luego, se realiza la aplicación de tres distintos tipos de regresión lineal funcional con respuesta escalar en los cuales interviene una única covariable explicativa, y se lleva a cabo la adaptación de un cuarto modelo que contempla dos covariables explicativas a la vez.

La metodología propuesta se utiliza con el propósito de escoger el mejor ajuste entre los modelos para predecir la precipitación (variable de respuesta escalar) en diferentes zonas con producción de maíz del Ecuador, mediante covariables explicativas funcionales como la temperatura y la velocidad del viento. Estos resultados serán de gran utilidad para calcular el valor de la precipitación efectiva, fracción de agua utilizada por el cultivo, del cual se obtiene el valor de la evapotranspiración real de la planta y, lo que es más importante, se determinan los requerimientos de riego en cada etapa de producción del maíz para entender el comportamiento de la productividad de este cultivo mediante su adecuado desarrollo fisiológico.

ABSTRACT

In the present work, Functional Linear Regression Models are used where the response variable belongs to the scalar field $(\mathbb{R})$, while the explanatory covariates have a functional structure in the square integrable space $(\mathcal{L}^2)$.

Previously, exploratory analysis and outlier detection techniques are discussed, within the context of Functional Data Analysis. Then, three different types of functional linear regression with scalar response are applied in which only one explanatory covariate is involved, and a fourth model that considers both explanatory covariates at the same time is adapted.

The proposed methodology is used with the purpose of choosing the best fit among models to predict precipitation (scalar response variable) in different corn producing zones of Ecuador, by means of functional explanatory covariates such as temperature and wind speed. These results will be very useful to calculate the value of effective precipitation, fraction of water used by the crop, from which the value of the real evapotranspiration of the plant is obtained and, more importantly, to determine the irrigation requirements at each stage of corn production to understand the behavior of the productivity of this crop through its adequate physiological development.

CAPÍTULO 1

INTRODUCCIÓN

1.1. INTRODUCCIÓN

Los desarrollos tecnológicos han hecho posible que los investigadores de muchas áreas dispongan de grandes volúmenes de información contenidos en curvas, superficies o datos pertenecientes a un espacio continuo, por lo que, existe un gran interés por el aprendizaje y desarrollo de herramientas estadísticas para su análisis. En el análisis multivariante la situación se torna compleja cuando se tratan variables correlacionadas por la gran cantidad de información que poseen como suele ocurrir con las series temporales o con los procesos estocásticos en tiempo continuo en donde los datos no son numerables.

El Análisis de Datos Funcionales *(FDA)* resulta ser una herramienta estadística de alto alcance, novedosa y cuyo objetivo es simplificar el estudio de variables que dependen de parámetros continuos gracias a que los datos se representan a través de curvas o en general de funciones; pues el análisis multivariante es insuficiente al procesar datos estructurados como funciones. Generalmente el *FDA* tiene una ventaja significativa que es la reducción apreciable de la dimensión que tienen este tipo de variables mediante un método de suavizamiento; ya que los datos observados inicialmente tienen una naturaleza discreta se desea utilizar un método para suavizar los datos iniciales con el fin de obtener una función como unidad básica de información (Carrillo Ramirez et al., 2017).

Por otro lado, la regresión es un procedimiento estadístico que consiste en estimar los parámetros que caracterizan la relación que existe entre una o más covariables explicativas con una variable de respuesta. Se plantean diferentes formas de estimar dicha relación mediante varios Modelos de Regresión Lineal Funcional *(FLR)* con Respuesta Escalar (Ramsay & Silverman, 2005). Los modelos de regresión se pueden aplicar en muchos campos, nos enfocaremos en el contexto agro-meteorológico ya que en Ecuador existe una gran variedad de cultivos, siendo

el maíz el segundo cultivo de mayor producción, cuya productividad o capacidad de producción se ve favorecida o afectada por factores climáticos como la cantidad de agua de lluvia o precipitación que se presenta desde el ciclo de siembra hasta la cosecha del cultivo.

El presente trabajo se encuentra estructurado en cuatro partes. En el Capítulo 1 se explica en qué consiste el caso de estudio para identificar las variables meteorológicas que intervienen en la productividad o capacidad de producción del maíz en Ecuador. Para ello, se detalla la selección de cantones ubicados estratégicamente por el alto nivel de producción de maíz que poseen.

En el Capítulo 2 se presenta el marco teórico del *FDA*, empezando con las definiciones principales, seguido del método de suavizamiento para la representación funcional de datos discretos mediante varios tipos de bases funcionales, y el estudio de las componentes funcionales.

En el Capítulo 3 se abordan los modelos *FLR* más importantes para estimar la relación entre una o más covariables explicativas funcionales y una variable de respuesta escalar.

En el Capítulo 4 se realiza la aplicación los modelos *FLR* con el fin de seleccionar y validar el mejor ajuste entre ellos para predecir la precipitación como variable de respuesta escalar, en cualquier lugar del país que tenga producción de maíz, a través de covariables explicativas funcionales como la temperatura y la velocidad del viento. De forma que, los resultados se aprovecharán para entender el comportamiento de la productividad del maíz a través de los requerimientos hídricos que se necesiten en las zonas de estudio.

Finalmente, en el Capítulo 5 se exponen las conclusiones y recomendaciones obtenidas por medio del *FDA* efectuado sobre las covariables funcionales, además los resultados proporcionados por los modelos *FLR*, y los requerimientos hídricos que se necesitan en las áreas de interés para comprender la productividad del maíz.

1.2. OBJETIVOS

1.2.1. OBJETIVO GENERAL

Predecir la precipitación utilizando modelos de regresión funcional con respuesta escalar para entender el comportamiento de la productividad del maíz en Ecuador.

1.2.2. OBJETIVOS ESPECÍFICOS

Seleccionar las variables meteorológicas más relevantes e influyentes en la producción agrícola de Ecuador a través del análisis de su comportamiento para estimar el mejor modelo de predicción de la precipitación en determinadas regiones del país.

Seleccionar el número óptimo de bases funcionales mediante técnicas de validación cruzada para encontrar la mejor representación funcional de las variables meteorológicas consideradas.

Validar el mejor ajuste de los modelos de regresión funcional a través de varios contrastes de hipótesis que permitan verificar la significancia de los modelos.

1.3. JUSTIFICACIÓN

Los modelos de regresión funcional se aplicarán con la finalidad de mejorar el problema de predicción que se presenta en el caso multivariante. Es decir, se desea encontrar una mejor relación entre las variables explicativas y de respuesta donde se considere la alta dimensionalidad. Por ejemplo, en el caso de Canadian Weather se consideró las temperaturas y el promedio de las precipitaciones anuales en cada estación meteorológica canadiense (Ramsay et al., 2009). De aquí, se observó la relación existente entre la variable explicativa (temperatura) y la variable de respuesta escalar (precipitación) mediante la regresión funcional, encontrando excelentes resultados de predicción climática para las estaciones de Canadá.

El uso de los modelos *FLR* permitirá obtener resultados predictivos de gran precisión que serán de utilidad para entender el comportamiento de la productividad agrícola del maíz en Ecuador. Con estos resultados se pueden establecer parámetros que permitan ilustrar escenarios (favorables y adversos) para poder afrontar de

mejor manera los temporales que se suscitan durante el año. Además, se podrán divulgar los resultados obtenidos para poner en práctica proyectos y/o estudios que permitan mejorar las condiciones socio-económicas del sector agrícola.

1.4. CASO DE ESTUDIO

Esta sección está organizada en tres partes. En primer lugar, se detalla particularmente la producción de maíz a nivel provincial y cantonal con el fin de determinar las zonas con alta capacidad productiva para este estudio.

Para continuar, se explica la relación entre la productividad de un cultivo y el balance hídrico que se necesita durante todo el ciclo de producción. Para ello, se describen los índices que intervienen en dicha relación y los cuales servirán para determinar los requerimientos hídricos imprescindibles en cada etapa del cultivo del maíz.

Por último, se especifican los datos recolectados de las variables meteorológicas con los que se trabajará el desarrollo del *FDA* y posteriormente la aplicación de los modelos *FLR*.

1.4.1. PRODUCCIÓN DE MAÍZ EN ECUADOR

En Ecuador existe una distribución del uso del suelo para sector agropecuario, formado por el sector agrícola (agricultura) y ganadero o pecuario (ganadería), que es esencial para la economía del país; aportando económicamente al país el $7,7\,\%$ del *PIB* según datos oficiales del Banco Central de Ecuador (BCE, 2021).

En el año 2020 la contribución agropecuaria alcanzó los $5,2$ millones de hectáreas (Ha) con producción de cultivos permanentes (caña, banano, palma africana, cacao) y transitorios (arroz, papa y maíz duro seco) según estudios del Instituto Nacional de Estadística y Censos (Márquez, 2021).

A continuación, se presentan los porcentajes de participación del uso del suelo proporcionados por la Encuesta de Superficie y Producción Agropecuaria Continua ESPAC (2020) para las regiones Sierra, Costa y Amazónica:

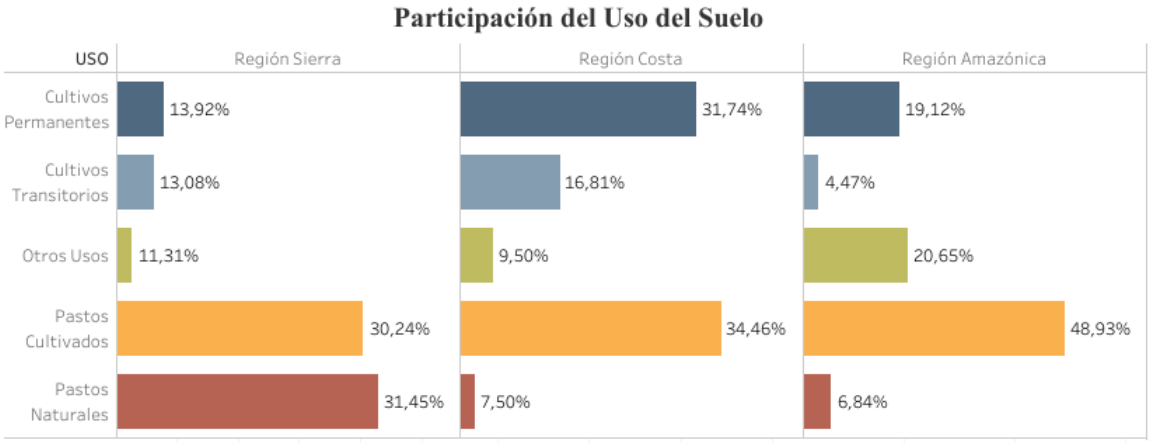

FIGURA 1.1. Participación de uso del suelo en las regiones Sierra, Costa y Amazónica.

En la figura (1.1) se observa que el mayor porcentaje de participación del uso del suelo se encuentra asignado a pastos cultivados, seguido de los cultivos permanentes y en tercer lugar se hallan los cultivos transitorios que representan el $15,76\,\%$ de la superficie de labor agropecuaria (Márquez, 2021). Estos últimos tienen una representación agrícola importante en la región Costa donde acumula el $16,81\,\%$, mientras que en la región Sierra y Amazónica existe un $13\,\%$ y $4,5\,\%$ de aportación respectivamente.

En este estudio se considera la producción del maíz, que es el segundo cultivo transitorio de mayor producción del país y cuya contribución es del $32,9\,\%$, solo por detrás del arroz que tiene el $38,2\,\%$ de participación (ESPAC, 2020). Es fundamental, mencionar que este tipo de cultivos se desarrollan en un ciclo corto de tiempo (frecuentemente dura pocos meses), es decir, una vez que dan su fruto se destruye la planta para realizar una nueva cosecha; lo que facilita el análisis de su productividad a través del desarrollo de la planta en cada etapa de producción.

En este sentido, en la siguiente tabla se muestran las provincias con mayor producción anual durante el 2020 de los tipos de maíz que se cultivan en el país:

Provincia	Tipo de Maíz	Uso de la Superficie (Ha)		Producción Anual (Tm)
		Sembrado	Cosechado	
Esmeraldas	Maíz duro seco	1.746	1.668	3.873
	Maíz duro choclo	311	311	180
Manabí	Maíz duro seco	10.4746	90.749	280.757
	Maíz duro choclo	714	668	1636
Los Ríos	Maíz duro seco	14.7434	144.109	642.761
	Maíz duro choclo	77	77	306
Santa Elena	Maíz duro seco	5.302	5.267	23.280
	Maíz duro choclo	210	210	884
Guayas	Maíz duro seco	58.866	55.511	247.712
	Maíz duro choclo	1.425	1.425	9.849
	Maíz suave seco	25	25	7
El Oro	Maíz duro seco	387	387	592
	Maíz duro choclo	63	63	74
	Maíz suave seco	666	610	172
Loja	Maíz duro seco	21.631	20.363	58.460
	Maíz suave choclo	893	884	2.634
	Maíz suave seco	867	837	1.443
Sucumbíos	Maíz duro seco	7.366	7.035	16.145
	Maíz duro choclo	551	551	465
Orellana	Maíz duro seco	9.339	9.057	19.564
Santo Domingo de los Tsáchilas	Maíz duro seco	1.161	1.126	1.974
	Maíz duro choclo	165	165	341

TABLA 1.1. Uso del suelo por hectáreas (Ha) y Producción anual por toneladas métricas (Tm) de cada tipo de maíz en las provincias con mayor productividad.

La información proporcionada por la tabla (1.1) muestra que se produce gran variedad de maíz; siendo el maíz duro seco el que tiene mayor cantidad de hectáreas para su siembra y cosecha y del cual se ha generado la mayor producción por toneladas métricas este último año. De esta manera, se presenta la participación de las provincias en la producción por toneladas métricas Tm de maíz duro seco:

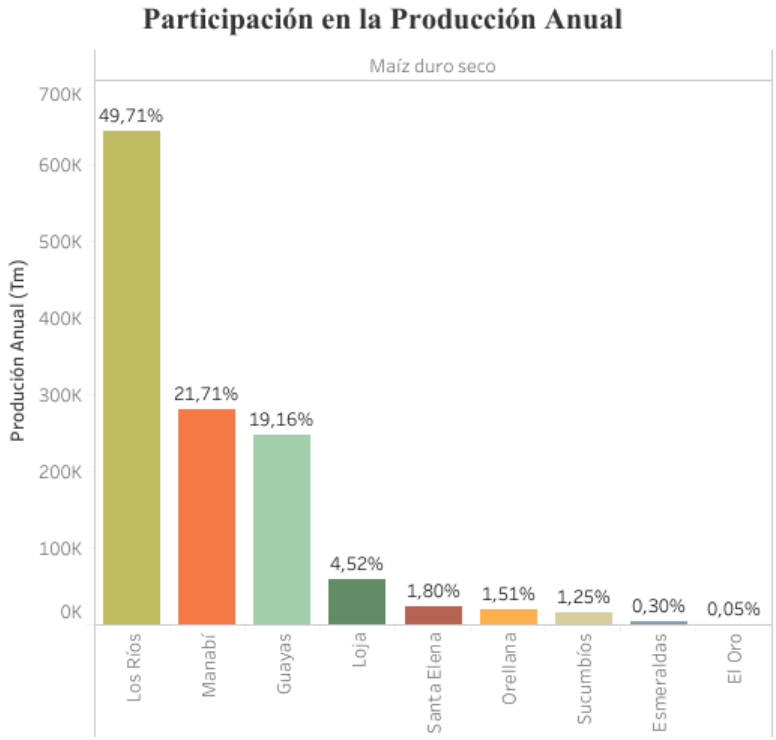

FIGURA 1.2. Participación de la producción anual en toneladas métricas (Tm) de maíz duro seco en las provincias con mayor productividad.

Este cultivo está presente con mayor concentración en las zonas costeras como en la provincia de los Ríos que tiene el $49,26\,\%$ de representación de la producción nacional, seguida de Manabí con el $22\,\%$ y Guayas con el $19\,\%$. Mientras que, las provincias de Loja, Santa Elena, Orellana, Sucumbíos, El Oro y Esmeraldas participan con menor porcentaje respectivamente.

En este contexto, para el desarrollo del caso de estudio se escogen escogen los siguientes lugares estratégicamente ubicados en las provincias mencionadas:

Provincia	Cantones	Coordenadas	
		Latitud	Longitud
Esmeraldas	San Lorenzo	1,29	-78,84
	Quinindé	0,33	-79,48
	Esmeraldas	0,96	-79,65
Manabí	Pedernales	0,08	-80,05
	Chone	-0,70	-80,09
	El Carmen	-0,27	-79,43
	Manta	-0,96	-80,71
	Portoviejo	-1,05	-80,45
	Jipijapa	-1,35	-80,58
Los Ríos	Buena Fe	-0,89	-79,49
	Quevedo	-1,03	-79,46
	Babahoyo	-1,80	-79,53
Guayas	Balzar	-1,37	-79,88
	Guayaquil	-2,20	-79,89
	Simón Bolívar	-2,02	-79,48
	Pedro Carbo	-1,82	-80,23
	Naranjal	-2,67	-79,62
Santa Elena	Santa Elena	-2,23	-80,85
El Oro	Pasaje	-3,33	-79,81
	Zaruma	-3,68	-79,62
	Arenillas	-3,55	-80,07
Loja	Loja	-3,99	-79,00
	Paltas	-4,07	-79,63
Sucumbíos	Shushufindi	-0,19	-76,65
Orellana	Loreto	-0,69	-77,31

TABLA 1.2. Coordenadas de los Cantones productores de maíz duro seco en las provincias con mayor productividad.

Los cantones 25 mostrados en la tabla (1.2) han sido elegidos por su participación en la producción de maíz duro seco a lo largo del 2020. Cabe destacar que estos cantones históricamente han sido los lugares con altos niveles de productividad ya que el clima es oportuno para producir el maíz (Márquez, 2021).

1.4.2. RELACIÓN ENTRE PRODUCTIVIDAD Y BALANCE HÍDRICO

En 1967 Turc planteó un índice en el que intervienen variables meteorologías para determinar la productividad agrícola de los cultivos por hectárea con mayor exactitud. Para ello, Turc comparaba el dato real de la productividad con el cálculo del índice y después predecía la productividad estimando dicho índice para las demás regiones (Estrada et al., 2013).

Tradicionalmente se han desarrollado mecanismos para detectar los beneficios que proporciona el uso del agua en los cultivos sembrados. Existen versiones modificadas del índice de Turc aplicadas en países andinos para cultivos como maíz, soya, etc., en estos estudios se ha encontrado dependencia de la cantidad de agua disponible sobre las áreas de estudio, es decir, que estos resultados dan evidencia de la importancia directa que tiene la precipitación en los niveles de productividad agrícola (Estrada, 2011).

En este marco, se plantea hallar un balance hídrico determinado por la pérdida de humedad en el suelo como consecuencia de la evaporación y transpiración del agua acumulada por la lluvia en una planta, proceso conocido como evapotranspiración, para verificar el rendimiento productivo que tendrá el cultivo desde la siembra hasta la cosecha. De manera que, se pueda establecer si la precipitación producida en los lugares donde se cultiva el maíz es suficiente o no para que el cultivo tenga un proceso de desarrollo fisiológico adecuado tal que su productividad se vea beneficiada.

El balance empieza con la estimación de la evapotranspiración de referencia (ET_o) medida en milímetros por día (mm/d) mediante:

$$ET_o = \frac{0,408\Delta(R_n - G) + \gamma\frac{900}{T+273}u(c_s - c_a)}{\Delta + \gamma(1 + 0,34u)} \tag{1.1}$$

Esta fórmula planteada por Penman-Monteith es más exacta para los cultivos de países andinos (FAO, 1977). A continuación, se describe cada término:

- R_n: Radiación neta en la superficie de cultivo medida en mega Joules por metro cuadrado por día $(MJm^{-2}d^{-1})$.

- G: Densidad del flujo de calor en la superficie del suelo cuya unidad de medida es $MJm^{-2}d^{-1}$.
- T: Temperatura media del aire a 2 metros de altura y se encuentra medida en grados Celsius $(°C)$.
- u: Velocidad del viento promedio a 2 metros de altura medida en metros por segundo (m/s).
- c_s: Presión del vapor de saturación medida en kilo pascal (kPa).
- c_a: Presión real del vapor cuya unidad de medida es kPa.
- Δ: Pendiente de la curva de la presión de vapor medida en kilo pascal por grado centígrado $(kPa/°C)$.
- γ: Constante psicrométrica cuya unidad de medida es $kPa/°C$.

La ET_o se define como la evapotranspiración que se produce bajo condiciones óptimas para satisfacer el requerimiento hídrico de la planta. Además, los principales factores meteorológicos que afectan la ET_o son la radiación, la temperatura, la humedad atmosférica y la velocidad del viento.

Luego, se define la evapotranspiración real (ET_c) como la evapotranspiración que realmente se produce en condiciones existentes de cada cultivo y es calculada por:

$$ET_c = K_c \cdot ET_o \tag{1.2}$$

donde K_c es la constante de la etapa en la que se encuentra el proceso de cultivo. En el caso del maíz se tiene:

- Etapa Inicial: $K_c = 0.3$.
- Etapa Desarrollo: $K_c = 0.8$.
- Etapa Media: $K_c = 1.2$.
- Etapa Maduración: $K_c = 0.35$.

A partir de estos indicadores se podrá determinar si se cumplen los requerimientos hídricos en las zonas de estudio o no, pues los niveles de productividad se ven favorecidos o afectados por estos indicadores.

En Ecuador se suelen realizar siembras por ciclos en épocas de lluvia ya que precisamente se aprovecha la lluvia para obtener un proceso adecuado del cultivo de modo que dichos niveles aumenten.

1.4.3. VARIABLES METEOROLÓGICAS

En la presente sección se analizan las variables meteorológicas más influyentes en la productividad agrícola del maíz duro seco para los 25 cantones de Ecuador mostrados en la taba (1.2), ya que tienen una importante característica socio-productiva. En este sentido, la precipitación, la temperatura y la velocidad del viento son las variables meteorológicas seleccionadas por su influencia directa en la productividad del cultivo y fundamentales para determinar el balance hídrico detallado anteriormente. En FAO (1977) se explica con mayor exactitud.

Para ello, se han tomado datos recolectados diariamente durante 10 años desde $(01/1/2010)$ hasta $(31/12/2020)$ para cada variable meteorológica mencionada de la NASA. En primer lugar, se considera el promedio de la precipitación por cada cantón con el fin de obtener un vector de dimensión (25). Posteriormente, se trabaja con el promedio diario de la temperatura hallado en cada uno de los 25 cantones obteniendo una estructura matricial de dimensión (365×25); de manera similar se obtiene el promedio diario para la velocidad del viento.

Más adelante, estos resultados permitirán construir una estructura funcional de los datos para el *FDA* y la aplicación de los Modelos *FLR*, estos tópicos se explican ampliamente en los capítulos posteriores.

1.4.3.1. Precipitación

La precipitación se define como la caída de agua desde la atmósfera hacia la superficie terrestre y es una parte importante del ciclo hidrológico. La cantidad de precipitación se mide en milímetros (mm) o, lo que es lo mismo, un milímetro de

lluvia representa un litro de agua por metro cuadrado (l/m^2) Meteoblue.

Para el desarrollo de los cultivos la precipitación es un factor muy importante ya que afecta a la humedad del suelo; este proceso de pérdida de humedad es la evapotranspiración (Jiménez et al., 2012). Por ello, se deben satisfacer las necesidades hídricas o la demanda de agua (por riego o por lluvia) de los cultivos medida en forma de evapotranspiración con el fin de que los cultivos se desarrollen adecuadamente. Por tanto, la precipitación es la variable más representativa relacionada a la cantidad de agua necesaria para los cultivos y es fundamental para el desarrollo de este trabajo.

Esto apoya el análisis de la precipitación en cada una de las superficies dadas en la tabla (1.2). Por lo cual, se muestra geográficamente el nivel que ha tenido esta variable en los distintos cantones por cada provincia de Ecuador.

- Precipitación en las provincias de Esmeraldas y Manabí

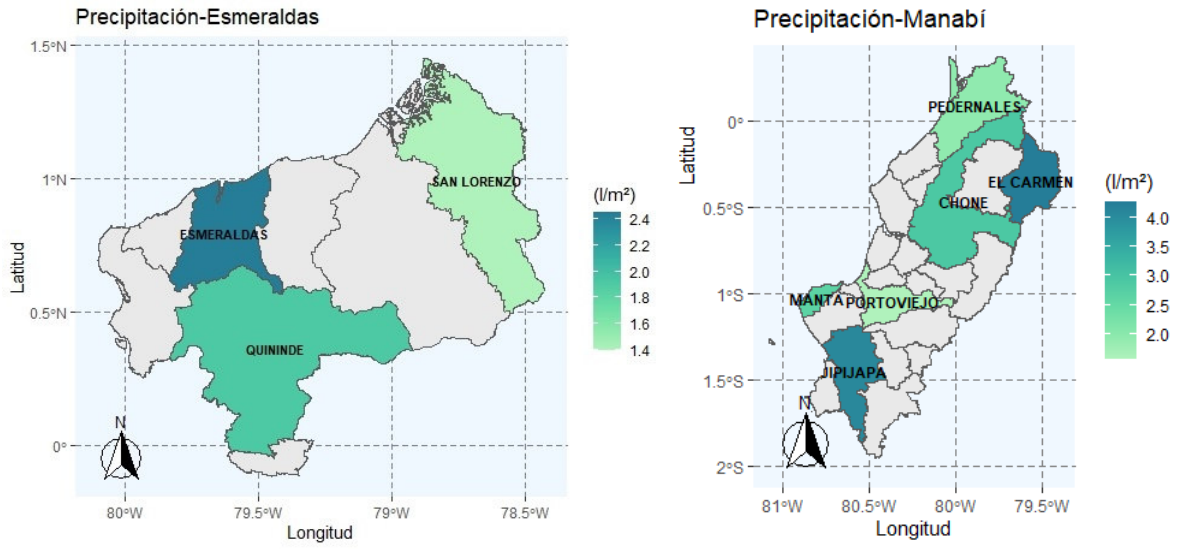

FIGURA 1.3. Precipitación promedio en los cantones de Esmeraldas y Manabí.

- Precipitación en las provincias de Santa Elena y Guayas

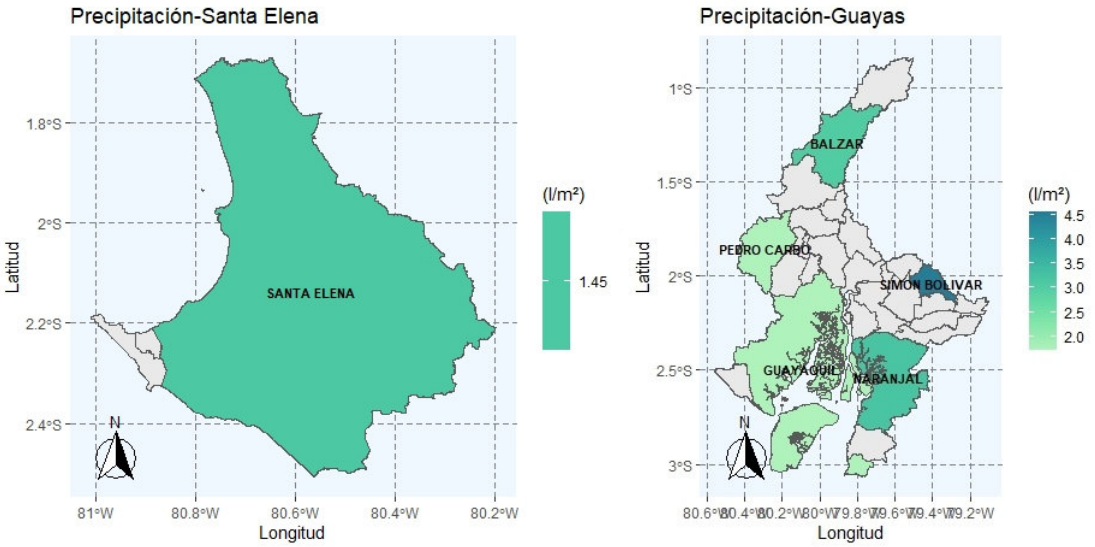

FIGURA 1.4. Precipitación promedio en los cantones de Santa Elena y Guayas.

- Precipitación en las provincias de Loja y Los Ríos

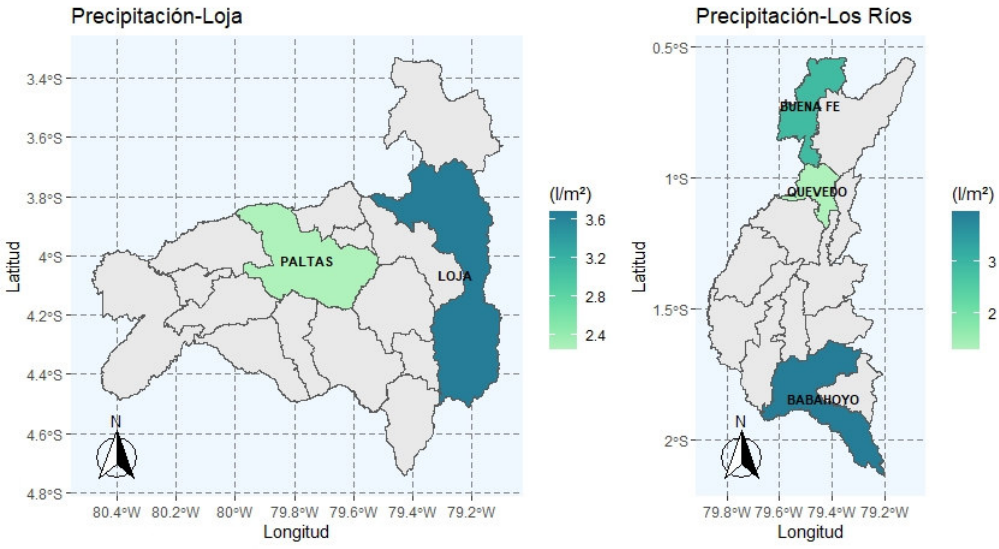

FIGURA 1.5. Precipitación promedio en los cantones de Loja y Los Ríos.

- Precipitación en las provincias de Sucumbíos y Orellana

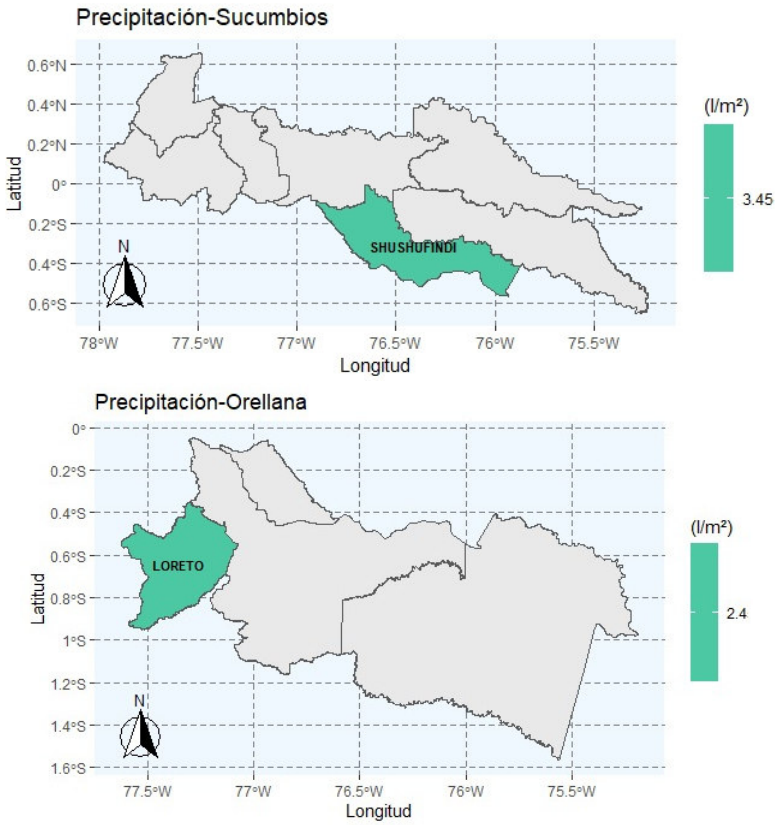

FIGURA 1.6. Precipitación promedio en los cantones de Sucumbíos y Orellana.

Analizando a nivel provincial, considerando el promedio de la precipitación de los cantones seleccionados para el estudio, se tiene que las provincias con mayores niveles de precipitación son Los Ríos y Sucumbíos con un promedio de $3,48\ mm$ y $3,46\ mm$ respectivamente. Mientras que los valores mínimos se registraron en las provincias de Santa Elena y El Oro.

Las medidas de precipitación, aunque fueron recolectadas sobre cantones que poseen similares condiciones agrícolas relacionadas al cultivo del maíz, presentan diferencias entre cada cantón de una misma provincia. Por ejemplo, en las provincias de Manabí y Los Ríos las precipitaciones entre cada cantón varían alrededor de $0,89\ mm$ y $0,74\ mm$, respectivamente; en el caso de la provincia de El Oro se tiene la menor variabilidad que se encuentra alrededor de $0,22\ mm$.

Estos niveles pueden incrementar o disminuir con el pasar del tiempo debido a la cantidad de agua que se pueda presentar en cada cantón.

1.4.3.2. Temperatura

Se define a la temperatura como la magnitud física que señala la energía interna de un cuerpo, es decir, indica el grado de calor (temperatura alta) o frío (temperatura baja) que tiene un objeto o el medio ambiente en general. Dicha energía interna se expresa en grados Celsius $(°C)$ Meteoblue.

La temperatura es un factor meteorológico importante para que los cultivos cumplan su ciclo y alcancen su rendimiento máximo; ya que tiene impactos directos en el desarrollo de las especies agrícolas (Jiménez et al., 2012).

A continuación, se muestra el comportamiento del promedio diario de la temperatura en las zonas de estudio:

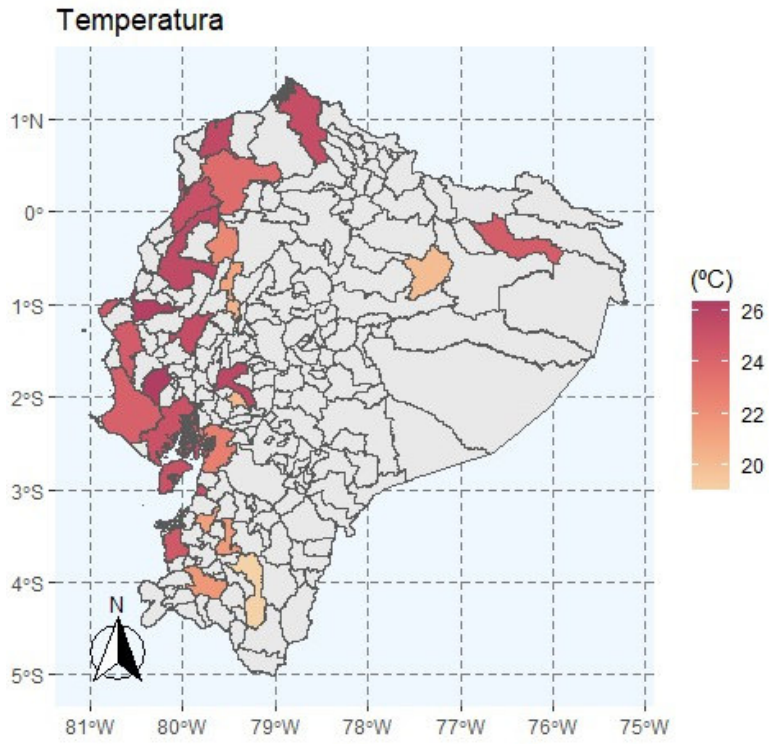

FIGURA 1.7. Temperatura promedio sobre las zonas de estudio.

Se puede ver que los niveles de temperatura son altos en la mayoría de los cantones. Sin embargo, existen algunos cantones ubicados al sur y al este del país que presentan niveles más bajos de temperatura.

En los cantones ubicados en las provincias de Esmeraldas, Manabí, Guayas, Santa Elena y Sucumbíos se presentan niveles de temperatura superiores a los $24°C$; mientras que, en Los Ríos, Loja, Orellana existe una oscilación de la temperatura entre los $20°C$ y $22°C$.

1.4.3.3. Velocidad del Viento

La velocidad del viento es un factor meteorológico que tiene dos características importantes: la dirección y la velocidad, de modo que se define como el movimiento del aire a una velocidad cambiante. Los fenómenos atmosféricos locales como nubes de tormenta son causa de los movimientos verticales del aire, y en sentido horizontal es la corriente de aire que se desplaza. La unidad básica de medida de esta variable es el metro por segundo (m/s) Meteoblue.

La importancia de la velocidad del viento en la agricultura se da en la toma de decisiones sobre los cultivos. En el caso de vientos fuertes sobre los cultivos derivan causas como: fumigaciones, la ruptura de ramas, tallos y tejidos, un mayor proceso de evapotranspiración, plagas e intoxicación. Mientras que, en el caso de vientos suaves se tienen efectos como: la renovación del aire para agilizar la evapotranspiración, se evita heladas nocturnas, sequía las cosechas y suelos encharcados, dispersa insectos benéficos, remueve humedad excesiva evitando ataque de patógenos (García de Pedraza, 1963).

En este contexto, se analiza la velocidad del viento debido a la variabilidad que presenta incluso entre cantones cercanos. Se visualiza el promedio diario de la velocidad del viento en las zonas de estudio en la siguiente figura:

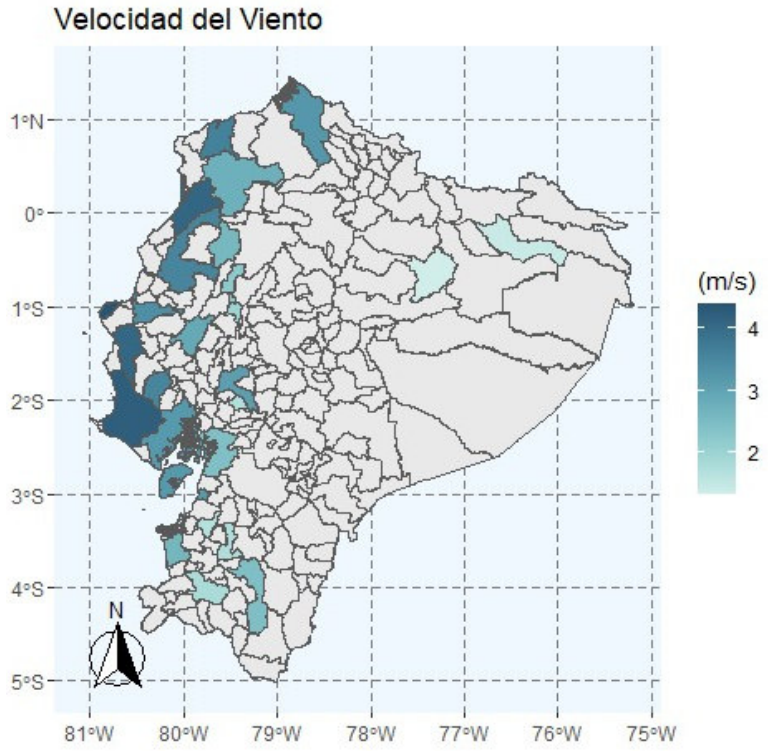

FIGURA 1.8. Velocidad del Viento promedio sobre las zonas de estudio.

Se puede observar que los cantones que se encuentran cerca del océano Pacífico tienen mayor nivel de velocidad del viento en comparación al resto de cantones. Se tiene que, los niveles de velocidad de viento más altos se hallan en Santa Elena alcanzando $4\ m/s$, los niveles medios de esta variable se localizan en las provincias de Esmeraldas, Manabí, Los Ríos y Guayas variando entre los $2\ m/s$ y $3\ m/s$; mientras que, en Loja, Sucumbíos y Orellana los niveles de esta variable son bajos y varían entre $1\ m/s$ y $2\ m/s$.

CAPÍTULO 2

ANÁLISIS DE DATOS FUNCIONALES

2.1. DEFINICIONES

El Análisis de Datos Funcionales *(FDA)* es una técnica desarrollada en el ámbito de la estadística que estudia la información contenida en curvas, es decir, muestras de funciones aleatorias y en general en cualquier elemento que esté variando sobre un espacio continuo. La herramienta estadística que sustenta de manera teórica el análisis de datos funcionales es la de los procesos estocásticos.

En el análisis práctico se empieza a trabajar con observaciones discretas de las curvas muestrales a partir de las cuales se reconstruye la forma funcional. Así, se mantienen las propiedades de las variables continuas y por ende se pierde menos información. El *FDA* cuenta con ventajas sustanciales, puesto que tiene hipótesis de menor exigencia a comparación de técnicas comunes como el análisis de series temporales que imponen, entre otros requerimientos, eliminar la tendencia o no estacionariedad de la serie, contar con observaciones igualmente espaciadas, y que la pertenencia del proceso sea de una clase específica.

Definición 2.1 *(Variable Aleatoria Funcional). Se dice que una variable* $\{\mathcal{X}(t)\}_{t\in[0,T]}$ *definida sobre un espacio de probabilidad* (Ω, A, P) *es una variable funcional si toma valores en un espacio infinito dimensional (espacio funcional), es decir, un espacio normado o semi-normado completo.*

Definición 2.2 *A una observación de* χ_i *de* $\mathcal{X}(t)$*, se le llama dato funcional.*

Definición 2.3 *Sea* Ω *un subconjunto compacto y medible de* $\mathbb{R}$ *con medida de Lebesgue positiva. Se define el espacio de Hilbert* $\{\mathcal{L}^2[T] : \ T = [0, T] \in \Omega\}$ *que es el espacio de funciones cuadrado integrable sobre el intervalo real* T *y se determina por:*

$$\mathcal{L}^2[T] = \left\{ f : T \to \mathbb{R} : \int_T f^2(t)dt < \infty \right\}$$

cuyo producto escalar usual esta definido por:

$$\langle f, g\rangle = \int_T f(t)g(t)dt \quad \forall f, g \in \mathcal{L}^2(T)$$

Este producto es el equivalente al producto interno de vectores en $\mathbb{R}^n$*.*

De manera que, el *FDA* explora los datos que son funciones pertenecientes al espacio $\mathcal{L}^2[T]$ con la finalidad de extender las metodologías del análisis de datos multivariado a datos funcionales. Por lo que, se puede extraer conceptos tales como:

- Media muestral: $\bar{\mathcal{X}}(t) = \frac{1}{n}\sum_{i=1}^{n}\chi_i(t)$
- Varianza muestral: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}[\chi_i(t) - \bar{\mathcal{X}}(t)]^2$

En este contexto, se considera la variable funcional $X_t = \{\mathcal{X}(t)\}_{t\in[0,T]}$ como un proceso estocástico en el espacio $\mathcal{L}^2[0,T]$ para definir:

Definición 2.4 *(Media Funcional). Sea* $\mathcal{S}_n = \{\chi_i(t)\}_{i=1}^n$ *una muestra de la variable funcional* X_t *se define la media de la siguiente forma:*

$$\mu(t) = E[X_t] \tag{2.1}$$

Definición 2.5 *(Varianza Funcional). Sea la variable funcional* X_t *la varianza está determinada por:*

$$Var(t) = E[(X_t - \mu(t))^2] \tag{2.2}$$

La varianza brinda información sobre la variación de las curvas en cierto punto t*.*

Definición 2.6 *(Función de Covarianza). Se requiere información sobre la relación existente entre los valores de las curvas en* t *y los valores en otro punto* s*. Para ello, se define la función de covarianza bivariada:*

$$Cov(s,t) = E[(X_s - \mu(s))(X_t - \mu(t))] \tag{2.3}$$

Definición 2.7 *(Función de Correlación). La función de correlación está dada por:*

$$r(s,t) = \frac{Cov(s,t)}{\sqrt{Var(s)Var(t)}}. \tag{2.4}$$

2.2. MÉTODOS DE SUAVIZAMIENTO

El suavizamiento es el método por el cual se identifica cada observación de alta dimensión con una función. El objetivo de este método es construir los datos funcionales a partir de sus observaciones discretas.

Por lo general, en la práctica se encuentran datos donde cada observación está dada por un vector real cuya dimensión es muy alta, como ejemplo, el precio de una acción por hora de todo el año ya que no se puede almacenar la información continua del precio. Es decir, es complicado contar con un conjunto de funciones de manera continua en el tiempo.

Por lo cual, el problema es que en vez de contar con observaciones continuas $\{\chi_1, ..., \chi_n\}$, se empezará trabajando con observaciones discretas $\{x_1(t_{i0}), ..., x_n(t_{im_i})\}$ donde $t_{ik} \in [0,T]$ es el momento $1 \leq i \leq n$ y $1 \leq k \leq m_i$. Es decir, se contará con observaciones de tales funciones en diferentes momentos del tiempo $\{t_{i0}..., t_{im_i}\}$ y con un distinto número de observaciones para cada individuo. De este modo, se considera que la muestra está determinada por los vectores $x_i = (x_i(t_{i0}), ..., x_i(t_{im_i}))'$ donde x_{ik} es el valor observado de la trayectoria muestral en el instante t_{ik}.

Frente a ello, se propone reconstruir la forma funcional de las trayectorias muestrales $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ que describen el comportamiento de los datos discretos a partir de una representación en términos de bases funcionales. Estas funciones recuperadas son consideradas datos funcionales.

Definición 2.8 *(Base funcional). Un conjunto de funciones* $\{\phi_1, \phi_2, ...\}$ *es una base funcional en* $\mathcal{L}^2[0,T]$ *si toda función* $\chi \in \mathcal{L}^2[0,T]$ *tiene una descomposición*

$$\chi(t) = \sum_{j=1}^{\infty} c_{ij}\phi_j(t) \tag{2.5}$$

tal que $c_{ij} \in \mathbb{R} \ \ \forall i = 1, ..., n.$

En la práctica, se debe escoger una cantidad finita (κ_x) de funciones base, tal que sea suficientemente grande para que los errores sean tan pequeños como se desee. Es importante notar que para los datos observados no se cuenta con el continuo de puntos de χ. De manera que, si se toman en cuenta los puntos observados $x_i = \chi(t_{ik})$ para $i = 1, ..., n$ y $k = 1, 2, ...m_i$ entonces la descomposición se puede expresar por:

$$x_i = \chi(t_{ik}) = \sum_{j=1}^{\kappa_x} c_{ij}\phi_j(t_{ik}) + \epsilon(t_{ik})$$

Esta ecuación indica que el suavizamiento de cada observación se puede considerar como m problemas de regresión lineal donde $\chi(t_{ik})$ corresponde a la variable independiente y $\{\phi_j(t_{ik})\}_{j=1}^{\kappa_x}$ corresponden a las variables explicativas (Ramsay et al., 2009). De manera que, los coeficientes c_{ij} se pueden estimar por mínimos cuadrados

$$\hat{\mathbf{c}} = \arg\min \left[\sum_{k=1}^{m_i} \left(\chi(t_{ik}) - \sum_{j=1}^{L} c_{ij}\phi_j(t_{ik}) \right)^2 \right]$$

Por lo que, el dado funcional suavizado esta determinado por:

$$\hat{\chi}(t) = \sum_{j=1}^{\kappa_x} c_{ij}\phi_j(t) = \hat{\mathbf{c}}^T \boldsymbol{\phi} \tag{2.6}$$

Cabe recalcar que para escoger tanto la base funcional como el parámetro κ_x depende de las características y el comportamiento de los datos que se desean suavizar.

En Ramsay & Dalzell (1991) se estudian los tipos de bases como: Bases de Fourier para la suavización de datos periódicos y Bases B-Splines para la suavización de datos no periódicos. Además, existen otros tipos de bases funcionales útiles como las bases Wavelets, Exponenciales y Polinomiales.

2.2.1. BASES DE FOURIER

Esta base consiste en el conjunto de las funciones seno y coseno de frecuencia creciente y está constituida por los siguientes elementos:

$$B_n = \{1, sin(\omega t), cos(\omega t), sin(2\omega t), cos(2\omega t), ..., sin(k\omega t), cos(k\omega t) : k \in \mathbb{N}\}$$

con $\omega = \frac{2\pi}{p}$ donde P es el periodo de oscilación y n es el número de elementos de la base.

Mediante el uso de la base de Fourier se aproxima el dato funcional $\chi(t)$ a partir de datos discretos (Ramsay & Silverman, 2005). Esta aproximación se logra de acuerdo a la siguiente expresión:

$$\hat{\chi}(t) \approx \frac{a_0}{2} + \sum_{i=1}^{N} \left[a_i \cos\left(\frac{2\pi i t}{N}\right) + b_i \sin\left(\frac{2\pi i t}{N}\right) \right]$$

donde a_0, a_i y b_i son constantes $\forall i = 1, ..., N$.

A este procedimiento también se lo conoce como suavizamiento de una curva, pues elimina pequeños movimientos en los datos para que conserven la forma correcta. Este tipo de base, por lo general, se usa cuando las trayectorias a estimar son regulares y tienen naturaleza periódica. Sin embargo, tienden a ser muy suaves lo que implica que no son apropiadas para ajustar datos que tomen valores abruptos, es decir, no son útiles cuando los datos presenten algún grado de discontinuidad en las trayectorias a aproximar.

2.2.2. BASES B-SPLINES

Este tipo de funciones son las más utilizadas para aproximar series de datos no periódicos. Antes de describir una base B-spline se explicará brevemente que es un spline (Marx & Eilers, 1999).

Un spline es una función que se construye por secciones, es decir, un spline está formado por trozos de polinomios que se conectan entre sí mediante un nodo (De Boor, 1977). Al número de secciones se le suele designar como L. Para poder determinar una función spline se debe tener un número de valores que sea superior a $L + g + 1$ donde g es el grado del polinomio utilizado en cada intervalo. Se dice orden del spline al número $g + 1$ y al número $L + g + 1$ se denomina grados de libertad de la familia de splines.

Definición 2.9 *(Función Spline). Es una función $\phi : T \to \mathbb{R}$ tal que existe una partición $\{\tau_l\}_{l=1}^{L}$ de T y funciones polinomiales $\{\beta_k\}_{k=1}^{L}$ donde cada $\beta_k : \tau_k \to \mathbb{R}$ de grado m cumple las siguientes condiciones:*

1. $\phi|_{\tau_k} = \beta_k$

2. *Para todo par de polinomios adyacentes* β_j, β_i*, sus derivadas coinciden en el punto* $t \in T$ *tal que* $\phi(t) = \beta_j(t) = \beta_i(t)$ *para* $q = 1, ..., m-2$.

En Hastie & Mallows (1993) se plantea usar las funciones spline para construir los datos funcionales. El conjunto de funciones splines conforman un espacio vectorial cuya dimensión está dada por los grados de libertad, por lo que, se pueden utilizar diversas bases para generar dicho espacio.

Entonces, se representa cualquier función spline $\beta(t)$ de la siguiente manera:

$$\beta(t) = \sum_{i=0}^{N+k} c_i B_{i,k}(t),\ t \in [t_0, t_{N+1}]$$

donde los c_i son llamados puntos control o puntos de Boor. Para un B-spline de grado k con N nodos interiores existen $M = N + k + 1$ puntos control.

2.2.3. BASES WAVELETS

Las bases Wavelets son usadas como funciones básicas a la hora de representar los datos funcionales. Se usan sobre todo para representar curvas muestrales con un carácter local muy fuerte. Estas se obtienen mediante dilataciones y traslaciones de una función madre apropiada ϕ dada por:

$$\phi_{kj}(t) = 2^{\frac{k}{2}} \phi(2^k y - j)$$

con $j, k \in \mathbb{R}$. A diferencia de las bases de Fourier, las bases Wavelets no asumen que los datos sean periódicos. Es decir, tiene cierta ventaja sobre Fourier ya que esta última presenta una buena localización en frecuencia, pero no en tiempo; por su parte Wavelets tienen mejor localización en tiempo y frecuencia (Haar, 1910).

2.2.4. BASES EXPONENCIALES Y POTENCIALES

Las bases Exponenciales consisten en una serie de funciones dadas por:

$$\phi_j(t) = \exp\{\lambda_j t - w\}$$

donde $\lambda_1 = 0$, λ_j todos distintos y $w \in [0, T]$ es un parámetro a elegir (Ramsay & Silverman, 2005). La expansión en términos de bases exponenciales se halla en las ecuaciones diferenciales lineales con coeficientes constantes.

Las bases Potenciales consisten funciones compuestas por monomios que se utilizan para construir series de potencias dadas por:

$$\phi_j(t) = (t - w)^j$$

con $j \in \mathbb{R}$ y $w \in [0, T]$ es un parámetro a elegir (Ramsay & Silverman, 2005).

2.3. CRITERIO DE VALIDACIÓN CRUZADA

En general, los datos funcionales tienen alta dimensión por lo que es difícil de manejar, y este método reduce los datos funcionales a un espacio de dimensión finita κ_x que tiene un manejo más fácil.

En principio, no existe una regla universal que permita realizar una elección óptima del número de parámetros de la base. Además, el tipo de base apropiada para los datos observados debe ser escogido adecuadamente para cumplir el objetivo; por ejemplo, es común usar bases de Fourier para datos periódicos y bases B-splines para datos no recurrentes.

Para elegir un número razonable de bases para los datos funcionales mediante el parámetro $\nu_1 = \kappa_x$, se consideran los criterios de validación cruzada y validación cruzada generalizada donde se tiene un parámetro de penalización $\nu_2 = \lambda$ que se incluye en este proceso. Ambos criterios se definen a continuación:

- Validación Cruzada:

$$CV(\nu) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{r}^{\nu}(x_i))^2}{1 - S_{ii}} \omega(x_i)$$

  donde $\hat{r}^{\nu}$ es la predicción en el punto t_i obtenida al omitir el *i-ésimo* par (x_i, y_i), S_{ii} es el elemento diagonal *i-ésimo* de la matriz de suavizamiento S (con $\nu = traza(S)$) y $\omega(x_i)$ es el peso de los datos x en punto t_i.

- Validación Cruzada General:

$$GCV(\nu) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{r}^{\nu}(x_i))^2 \omega(x_i) \Xi(\nu)$$

  donde Ξ denota el tipo de función penalizante cuyos tipos son:

1) Validación cruzada generalizada *(GCV)*: $\Xi(\nu) = (1 - tr(S)n^{-1})^{-2}$
2) Criterio de información de Akaike *(AIC)*: $\Xi(\nu) = exp\,(2tr(S)n^{-1})$
3) Error de predicción finita *(FPE)*: $\Xi(\nu) = \frac{(1+tr(S)n^{-1})}{(1-tr(S)n^{-1})}$
4) Selector de modelo de Shibata *(Shibata)*: $\Xi(\nu) = (1 + 2tr(S)n^{-1})$
5) Selector de ancho de banda de Rice *(Rice)*: $\Xi(\nu) = (1 - 2tr(S)n^{-1})^{-1}$

Estos criterios se dan en Härdle (1990) y se aplican en Oviedo de la Fuente (2018).

2.4. MEDIDAS DE PROFUNDIDAD FUNCIONAL

Las medidas de profundidad se utilizan para resumir un conjunto de datos funcionales. El concepto de profundidad mide que tan central o profundo es un dato con respecto a una determinada población. Es decir, es una herramienta estadística que proporciona información acerca del orden interno de los datos, de modo que se clasifican del menos al más profundo.

De manera similar a la estimación puntual univariante clásica, la mediana sería el punto más profundo. Por supuesto, existen otras posibilidades, bien orientadas a lograr una mayor robustez o bien a tener en cuenta diferentes aspectos de la noción de "centralidad" para ordenar datos funcionales (Cuevas et al., 2007).

En primer lugar, se define el concepto de profundidad funcional, propuesto en Fraiman & Muniz (2001) y basado en la mediana funcional.

Definición 2.10 *(Profundidad de Fraiman-Muniz (FMD)). Sea $F_{n,t}$ la distribución empírica de la muestra $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ i.i.d. de la variable aleatoria funcional X_t. Para cada $t \in [0, T]$ la profundidad univariante $Z_i(t)$ de los datos $\{\chi_i(t)\}$ está dada por:*

$$Z_i(t) = 1 - \left|\frac{1}{2} - F_{n,t}(\chi_i(t))\right|$$

Luego, la profundidad FMD se define como el promedio de la profundidad univariante a lo largo del dominio del dato funcional:

$$FMD(\chi_i) = \int_0^1 Z_i(t)dt \;\; \forall i = 1, ..., n$$

La *FMD* coincidirá con los datos más profundos, es decir, se ordena la función $\chi_i(t)$ para la cual $FMD(\chi_i)$ es máxima (Fraiman & Muniz, 2001). De aquí, se deriva la definición de medidas robustas de localización que sirven para hallar el elemento que maximiza la profundidad. Para definir los siguientes conceptos se considera a $\chi_{(1)} = \chi_1$ como la curva más profunda y a $\chi_{(n)} = \chi_n$ como la curva menos profunda.

Definición 2.11 *(Media Recortada Funcional (FTM)). Se define la media recortada como la media de la mayor cantidad curvas más profundas ($n - \alpha n$) por:*

$$FTM_\alpha(\chi_i) = \frac{1}{n - \alpha n} \sum_{i=1}^{n-[\alpha n]} \chi_i$$

donde $0 < \alpha < (n-1)/n$ tal que $[\cdot]$ es la parte entera. De modo que, considera el promedio del $100(1-\alpha)\,\%$ de las funciones más profundas de la muestra. Es decir, elimina un porcentaje de datos menos profundos y estima la media con los restantes.

$$FMED_\alpha(\chi_i) = \chi_{(1)}$$

(Varianza Recortada Funcional). A partir de los conceptos anteriores se define una medida robusta para la varianza marginal dada por:

$$\hat{\sigma}^2_{TSD,\alpha} = \left(\frac{1}{n - \alpha n} \sum_{i=1}^{n-[\alpha n]} (\chi_i(t) - FTM_\alpha)^2 \right)^{1/2}$$

Ahora, en Cuevas et al. (2007) se usa una medida de profundidad basada en qué tan rodeadas están las curvas respecto a una distancia (métrica o semi-métrica), seleccionando la trayectoria más densamente rodeada por otras trayectorias del proceso.

Definición 2.12 *(Profundidad Modal (MD)). Sea la muestra $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ i.i.d. de la variable aleatoria funcional X_t. Se define el núcleo por:*

$$\begin{aligned} K : \mathbb{R} &\longrightarrow \mathbb{R} \\ t &\longmapsto K(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{-t^2}{2} \right\} \end{aligned}$$

La h-profundidad sobre una vecindad z se determina mediante:

$$MD_h(z) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\|z - x_i\|}{h}\right)$$
$$= n^{-1}\sum_{i=1}^{n} K_h\left(\|z - x_i\|\right)$$

donde $K_h(t) = h^{-1}K(\frac{t}{h})$ es el núcleo reescalado y h es un parámetro de ajuste.

En Cuevas et al. (2007) también considera la medida de profundidad calculada a través de Proyecciones Aleatorias *(RP)*.

Definición 2.13 *(Profundidad sobre Proyecciones Aleatorias (RPD)). Sea la muestra $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ i.i.d. de la variable aleatoria funcional X_t. Sea $h \in \mathcal{H}$ una dirección del proceso de dirección independiente $\mathcal{H}$. La proyección $\{\chi_i\}$ a lo largo de la dirección h está dada por:*

$$P_i^h = \langle h, \chi_i \rangle = \int_0^T h(t)\chi_i(t)dt$$

Se define la RPD como:

$$RPD(\chi_i, h) = \mathcal{D}(P_i^h)$$

donde $\mathcal{D}$ es la medida de profundidad univariante. Si se considera una colección de proyecciones aleatorias $\{h_j\}_{j=1}^m$ la profundidad se obtiene usando todas las proyecciones

$$RPD(\chi_i, \{h_j\}_{j=1}^m) = \frac{1}{m}\sum_{j=1}^{m} \mathcal{D}(P_i^{h_j})$$

Por otra parte, en Cuesta-Albertos & Nieto-Reyes (2008) se presenta una variante de la profundidad de proyecciones aleatorias.

Definición 2.14 *(Profundidad Aleatoria de Tukey (RTD)). Se define como una variante de RPD dada por:*

$$RTD(\chi_i, \{h_j\}_{j=1}^m) = \min \mathcal{D}(P_i^{h_j})$$

Estas medidas son estimadores orientados a tomar en cuenta la tendencia central del proceso, por lo cual se verifica el cumplimiento de cada uno de ellos si se

considera que el objetivo es la función media $\{\mu(t) = E[X_t]\}$ (Cuesta-Albertos & Nieto-Reyes, 2008).

2.5. BOOTSTRAP PARA BANDAS DE CONFIANZA

Para evaluar la precisión de un estimador de localización como los vistos en la sección anterior se calculan las Bandas de Confianza *(BC)*. El procedimiento de suavizamiento Bootstrap es la metodología de remuestreo para datos funcionales que permite calcular las *BC* para los estimadores, de modo que este procedimiento llena los huecos en el espacio funcional teniendo en cuenta la estructura de covarianza en el suavizamiento (Cuevas et al., 2007).

Las *BC* son aquellos remuestreos que están a cierta distancia del estimador. Así, el suavizamiento bootstrap se construyen a partir de la información de muestreo que viene dada por datos funcionales $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ obtenidas como observaciones i.i.d. de un proceso estocástico X_t, con trayectorias continuas en $[0, T]$. El procedimiento para la construcción de las *BC* se tiene a continuación:

1) En primer lugar, se calcula b muestras bootstrap a partir de $\mathcal{S}_n$ de donde se obtiene $\mathcal{S}_n^{*j} = \{\chi_1^*(t), ..., \chi_n^*(t)\}$ con $j = 1, ..., b$ tal que

$$\chi_i^*(t) = \chi_i(t) + Z(t)$$

donde $Z(t)$ se distribuye normalmente con media 0 y matriz de covarianza $\gamma\Sigma_{\mathcal{S}_n}$, siendo $\Sigma_{\mathcal{S}_n}$ la matriz de covarianza de $\mathcal{S}_n$ y γ es el parámetro de suavizado.

2) Se toma en cuenta las muestras $\mathcal{S}_n$ y $\mathcal{S}_n^{*j}$ para calcular las estimaciones $\hat{\sigma}(\mathcal{S}_n)$ y $\hat{\sigma}(\mathcal{S}_n^{*j}$ respectivamente. Es decir, obtiene el estimador $T(\mathcal{S}_n^{*j})$ definido a través del cálculo del valor $D(\mathcal{S}_n)$.

3) Luego, usando la norma $\mathcal{L}^2$ se obtiene las distancias $d\left(\hat{\sigma}(\mathcal{S}_n), \hat{\sigma}(\mathcal{S}_n^{*j})\right)$.

4) Por último, se calcula la *BC* correspondiente a un nivel de confianza $(1-\alpha)\,\%$ de las réplicas bootstrap $T(\mathcal{S}_n^{*j})$ que están a una distancia $d\left(\hat{\sigma}(\mathcal{S}_n), \hat{\sigma}(\mathcal{S}_n^{*j})\right)$ de su promedio menor que $D(\mathcal{S}_n)$.

De esta manera, se representa las *BC* con nivel de significancia $\alpha = 5\,\%$. El estadístico utilizado suele ser la media funcional pero también se pueden usar las funciones basadas en la profundidad.

2.6. DETECCIÓN DE DATOS ATÍPICOS FUNCIONALES

En general en un análisis exploratorio la detección de los valores atípicos es importante. Este tipo de valores pueden sesgar los resultados de los modelos que se ponen a prueba, obteniéndose curvas atípicas. En Febrero-Bande et al. (2010) se considera una relación inversa entre la profundidad y lo atípico para identificar valores atípicos en conjuntos de datos funcionales, pues la curva correspondiente tendrá una profundidad significativamente baja. Bajo este criterio, las curvas con profundidades más bajas serán detectadas como atípicas.

Para realizar la detección de datos atípicos funcionales, se sugiere seguir el siguiente procedimiento:

1) Se obtiene las profundidades funcionales $D_n(\chi_i), ..., D_n(\chi_n)$ mediante las medidas de profundidad definidas anteriormente como: FMD, MD, RPD o RTD.

2) Si $\{\chi_{i_1}, ..., \chi_{i_k}\}$ son la k curvas tal que $D_n(\chi_{i_k}) \leq C$, para un punto de corte dado C. Entonces si consideramos a $\{\chi_i, ..., \chi_n\}$ como atípicos, se los elimina de la muestra.

3) Volver al paso 1 con el nuevo conjunto de datos después de eliminar los valores atípicos encontrados en el paso 2. Se repite esto hasta que no se encuentren más valores atípicos.

El procedimiento para seleccionar C de modo que, en ausencia de valores atípicos: $Pr(D_n(\chi_i) \leq C) = \alpha\ , i = 1, ..., n$, por defecto $\alpha = 0.05$. Por tanto, el C tomado es el percentil α de la distribución de la profundidad funcional considerada.

2.7. COMPONENTES FUNCIONALES

2.7.1. COMPONENTES PRINCIPALES FUNCIONALES (FPC)

El Análisis de Componentes Principales Funcionales *(FPCA)* es una extensión del Análisis de Componentes Principales *(PCA)* del caso multivariado. Donde el objetivo general es generar un nuevo conjunto de variables no correlacionadas a partir de las variables originales. Es decir, se requiere observar los atributos principales que caracterizan a las observaciones funcionales χ_i de manera que se pueda reducir la dimensionalidad de los datos para comprender el comportamiento de estos utilizando una menor cantidad de parámetros.

En el caso multivariado las componentes principales *(PC)* corresponden a las direcciones ortogonales que explican mayor parte de la varianza de las observaciones. Mientras que en el caso funcional para determinar las componentes principales funcionales *(FPC)* se debe hallar una descomposición ortogonal para la función de covarianza bivariada. Para continuar con el *FPCA* es conveniente enunciar el siguiente lema y teorema.

Lema 2.1 *(Lema de Mercer). Sea la función de covarianza dada en* (2.3)*. Entonces existe una sucesión* $\{\psi_n\}$ *de funciones continuas y una sucesión decreciente* $\{\lambda_n\}$ *positiva tal que*

$$\int_0^T Cov(s,t)\psi_j(t)dt = \lambda_j\psi_j(t)$$

Además,

$$Cov(s,t) = \sum_{j=1}^{\infty} \lambda_j\psi_j(t)\psi_k(t)$$

donde la serie converge de forma uniforme, por lo tanto

$$\sum_{i=1}^{\infty} \lambda_i = \int Cov(t,t)dt < \infty$$

Teorema 2.2 *(Expansion de Karhunen-Loeve). Sea* $X_t = \{\mathcal{X}(t)\}_{t\in[0,T]}$ *un proceso estocástico cuadrado integrable de media cero definido sobre un espacio de probabilidad, con función de covarianza continua. Entonces*

$$X_t = \sum_{i=1}^{\infty} \eta_j\psi_j(t)$$

donde η_j es una sucesión real de variables aleatorias de media cero tal que

$$E[\eta_j \eta_i] = \lambda_j \delta_{ij}$$

y donde la secuencia (λ_j, ψ_j) es definida por el lema (2.1)*. La serie converge uniformemente en $[0, T]$.*

Ahora, para hallar las *FPC* se considera la variable funcional $X_t = \{\mathcal{X}(t)\}_{t\in[0,T]}$ que es un proceso estocástico de segundo orden, del cual se dispone de una muestra aleatoria $\mathcal{S}_n\{\chi_i(t)\}_{i=1}^n$ de n funciones independientes cuadrado integrable, cuya media es constante e igual a cero, $\mu(t) = 0\ \forall t \in T$. En el caso, de un proceso no centrado se debería considerar $\{X_t - \mu(t)\}$.

Se procede a definir los operadores que sirven para la determinar los componentes dominantes de la función (2.3) por medio de una descomposición ortogonal como se explica en Ramsay & Dalzell (1991).

Definición 2.15 *(Operador Lineal). Se define el siguiente operador:*

$$\begin{aligned} \mathcal{C} : \mathcal{L}^2[0,T] &\longrightarrow \mathbb{R}^p \\ e(t) &\longmapsto f_i = \langle e, \chi_i \rangle = \int_0^T \chi_i(t) e(t) dt \end{aligned} \tag{2.7}$$

donde f_i es la i-ésima coordenada del vector $f \in \mathbb{R}^p$.

De este operador se obtiene un vector $f \in \mathbb{R}^p$ a partir de una función $e(t) \in \mathcal{L}^2[0,T]$ cuya componente i-ésima mide el ángulo que forma entre $e(t)$ y $\chi_i(t)$. Si $e(t)$ es ortogonal a algunas de las $\chi_i(t)$ las componentes correspondientes serán nulas.

Definición 2.16 *El operador adjunto $\mathcal{C}^\dagger$ se define de la siguiente manera:*

$$\begin{aligned} \mathcal{C}^\dagger : \mathbb{R}^p &\longrightarrow \mathcal{L}^2[0,T] \\ t &\longmapsto e(t) = \mathcal{C}^\dagger f = \sum_i^n \chi_i(t) f_i \end{aligned} \tag{2.8}$$

Este operador obtiene una curva del espacio $\mathcal{L}^2[0,T]$ que viene dada como una combinación lineal de las $\chi_i(t)$ ponderados por las componentes de f.

El *FPCA* requiere una sucesión de curvas $\{\xi_j(t)\}_{j=1}^{p}$ con $\|\xi_j(t)\| = 1$ del espacio $\mathcal{L}^2[0,T]$, cuya transformación dada por el operador (2.7) tenga módulo máximo y sean ortogonales entre sí. De esta manera, se tiene el siguiente problema:

$$\begin{aligned} \underset{\xi_j}{\text{máx}} \quad & \langle \mathcal{C}\xi_j, \mathcal{C}\xi_j \rangle \\ \text{s.a.} \quad & \|\xi_j\|^2 = 1 \\ & \langle \xi_j, \xi_k \rangle = 0 \ \forall k < j. \end{aligned} \tag{2.9}$$

De aquí, se obtiene el operador endomórfico resultante de la composición entre los operadores (2.7) y (2.8).

Definición 2.17 *(Operador de Covarianza). Se define mediante:*

$$\begin{aligned} V = \mathcal{C}^\dagger \circ \mathcal{C} : \mathcal{L}^2[0,T] &\longrightarrow \mathcal{L}^2[0,T] \\ e(t) &\longmapsto V(e(t)) = \int_0^T \left[\sum_{i=1}^{n} \chi_i(t)\chi_i(s) \right] e(s)ds \end{aligned} \tag{2.10}$$

De esta manera, el operador de covarianza (2.10) tiene la siguiente estructura:

$$V(e(t)) = \int_0^T K(s,t)e(s)ds,$$

donde $K(s,t)$ es el núcleo del operador V que coincide con la $Cov(s,t)$ definida en (2.3). Se nota que $\frac{K(s,t)}{(n-1)}$ es la covarianza intrapuntos del conjunto de funciones que determinan $\mathcal{L}^2[0,T]$.

Ahora, el problema (2.9) puede expresarse como:

$$\begin{aligned} \underset{\xi_j}{\text{máx}} \quad & \langle \xi_j, \mathcal{C}^\dagger \mathcal{C}\xi_j \rangle = \langle \xi_j, V\xi_j \rangle \\ \text{s.a.} \quad & \|\xi_j\|^2 = 1 \\ & \langle \xi_j, \xi_k \rangle = 0 \ \ \forall k < j. \end{aligned}$$

Por el Lema de Mercer (2.1) se realiza una descomposición ortogonal del núcleo de la siguiente forma:

$$K(s,t) = Cov(s,t) = \sum_{j=1}^{p} \lambda_j \xi_j(t)\xi_j(s)$$

siendo $\{\xi_j(t)\}_{j=1}^{p}$ la sucesión de p funciones propias de Cov que constituyen una base ortonormal en $\mathcal{L}^2[0,T]$. Éstas están asociadas a la sucesión decreciente $\{\lambda_j\}_{j=1}^{p}$

de valores propios que indican la cantidad de varianza que se le atribuye a cada componente principal.

Por otro lado, el teorema (2.2) permite solventar el problema del infinito usando el desarrollo de Karhunen-Loeve truncado de orden n para expresar cualquier curva muestral $\chi_i(t)\ \forall i = 1, ..., n$ del proceso X_t de la siguiente manera:

$$\chi_i = \sum_{j=1}^{p} \xi_j(t) f_{ij} \tag{2.11}$$

donde el conjunto $\{\xi_j\}_{j=1}^{p}$ son las componentes principales funcionales y forman un sistema ortonormal de $\mathcal{L}^2[0.T]$ y $\{f_{ij}\}$ son los puntajes o score de las componentes principales no correlacionadas con media cero estimadas mediante:

$$f_{ij} = \int_0^T \xi_j(t)\chi_i(t)dt$$

De donde se tiene la cantidad relativa que existe en la componente j para la observación i. Y representan la proyección de χ_i sobre la *j-esima* autofunción $\xi_j(t)$.

Las soluciones al problema (2.9) son las autofunciones del operador V definido en (2.10). Estas autofunciones se obtienen resolviendo la ecuación integral de Fredholm:

$$\begin{aligned} (V\xi_j)(t) \ &= \int_0^T Cov(s,t)\xi_j(t)dt = \langle Cov(s,\cdot), \xi_j\rangle = \lambda_j\xi_j(s) \\ \text{s.a.} \qquad & \|\xi_j\|^2 = 1 \\ & \langle \xi_j, \xi_k\rangle = 0\ \forall k < j. \end{aligned} \tag{2.12}$$

Se considera que $\lambda_j = Var[\xi_j]$ y la sucesión $\{\xi_j\}_{j=1}^{p}$ es la base ortonormal de $\mathcal{L}^2[0,T]$. Para resolver el problema planteado en (2.12) se toma en cuenta la representación en bases funcionales (2.6) de cada dato funcional dada por:

$$\chi_i(t) = \mathbf{c}_i\boldsymbol{\phi}^T.$$

Además, se determina una representación en la misma base $\{\phi_j\}_{j=1}^{L}$ para las componentes principales. Para ello, se define los coeficientes $\mathbf{d}_j \in \mathbb{R}^L$ tal que

$$\xi_j = \mathbf{d}_j\boldsymbol{\phi}^T\ \forall j.$$

De esta manera, asumiendo que los datos están centrados y teniendo en cuenta la definición de $Cov(s,t)$ en (2.3), el problema de encontrar las componentes principales se convierte en solucionar:

$$\int_0^T Cov(s,t)\xi_j(t)dt = \lambda_j\xi_j(s)$$
$$\int_0^T \frac{1}{n-1}\sum_{i=1}^{n}\chi_i(s)\chi_i(t)\xi_t = \lambda_j\xi_j(s)$$
$$\int_0^T \frac{1}{n-1}\sum_{i=1}^{n}[\mathbf{c}_i^T\boldsymbol{\phi}(s)][\boldsymbol{\phi}^T(t)\mathbf{c}_i][\mathbf{d}_j^T\boldsymbol{\phi}(t)]dt = \lambda_j\mathbf{d}_j^T\boldsymbol{\phi}(s)$$
$$\left(\frac{1}{n-1}\sum_{i=1}^{n}\mathbf{c}_i\mathbf{c}_i^T\right)\left(\int_0^T \boldsymbol{\phi}(s)\boldsymbol{\phi}^T(t)dt\right)\mathbf{d}_j^T\boldsymbol{\phi}(s) = \lambda_j\mathbf{d}_j^T\boldsymbol{\phi}(s)$$
$$\left(\frac{1}{n-1}\sum_{i=1}^{n}\mathbf{c}_i\mathbf{c}_i^T\right)\left(\int_0^T \boldsymbol{\phi}(s)\boldsymbol{\phi}^T(t)dt\right)\mathbf{d}_j^T = \lambda_j\mathbf{d}_j^T$$
$$\mathbf{C}\Phi\mathbf{d}_j^T = \lambda_j\mathbf{d}_j^T$$

De donde

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{c}_i\mathbf{c}_i^T \text{ y } \Phi = \int \phi(s)\phi^T(t)dt$$

Así, el problema de solucionar una ecuación de funciones y valores propios para determinar *FPC* se transforma en un problema de encontrar vectores y valores propios en un contexto vectorial sujeto a condiciones de ortogonalidad.

La gran ventaja que nos proporciona el *FPCA* es que permite reducir un problema infinito dimensional a un conjunto de curvas finitas, resultado de realizar combinaciones lineales con las componentes principales.

2.7.2. MÍNIMOS CUADRADOS PARCIALES FUNCIONALES (FPLS)

Las componentes principales funcionales *FPC* son una buena solución para representar los datos en una dimensión reducida espaciada (Martens & Naes, 1991). Pero cuando la variable funcional X_t se relaciona con una variable escalar Y, se puede utilizar directamente la relación existente mediante Mínimos Cuadrados Parciales Funcionales *(FPLS)*.

El enfoque *PLS* ofrece una buena alternativa al método *PCA* al reemplazar el criterio de mínimos cuadrados por la covarianza máxima entre un proceso X_t y una variable

escalar $Y = (y_1, ..., y_n)$ (Preda & Saporta, 2005). Es decir, la idea básica es construir un conjunto de componentes *PLS* funcionales en el espacio lineal generado por X_t, teniendo en cuenta la correlación entre un proceso X_t y una variable escalar Y.

En este caso, se asume que el proceso de segundo orden $X_t = \{\mathcal{X}(t)\}_{t\in[0,T]}$ dispone de una muestra aleatoria de n funciones independientes $\mathcal{S}_n = \{\chi_i(t)\}_{i=1}^n$ cuadrado integrable y centrado. Sea $Y = (y_1, ..., y_n)$ una variable aleatoria real. Sin pérdida de generalidad, se considera que $E[y_i] = 0$ y $E[X_t] = 0,\ \forall t \in T$.

Para encontrar las componentes *FPLS* se introducen los siguientes operadores que permitirán maximizar la covarianza entre X_t e Y.

Definición 2.18 *(Operador de Covarianza Cruzada). Se define el operador de varianza cruzada C_{yx} por:*

$$\begin{aligned} C_{yx} : \mathcal{L}^2[0,T] &\longrightarrow \mathbb{R}^p \\ e(t) &\longmapsto z_i = \int_0^T Cov(X_t, y_i)e(t)dt \\ e(t) &\longmapsto z_i = \int_0^T E[X_t \cdot y_i]e(t)dt \end{aligned} \tag{2.13}$$

De manera análoga, se define el operador adjunto C_{xy}.

Definición 2.19 *El operador adjunto C_{xy} se define de la siguiente manera:*

$$\begin{aligned} C_{xy} : \mathbb{R}^p &\longrightarrow \mathcal{L}^2[0,T] \\ z_i &\longmapsto e(t) = Cov(X_t, y_i) \cdot z_i \\ z_i &\longmapsto e(t) = \sum_{i=1}^p E[X_t \cdot y_i]z_i \end{aligned} \tag{2.14}$$

La metodología *PLS* consiste en penalizar el criterio de mínimos cuadrados maximizando la covarianza en lugar del coeficiente de correlación (Tenenhaus, 1998). Para obtener las componentes *FPLS* se utiliza Criterio de Tucker dado por:

$$\begin{aligned} \underset{\kappa\in\mathcal{L}^2[T]}{\text{máx}} \quad & Cov^2\left(\int_0^T X_t\kappa(t)dt, \sum_{i=1}^p c_i y_i\right) \\ \text{s.a.} \quad & \|\kappa\|^2 = 1 \\ & \|c\|^2 = 1. \end{aligned} \tag{2.15}$$

De donde se tiene que

$$\kappa(t) = \frac{Cov(X_t, y)}{\|Cov(X_t, y)\|} = \frac{E[X_t \cdot y]}{\sqrt{\int_0^T E^2[X_t \cdot y]}} \tag{2.16}$$

El hecho de que los componentes *PLS* son soluciones al criterio de Tucker fue descubierto posteriormente por Stone & Brooks (1990).

Ahora, se define el siguiente operador mediante la composición de los operadores (2.13) y (2.14).

Definición 2.20 *(Operador Composición). Se define el operador mediante la composición de los operadores anteriormente definidos.*

$$\begin{aligned} \mathcal{U} = C_{xy} \circ C_{yx} : \mathcal{L}^2[0,T] &\longrightarrow \mathcal{L}^2[0,T] \\ z(t) &\longmapsto U(z(t)) = \int_0^T \left[E(X_t \cdot z(t))\right] X_t dt \end{aligned} \tag{2.17}$$

De manera que, el problema (2.15) se puede reescribir como:

$$\begin{aligned} \underset{\kappa_h \in \mathcal{L}^2[T]}{\text{máx}} \quad & \frac{\langle \mathcal{U}\kappa, \kappa \rangle}{\langle \kappa, \kappa \rangle} \\ \text{s.a.} \quad & \|\kappa\|^2 = 1 \end{aligned}$$

Cuyas soluciones son las autofunciones del operador $\mathcal{U}$. Entonces, la solución del problema (2.15) es la función propia del operador $\mathcal{U}$ asociada a su valor propio más grande λ_1 está dada por

$$\mathcal{U}\kappa_1 = \lambda_1 \kappa_1$$

tal que, la primera componente *FPLS* está definida como

$$\eta_1 = \int_0^T \kappa_1(t)\chi_i(t)dt \tag{2.18}$$

Una vez determinadas las componentes *FPLS* se puede expresar cualquier curva muestral $\mathcal{S}_n = \{\chi_i(t)\}_{i=1}^n$ del proceso X_t y cualquier y_i de Y utilizando el teorema Karhunen-Loeve (2.2) truncado de orden n de la siguiente manera:

$$\begin{aligned} \chi_i &= \sum_{h=1}^{p} p_h(t)\eta_h \\ y_i &= \sum_{h=1}^{p} c_h(t)\eta_h \end{aligned} \tag{2.19}$$

Puesto que, la forma de obtener las componentes *FPLS* es recursiva. Sean χ_h y y_h los residuos de las regresiones lineales dadas por:

$$\begin{aligned} \chi_h(t) &= \chi_{h-1}(t) - p_h(t)\eta_h \\ y_h(t) &= y_{h-1}(t) - c_h(t)\eta_h \end{aligned} \tag{2.20}$$

De donde

$$\begin{aligned} p_h(t) &= \frac{Cov(\chi_{h-1}, \eta_h)}{Var(\eta_h)} = \frac{E[\chi_{h-1}\eta_h]}{E[\eta_h^2]} \\ c_h(t) &= \frac{Cov(y_{h-1}(t)\eta_h)}{Var(\eta_h)} = \frac{E[y_{h-1}\eta_h]}{E[\eta_h^2]} \end{aligned} \tag{2.21}$$

Se obtiene un conjunto $\{\eta_h\}_{h=1}^{p}$ de componentes $FPLS$ usando un procedimiento iterativo (Apostol & Pedra, 2010). En cada paso, la componente PLS se define como la combinación lineal de cualquier curva muestral $\chi_h(t)$ del proceso X_t obtenida mediante las funciones κ_h que alcanza la máxima covarianza. De esta manera, se define la *h-ésima* componente *FPLS* por:

$$\eta_h = \int_0^T \kappa_h(t)\chi_{h-1}(t)dt \tag{2.22}$$

donde se tiene que

$$\kappa_h(t) = \frac{Cov(\chi_{h-1}(t), y_{h-1})}{\|Cov(\chi_{h-1}(t), y_{h-1})\|} \tag{2.23}$$

De manera que, el problema concluye con la solución de la siguiente ecuación:

$$\mathcal{U}_{h-1}\kappa_h = \lambda_h\kappa_h.$$

Este proceso además de reducir un problema infinito dimensional proporciona la ventaja de obtener directamente las componentes sin necesidad de realizar combinaciones lineales entre las componentes, al determinar las componentes *FPLS* por medio de vectores y valores propios en un contexto vectorial sujeto al problema de maximización de la covarianza.

CAPÍTULO 3

MODELOS DE REGRESIÓN FUNCIONAL CON RESPUESTA ESCALAR

En general los modelos de regresión se utilizan para estudiar la relación entre una o más covariables explicativas (independientes) y una variable de respuesta (dependiente). En este trabajo, se tiene la finalidad de modelar una variable de respuesta escalar a partir de covariables explicativas funcionales.

Por estas razones, se analizará teóricamente los Modelo de Regresión Lineal Funcional *(FLR)* con Respuesta Escalar tales como: el modelo *FLR* con Representación en Bases detallado en Ramsay et al. (2009), el modelo *FLR* con Base por Componentes Principales Funcionales *(FPC)* propuesto por Cardot & Sarda (2006), el modelo *FLR* con Base por Mínimos Cuadrados Parciales Funcionales *(FPLS)* planteado en Preda & Saporta (2005) y la adaptación del modelo *FLR* con Representación en Bases usando dos covariables funcionales.

3.1. MODELO LINEAL FUNCIONAL CON UNA COVARIABLE FUNCIONAL

Este modelo *FLR* clásico busca estimar la relación entre variable de respuesta escalar Y y una covariable funcional única $X_t = \{\mathcal{X}(t)\}_{t\in[0,T]}$ que es un proceso estocástico de segundo orden. Sin pérdida de generalidad, se considera que $E[Y] = 0$ y $E[X_t] = 0\ \forall t \in [0,T]$. En el caso, de un proceso no centrado se debería usar $\{X_t - \mu(t)\}$. De tal manera que, el modelo *FLR* con respuesta escalar Y y covariable funcional explicativa X_t está determinado por:

$$Y = \alpha + \int_0^T X_t\beta(t)dt + \epsilon$$

En otras palabras, sea $\{y_i\}_{i=1}^n$ las respuestas escalares correspondientes a Y y las curvas muestrales $\mathcal{S}_n = \{\chi_i(t)\}_{i=1}^n$ en $\mathcal{L}^2[0,T]$. El modelo *FLR* se puede expresar mediante:

$$y_i = \alpha + \int_0^T \chi_i(t)\beta(t)dt + \epsilon_i \tag{3.1}$$

donde $\beta(t) \in \mathcal{L}^2[0,T]$ es el parámetro funcional y $\epsilon_i \sim N(0,\sigma^2)$ es el residual tal que $Cov(\epsilon_i, \epsilon_j) = 0 \; \forall i \neq j$.

A continuación, se plantean los modelos más importantes para conseguir la estimación buscada.

3.1.1. MODELO DE REGRESIÓN LINEAL FUNCIONAL (FLR) CON REPRESENTACIÓN EN BASES

Dado el modelo de regresión (3.1) se propone estimar la función $\beta(t)$ por medio una base funcional B-spline lo suficientemente grande. Para ello, se considera la base funcional $\{\theta_j\}_{j=1}^{\kappa_B}$ tal que

$$\hat{\beta}(t) = \sum_{j=1}^{\kappa_B} b_j \theta_j(t) = \boldsymbol{\theta}^T \mathbf{b}$$

Además, se toma en cuenta el suavizamiento de los datos funcionales χ_i dado por

$$\chi_i(t) = \sum_{j=1}^{\kappa_x} c_{ij} \phi_j(t) = \mathbf{c}_i^T \boldsymbol{\phi}$$

Si se reemplaza estas representaciones en el modelo (3.1) se obtiene que

$$\begin{aligned} y_i &= \alpha + \int_0^T [\mathbf{c}_i^T \boldsymbol{\phi}][\boldsymbol{\theta}^T(t)\mathbf{b}]dt + \epsilon_i \\ &= \alpha + \mathbf{c}_i^T \left[\int_0^T \boldsymbol{\phi}(t)\boldsymbol{\theta}^T(t)dt\right] \mathbf{b} + \epsilon_i \\ &= \alpha + \mathbf{c}_i^T [J_{\boldsymbol{\phi\theta}}]\mathbf{b} + \epsilon_i \end{aligned}$$

tal que $J_{\phi\theta}$ es la matriz de productos internos dados por $\boldsymbol{\phi}(t)$ y $\boldsymbol{\theta}(t)$ de tamaño $(\kappa_x \times \kappa_B)$. Así, se considera la matriz $Z_i = [\mathbf{c}_i^T J_{\phi\theta}]$ y el vector $\mathbf{b} = (b_1, ..., b_{\kappa_B})$ de tal modo que se reduce el problema de la siguiente forma:

$$y_i = \alpha + Z_i \mathbf{b} + \epsilon_i$$

Este modelo de regresión permite estimar α y los coeficientes $\mathbf{b}_j$ de $\beta(t)$ mediante mínimos cuadrados obteniendo así la estimación

$$\hat{\mathbf{b}} = (Z^T Z)^{-1} Z^T Y$$

Esta metodología trae consigo la dificultad de determinar la inversa de $(Z^T Z)$, ya que la dimensión de Z está dada por el número de elementos de la base κ_B elegida

para la estimación $\hat{\beta}(t)$. De aquí, si se aumenta demasiado el número de bases κ_B provoca una sobreparametrización en el modelo y en el caso de reducir tanto el número de bases κ_B se pierde información en el modelo, es decir, el modelo no es suficientemente flexible.

Para afrontar estas eventualidades Ramsay & Silverman (2005) y Cardot & Sarda (2006) proponen considerar una base lo suficientemente grande pero aplicando una penalización sobre el comportamiento de la función de regresión para evitar fluctuaciones locales que puedan generarse de forma excesiva. De esta forma, considerando el modelo (3.1), se requiere encontrar (α, β_i) tales que minimicen la siguiente expresión:

$$\sum_{i=1}^{n}\left[y_i - \alpha - \int_0^T \chi_i(t)\beta(t)dt\right]^2 + \lambda\int_0^T\left[\frac{d^2}{dt^2}(\beta(t))\right]^2 dt$$

donde $\lambda \in \mathbb{R}$ es un parámetro cuya selección se realiza por validación cruzada. En este modelo con penalización se puede estimar los coeficientes $\beta_i(t)$ mediante

$$\hat{\mathbf{b}} = (Z^T Z - \lambda R)^{-1} Z^T Y$$

tal que se tiene la matriz $R = \int_0^T \left[\frac{d^2}{dt^2}(\beta(t))\right]^2 dt$.

En este trabajo, para el caso de estudio sobre la estimación de la Precipitación se plantea utilizar el modelo estudiado a partir de la Temperatura como covariable explicativa funcional.

Por lo que, se plantea el siguiente problema:

$$y_i = c_0 + \int_0^T \chi_i(t)\beta(t)dt + \epsilon_i \tag{3.2}$$

donde $\{y_i\}_{i=1}^n$ es la muestra de la variable escalar Y representada por la Precipitación y la muestra $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ corresponde a la Temperatura como covariable explicativa funcional. De manera análoga, se realizará el mismo proceso usando la Velocidad del Viento como covariable explicativa funcional.

3.1.2. MODELO DE REGRESIÓN LINEAL FUNCIONAL (FLR) CON BASE POR COMPONENTES PRINCIPALES FUNCIONALES

En la Sección 2.7.1 se estudiaron las componentes *FPC* de la muestra $\mathcal{S}_n = \{\chi\}_{i=1}^n$

del proceso $X_t = \{\mathcal{X}(t)\}_{t\in[0,T]}$ que son combinaciones lineales dadas por las autofunciones del operador de covarianza y constituyen una base ortonormal de $\mathcal{L}^2$.

Este método permite asociar un modelo de regresión multivariado con un modelo de regresión funcional mediante el uso de los puntajes de las componentes *FPC*. A partir de las observaciones funcionales $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ y de la función de covarianza $Cov(s,t)$ dada en (2.3) se utilizan las primeras componentes principales funcionales $\{\xi_j(t)\}_{j=1}^p$, de modo que se tiene la descomposición (2.11) dada por

$$\chi_i(t) = \sum_{j=1}^{p} \xi_j(t) f_{ij}$$

El modelo asociado a la regresión funcional (3.1) se expresa de la siguiente forma:

$$y_i = b_0 + \sum_{j=1}^{p} b_j(t) f_{ij} + \epsilon_i$$

Este modelo es un modelo de regresión múltiple multivariado estándar que se resuelve por mínimo cuadrados. Así, la estimación de $\mathbf{b} = (b_1, ...b_p)$ se obtiene por

$$\hat{\mathbf{b}} = (F^T F)^{-1} F^T Y$$

donde $F = \{f_{ij}\}_{nxp}$. Puesto que, las componentes *FPC* forma un sistema ortogonal, se escoge

$$\beta_{FPC}(t) = \sum_{j=1}^{p} b_j \xi_j(t)$$

de manera que se puede llegar al modelo (3.1) mediante:

$$y_i = b_0 + \int_0^T [\chi_i(t)][\sum_{j=1}^{p} b_j \xi_j(t)]dt + \epsilon_i$$

Así, después de realizar la regresión funcional por medio de componentes *FPC*, se recupera la función de regresión funcional estimada, determinada por

$$\hat{\beta}_{FPC}(t) = \sum_{j=1}^{p} \hat{b}_j \xi_j(t)$$

En este trabajo, para el caso de estudio sobre la estimación de la Precipitación se plantea utilizar las bases determinadas por las componentes *FPC* del modelo estudiado a partir de la Temperatura como covariable explicativa funcional.

Por lo que, se plantea el siguiente problema:

$$y_i = c_0 + \int_0^T [\chi_i(t)][\beta_{FPC}(t)]dt + \epsilon_i \tag{3.3}$$

donde $\{y_i\}_{i=1}^n$ es la muestra de la variable escalar Y representada por la Precipitación y la muestra $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ corresponde a la Temperatura como covariable explicativa funcional. De igual forma, se realizará el mismo proceso usando la Velocidad del Viento como covariable explicativa funcional.

3.1.3. MODELO DE REGRESIÓN LINEAL FUNCIONAL (FLR) CON BASE POR MÍNIMOS CUADRADOS PARCIALES FUNCIONALES

Este proceso se basa en descomponer la variable respuesta y la covariable funcional en términos de variables aleatorias, de modo que se maximice la predicción. El uso del criterio de mínimos cuadrados para estimar el modelo (3.1) da como resultado un problema mal planteado debido a que el operador de covarianza (2.10), en general, no es invertible (Aguilera et al., 2016).

De hecho, la estimación de la función del coeficiente de regresión $\beta(t)$ bajo el criterio de mínimos cuadrados da como resultado la ecuación integral de Wiener Hopf:

$$E(Y \cdot X_t) = \int_0^T E[X_t \cdot X_s]\beta(s)ds \tag{3.4}$$

El enfoque *PLS* es una solución eficiente al problema inverso encontrado en (3.4) (Preda & Saporta, 2005).

En la Sección 2.7.2 se determinó el conjunto $\{\eta_h\}_{h=1}^p$ de componentes *FPLS* usando el procedimiento iterativo de las ecuaciones dadas en (2.20) que se tratan en Apostol & Pedra (2010). A partir de las observaciones funcionales $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ se usan los puntajes de las componentes *FPLS* dadas por la descomposición (2.19) como se muestra a continuación:

$$\chi_i = \sum_{h=1}^{p} p_h(t)\eta_h$$

El modelo asociado a la de regresión funcional (3.1) se expresa mediante:

$$y_i = b_o + \sum_{h=1}^{p} b_h\langle\chi_i(t), \kappa_h(t)\rangle + \epsilon_i$$

De aquí, se puede reescribir η_h como

$$\eta_h = \langle \chi_{h-1}, \kappa_h \rangle = \langle \chi_i, \phi_h \rangle$$

donde

$$\phi_h = \kappa_h - \langle p_1, \kappa_h \rangle \phi_h - ... - \langle p_{h-1}, \kappa_h \rangle \phi_{h-1}$$

siendo $\phi_1 = \kappa_1$. De tal modo que

$$y_i = b_0 + \sum_{h=1}^{p} b_h \langle \chi_i(t), \phi_h(t) \rangle + \epsilon_i$$

En otras palabras, variable de respuesta y_h se reescribe de la siguiente manera:

$$y_i = b_0 + \langle \chi_i(t), \sum_{h=1}^{p} b_h \phi_h(t) \rangle + \epsilon_i$$

Este modelo es un modelo de regresión múltiple multivariado estándar que se resuelve por mínimo cuadrados. De esta manera, la estimación de $\mathbf{b} = (b_1, ...b_p)$ está dada por

$$\hat{\mathbf{b}} = (S^T S)^{-1} S^T Y$$

donde $S = \{\phi_h\}_{nxp}$. Por lo cual, se considera que

$$\beta_{FPLS}(t) = \sum_{h=1}^{p} b_h \phi_h(t)$$

de modo que se puede llegar al modelo (3.1) mediante:

$$y_i = b_0 + \int_0^T [\chi_i(t)][\sum_{h=1}^{p} b_h \phi_h(t)]dt + \epsilon_i$$

Luego de realizar la regresión funcional por medio de componentes *FPLS*, se puede recuperar la función estimada, dada por

$$\hat{\beta}_{FPLS}(t) = \sum_{h=1}^{p} b_h \phi_h(t)$$

En este trabajo, para el caso de estudio sobre la estimación de la Precipitación se plantea utilizar bases determinadas por las componentes *FPLS* del modelo estudiado a partir de la Temperatura como covariable explicativa funcional.

Por lo que, se plantea el siguiente problema:

$$y_i = c_0 + \int_0^T [\chi_i(t)][\beta_{FPLS}(t)]dt + \epsilon_i \tag{3.5}$$

donde $\{y_i\}_{i=1}^n$ es la muestra de la variable escalar Y representada por la Precipitación y la muestra $\mathcal{S}_n = \{\chi_i\}_{i=1}^n$ corresponde a la Temperatura como covariable explicativa funcional. De manera similar, se realizará el mismo proceso utilizando la Velocidad del Viento como covariable explicativa funcional.

3.2. MODELO LINEAL FUNCIONAL CON MÁS DE UNA COVARIABLE FUNCIONAL

El modelo *FLR* estudiado en la Sección 3.1.1 es generalizado para estimar la relación entre una variable de respuesta escalar Y y un número finito q de covariables funcionales $X_t^j = \{\mathcal{X}^j(t)\}_{j=1}^q$. Sin pérdida de generalidad, se considera que $E[Y] = 0$ y $E[X_t^j] = 0 \; \forall t \in [0, T]$. En el caso, de un proceso no centrado se debería usar $\{X_t^j - \mu(t)\}$.

Así, el modelo *FLR* con respuesta escalar Y y un número finito q de covariables explicativas funcionales X_t^j está determinado por

$$Y = \alpha + \sum_{j=1}^{q} \int_0^T X_t^j \beta^j(t) dt + \epsilon$$

Se considera $\{y_i\}_{i=1}^n$ una muestra de la variable de respuesta escalar Y y un conjunto de curvas muestrales $\{\mathcal{S}_n^1, ..., \mathcal{S}_n^q\}$ tal que $\{\mathcal{S}_n^j\}_{j=1}^q = \{\chi_i^j(t)\}_{i=1}^n$ en $\mathcal{L}^2[0, T]$.

La regresión se puede expresar mediante:

$$y_i = \alpha + \sum_{j=1}^{q} \int_0^T \chi_i^j(t) \beta^j(t) dt + \epsilon_i \tag{3.6}$$

donde $\beta^j(t) \in \mathcal{L}^2[0, T]$ es el parámetro funcional correspondiente a cada variable funcional y $\epsilon_i \sim N(0, \sigma^2)$ es el residual tal que $Cov(\epsilon_i, \epsilon_j) = 0 \; \forall i \neq j$.

Este modelo ha sido planteado en Ramsay & Silverman (2005) y los coeficientes funcionales pueden estimarse de diferentes formas, como se puede ver en las secciones previas.

En este trabajo, para el caso de estudio sobre la estimación de la Precipitación se plantea realizar una adaptación del modelo dado por la ecuación (3.6) usando la Temperatura y la Velocidad del Viento como covariables explicativas funcionales.

Por lo que, se plantea el siguiente problema:

$$y_i = \alpha + \int_0^T \chi_i^1(t)\beta^1(t)dt + \int_0^T \chi_i^2(t)\beta^2(t)dt + \epsilon_i \tag{3.7}$$

donde $\{y_i\}_{i=1}^n$ es la muestra de la variable escalar Y representada por la Precipitación y las muestras $\mathcal{S}_n^1 = \{\chi_i^1\}_{i=1}^n$, $\mathcal{S}_n^2 = \{\chi_i^2\}_{i=1}^n$ correspondes a las variables explicativas funcionales de la Temperatura y la Velocidad del Viento respectivamente. De modo que, la estimación de $\beta^1(t)$ y $\beta^2(t)$ será calculada mediante representación en bases funcionales.

3.3. MEDIDAS DE INFLUENCIA FUNCIONAL

Las medidas de influencia funcional son las siguientes:

1) La medida funcional de predicción de Cook *(CP)* permite detectar observaciones cuya eliminación puede implicar cambios importantes en la predicción del resto de los datos y se define de la siguiente manera:

$$CP_i = \frac{(\hat{y} - \hat{y}_{-i})^T(\hat{y} - \hat{y}_{-i})}{S_R^2}$$

donde $\hat{y}_{-i}$ es la predicción de la respuesta y excluyendo la i-ésima observación (X_i, y_i) en la estimación.

2) La medida funcional de Cook para la estimación *(CE)* permite detectar observaciones cuya eliminación puede implicar cambios importantes en la estimación.

$$CE_i = \frac{||\hat{\beta} - \hat{\beta}_{-i}||^2}{\frac{S_R^2}{n}\sum_{k=1}^{k_x}\frac{1}{\lambda_k}}$$

donde $\hat{\beta}_{-i}$ es la estimación del parámetro β excluyendo la i-ésima observación (X_i, y_i) en el proceso.

3) La medida funcional de Peña para la predicción *(P)* permite detectar las observaciones cuya predicción se ve más afectada por la eliminación de otros datos.

$$P_i = \frac{(\hat{y}_i - \hat{y}_{(-1,i)}, ..., \hat{y}_i - \hat{y}_{(-n,i)})^T(\hat{y}_i - \hat{y}_{(-1,i)}, ..., \hat{y}_i - \hat{y}_{(-n,i)})}{S_R^2 h_{ii}}$$

donde $\hat{y}_{(-n,i)}$ es el i-ésimo componente del vector de predicción $\hat{y}_{(-n)}$ para $h = 1, ..., n$.

En Febrero-Bande et al. (2010) se discuten estas medidas que sirven para encontrar las curvas funcionales influyentes en la estimación de los modelos de regresión. Es importante recalcar que, las curvas influyentes halladas en ocasiones coinciden con las curvas atípicas; en otros casos a pesar de encontrar una curva influyente no necesariamente esta curva es un dato atípico funcional.

3.4. NÚMERO ÓPTIMO DE COMPONENTES FUNCIONALES EN LOS MODELOS FLR

En la regresión por *FPC* se presenta un dilema al decidir qué componentes usar en la regresión. Este problema se resuelve eligiendo un subconjunto óptimo de *FPC* que sean los mejores estimadores para la respuesta.

Los métodos para elegir el número de componentes son los siguientes:

- Validación cruzada predictiva:

$$PCV(k_n) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left\langle \chi_i, \hat{\beta}_{(-i,k_n)} \right\rangle\right)^2$$

- Criterios de selección del modelo:

$$MSC(k_n) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left\langle \chi_i, \hat{\beta}_{(-i,k_n)} \right\rangle\right)^2\right] + p_n\frac{k}{n}$$

  donde el valor $p_n =$ depende del criterio que se escoja. Los criterios son los siguientes:

  1) Criterio de Información de Akaike *(AIC)*: $p_n = 2$
  2) Criterio de Información de Akaike corregido *(AICc)*: $p_n = \frac{2n}{n-k_n-2}$
  3) Criterio de Información de Schwarz *(SIC)*: $p_n = \frac{\log(n)}{n}$
  4) Criterio de Información de Schwarz corregido *(SICc)*: $p_n = \frac{\log(n)}{n-k_n-2}$

En el caso de la regresión por *FPLS*, cada componente se obtiene maximizando la covarianza entre X_t e Y, de modo que se puede seleccionar directamente la primera componente *PLS* funcional mediante Validación cruzada *(CV)* o Criterios de selección del modelo *(MSC)* en lugar de utilizar la mejor combinación de los componentes. Estas medidas se discuten en Febrero-Bande et al. (2010).

3.5. EVALUACIÓN DE LOS MODELOS FLR

Luego de estimar los parámetros de cualquier modelo se debe realizar una evaluación con el fin de saber que tan bien se ajusta el modelo a los datos (bondad de ajuste), y proceder a comparar resultados para seleccionar el mejor ajuste.

Bajo este enfoque, es útil introducir medidas de error en las predicciones que estimen el rendimiento y evalúen el ajuste del modelo a los datos para conocer la capacidad predictiva del mismo. Estas medidas son:

1) **Error Absoluto Medio:** Es una medida del error absoluto de predicción, y se encuentra definida por

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

donde y_i son las observaciones reales, $\hat{y}_i$ son los valores predichos y n es el tamaño de la muestra. Es una medida robusta para los valores atípicos donde todas las diferencias individuales se ponderan por igual.

2) **Error Porcentual Absoluto Medio:** Es una medida del error en términos porcentuales. Se define por

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

Esta medida expresa los errores como un porcentaje, lo que es una ventaja pues dispone una forma más intuitiva y fácil de juzgar el alcance de los errores del modelo.

3) **Suma de Cuadrados del Error de Predicción:** Es una medida de la desviación entre los valores ajustados y los valores observados del error de pronós-

tico. Se define por

$$PRESS = \sum_{i=1}^{n} |y_i - \hat{y}_i|^2$$

Mientras más pequeña sea esta medida mejor será la capacidad de predicción del modelo.

La evaluación continua a través del uso de medidas de bondad de ajuste, las cuales resumen la discrepancia entre los valores observados y predichos del modelo. Estas medidas permiten comparar varios modelos a la vez para determinar el mejor de los ajustes. Las medidas comunes de bondad de ajuste son:

1) **Coeficiente de determinación ajustado:** Es un cociente de varianza en lugar de un cociente de sumas de cuadrados, y se define por

$$R^2 ajustado = 1 - \frac{SSR/(n-(k+1))}{SSY/(n-1)}$$

donde $SSR/(n-(k+1))$ es la varianza de los residuales y $SSY/(n-1)$ es constante e independiente de la cantidad de variables en el modelo.

Este coeficiente resulta ser independiente del número de variables explicativas (independientes entre ellas), y sólo aumenta si al agregar una variable al modelo se reduce el cuadrado medio residual Aparicio et al. (2004). Se usa para comparar modelos totalmente independientes entre sí.

2) **Criterio de información de Akaike:** Es una medida de la calidad relativa de un modelo estadístico, y está definida por

$$AIC = 2k - 2\ln\Big(L[\hat{\beta}(k)]\Big)$$

donde $(L[\hat{\beta}(k)]$ es la función de máxima verosimilitud de las observaciones y k es el número de parámetros independientes estimados dentro del modelo. El primer término indica una penalización que crece conforme aumenta el número de parámetros; mientras que el segundo término puede ser interpretado como una medida de bondad de ajuste.

El modelo que mejor se ajusta tienen el menor valor de AIC, es decir, no se pretende identificar el modelo verdadero, sino que es una herramienta alternativa para seleccionar el mejor modelo (Aparicio et al., 2004).

3) **Criterio de información de Akaike corregido:** Es una medida de corrección de segundo orden del valor AIC, y se define por

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

donde n es el tamaño de la muestra. Este coeficiente se usa cuando el conjunto de datos es más pequeño (Aparicio et al., 2004).

4) **Criterio de información Bayesiano/Schwarz:** Es una medida alternativa con un enfoque bayesiano, y está definido por

$$BIC = k\ln(n) - 2\ln\Big(L[\hat{\beta}(k)]\Big)$$

Este coeficiente penaliza el número de parámetros con $\ln(n)$ y es mucho más fiable al aumentar el tamaño muestral. Estima la pérdida de información al aproximar un modelo real con un modelo estimado. El mejor modelo será aquel que tenga el menor valor BIC (Aparicio et al., 2004).

3.6. VALIDACIÓN DE LOS MODELOS FLR

Estos modelos requieren el cumplimiento de ciertas condiciones para que los resultados obtenidos sean fiables y se puedan usar más adelante para su aplicación. Es primordial conocer, diagnosticar y solucionar los supuestos que se deben satisfacer, de este modo, en cada caso se define la hipótesis nula (H_o) y alternativa (H_a) con el fin de realizar un contraste mediante pruebas estadísticas al $95\,\%$ de confianza.

Para comenzar se pretende determinar si los coeficientes estimados $\hat{b}_j$ del parámetro funcional $\beta(t)$, del modelo representado por un número κ_B de bases funcionales mostrado en la Sección 3.1.1, son realmente significativos (distintos de cero) para explicar la variable de respuesta escalar o si, por el contrario, se consideran nulos. Para resolver esta incógnita se plantea usar el siguiente contraste de hipótesis:

$$H_o : b_j = 0 \ \ \forall\ j = 1, ..., \kappa_B$$

$$H_a : b_j \neq 0 \ \ \forall\ j = 1, ..., \kappa_B$$

El estadístico de prueba que permite determinar este contraste es

$$t_c = \frac{\hat{b}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

donde C_{jj} es elemento de la diagonal $(Z^T Z)^{-1}$ correspondiente a $\hat{b}_j$.

De manera que, se rechazará la hipótesis H_o si $|t_c| > t_{\alpha/2,n-k-1}$ o si el $valor\ p$ asociado al estadístico t_c es menor que la significancia $\alpha = 0.05$. En otras palabras, si se cumplen las condiciones se aceptará la hipótesis H_a de que los parámetros son significativos.

En el caso de los modelos dados por componentes *FPC* y *FPLS* se plantea determinar si los coeficientes de las componentes halladas por los modelos son significativos o no mediante el mismo contraste.

Los supuestos que se deben cumplir se describen a continuación:

1) **Supuesto de Normalidad de los Residuos:** Se desea que los residuos sigan una distribución normal, en otras palabras, se requiere que $\epsilon_i \sim N(0; \sigma^2)$. Para ello, el contraste que se usará está dado por

$$H_o : \text{LA DISTRIBUCIÓN ES NORMAL}$$

$$H_a : \text{LA DISTRIBUCIÓN NO ES NORMAL}$$

Este supuesto se analiza a través del gráfico de los residuos. Se verifica analíticamente por el estadístico del test de Kolgomorov-Smirnov, definido por

$$D = \max |F_n(x) - F_o(x)|$$

donde $F_n(x)$ y $F_o(x)$ son las funciones de distribución muestral y teórica respectivamente, este se usa cuando existe una muestra mayor a 50.

Cuando la muestra es máxima de 50 se usa el test de Shapiro-Wilks cuyo estadístico de prueba es

$$W = \frac{D^2}{NS^2}$$

donde D es la suma de las diferencias corregidas y S^2 es la varianza muestral (Faraway, 2014).

En el caso de estudio se cuenta con 25 curvas por cada variable meteorológica de modo que se utilizará el test de Shapiro-Wilks, y se rechazará H_o de

normalidad si el $valor\ p$ asociado al estadístico de la prueba W es menor que la significancia $\alpha = 0,05$, es decir, se aceptará H_a de que la distribución de los residuos no es normal.

2) **Supuesto de Autocorrelación, Independencia de los Residuos**: Se procura que los residuos sean independientes entre sí de modo que no exista ningún tipo de correlación entre ellos o, lo que es lo mismo, que los residuos no tengan autocorrelación. En otras palabras, se quiere verificar que $Cov(\epsilon_i, \epsilon_j) = 0\ \forall i \neq j$ por medio del contraste definido por

H_o : RESIDUOS INDEPENDIENTES = NO AUTOCORRELACIÓN DE LOS RESIDUOS

H_a : RESIDUOS NO INDEPENDIENTES = AUTOCORRELACIÓN DE LOS RESIDUOS

Una forma de diagnosticar la autocorrelación es por el test de Durbin-Watson (Faraway, 2014). Donde se considera que el residual ϵ_i asociado a la observación en el tiempo t tiene la varianza y la media constantes e independientemente de t; el residual se define por $\epsilon_i = \rho\epsilon_{i-1} + z_i\ \forall i = 1, ..., n$ donde la variable aleatoria $z_i \sim N(0; \sigma^2)$ y el parámetro de autocorrelación es ρ. De esta manera, se reescriben las hipótesis como

$$H_o : \rho = 0$$

$$H_a : \rho \neq 0$$

Se usa el estadístico de Durbin-Watson dado por

$$d = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}\epsilon_i^2}$$

Durbin y Watson demostraron que d se encuentra entre dos cotas halladas en función del valor de probabilidad $\alpha = 0.05$, denominadas como cota inferior (d_L) y cota superior (d_U), tal que estos son los límites que determinarán, según se encuentre el valor del estadístico d, si existe autocorrelación o no. La decisión depende de los siguientes puntos:

a) Si $d < d_L$ se rechaza H_o.

b) Si $d > d_U$ no se puede rechazar H_o.

c) Si $d_L \leq d \leq d_U$ la prueba no es concluyente.

Si los valores del estadístico son pequeños implica que se debe rechazar H_o, pues la autocorrelación positiva indica que las diferencias entre los residuales serán pequeñas.

Otra forma de diagnosticar la autocorrelación es mediante el test de Box-Pierce $\&$ Ljung-Box. En primer lugar, Box y Pierce desarrollaron un estadístico que se define como

$$Q_{BP} = n \sum_{j=1}^{p} \rho_j^2$$

donde $p_j = \frac{\sum_{i=j+1}^{n} \epsilon_i \epsilon_{i-1}}{\sum_{i=1}^{n} \epsilon_{i-1}^2}$. El estadístico Q_{BP} está basado en los cuadrados de los primeros coeficientes de autocorrelación de los residuos, y sigue una distribución χ_p^2 con p grados de libertad.

Posteriormente este estadístico fue revisado por Ljung-Box obteniéndose un refinamiento para mejores resultados, y está definido por

$$Q_{LB} = n(n+2) \sum_{j=1}^{p} \frac{\rho_j^2}{n-j}$$

donde Q_{LB} tiene la misma distribución χ_p^2 con p grados de libertad.

Estos contrastes se han revelado potentes, bajo la hipótesis H_o independencia de los residuos. De modo que, si $Q_{BP} > \chi_{\alpha,p}^2$ o si $Q_{LB} > \chi_{\alpha,p}^2$ se rechazará la hipótesis H_o, en otras palabras, si el $valor\ p$ asociado al estadístico Q_{BP} o Q_{LB} es menor que la significancia $\alpha = 0,05$ se aceptará la hipótesis H_a de que los residuos tienen autocorrelación o bien que los residuos no son independientes.

3) **Supuesto de Homocedasticidad de los Residuos:** Se asume que la varianza de los errores es constante a lo largo de las observaciones, es decir, se requiere que $Var(\epsilon_i) = \sigma^2$. Esto se conoce como homocedasticidad, y es una cualidad necesaria para que los coeficientes estimados sean eficientes, lineales e insesgados; caso contrario será heterocedasticidad. Para verificar este supuesto se usa el contraste dado por

H_o : HOMOCEDASTICIDAD = VARIANZA CONSTANTE DE LOS RESIDUOS

H_a : HETEROCEDASTICIDAD $=$ VARIANZA NO CONSTANTE DE LOS RESIDUOS

Esto se diagnostica mediante el gráfico de residuos frente a los valores predichos. Se comprueba analíticamente usando el test de Breusch-Pagan/Godfrey cuyo estadístico de prueba es $\chi_c^2 = nR^2$ donde R^2 es el coeficiente de determinación y este sigue una distribución χ_{p-1}^2 con tantos grados de libertad o como variables explicativas introducidas para justificar la falta de varianza constante. En Aparicio et al. (2004) se describen distintas pruebas que podrían usarse para reconocer la heterocedasticidad.

De manera que, se rechazará H_o si $\chi_c^2 > \chi_{\alpha,p-1}^2$ o si el $valor\ p$ asociado al estadístico χ_c^2 es menor que la significancia $\alpha = 0,05$, es decir, se aceptará H_a de que los residuos no tienen varianza constante.

4) **Supuesto Especificación o Linealidad:** Se desea que la relación entre la variable explicativa y la respuesta escalar sea lineal. Esto puede confundirse con el problema de la autocorrelación. Para verificar la linealidad del modelo se propone el contraste dado por

H_o : MODELO CORRECTAMENTE ESPECIFICADO $=$ LINEAL

H_a : MODELO NO ESPECIFICADO $=$ NO LINEAL

Para diagnosticar esto se usa el test de Ramsey (Faraway, 2014). Esta prueba contrasta si es necesario introducir términos cuadráticos o cúbicos para que desaparezcan los patrones sistemáticos en los residuos. El estadístico de prueba está definido por

$$FR = \frac{(R^2 - R_0^2)/p}{(1 - R^2)/(n - k - p)}$$

donde p y $T-k-p$ son grados de libertad, siendo p el número de restricciones impuestas a los coeficientes estimados, k el número de parámetros dados por una regresión auxiliar, R^2 y R_0^2 son los coeficientes de determinación de modelo y de la regresión auxiliar respectivamente.

Se rechazará H_o si $FR > F_{\alpha,p,n-k-p}$ o si el $valor\ p$ asociado al estadístico de prueba FR es menor que la significancia $\alpha = 0,05$, es decir, se aceptará H_a de que el modelo no tiene una correcta especificación.

CAPÍTULO 4

APLICACIÓN DE LOS MODELOS

Este capítulo se centra en la aplicación de las metodologías estudiadas para el desarrollo del *FDA* y de los modelos *FLR* con respuesta escalar orientado al campo de la meteorología con datos recolectados en las provincias de mayor concentración de producción de maíz duro seco en Ecuador. Para ello, se toma en cuenta el Caso de Estudio visto en Sección 1.4 donde se expuso la selección de variables meteorológicas para la predicción de la Precipitación como variable de respuesta escalar mediante el uso de la Temperatura y Velocidad del Viento como covariables explicativas funcionales.

El análisis se desarrollará en diferentes etapas. En primer lugar, se puntualizan las medidas adoptadas en la creación de los datos funcionales y las técnicas aplicadas en la reducción dimensional. Se prosigue con el análisis exploratorio de los datos funcionales. Luego, se aplicarán los modelos de regresión funcional con respuesta escalar. Finalmente, se escogerá y validará el mejor ajuste entre todas las estimaciones dadas por cada uno de los modelos planteados.

El desarrollo de todos los métodos se halla en el entorno de programación `R`. Específicamente, se utiliza el paquete `fda.usc` propuesto en Febrero Bande & Oviedo de la Fuente (2012).

4.1. SUAVIZAMIENTO PARA DATOS FUNCIONALES

Se empieza detallando el proceso de suavizamiento para determinar los datos funcionales, que se usarán para el *FDA*, a partir de observaciones discretas.

Los datos iniciales presentan una correspondencia directa por cada unidad de tiempo (días), es decir, por cada día se tiene una observación, sin embargo, cuando se suavizan los datos se obtiene una función que depende del tiempo creando una representación continua

Desde esta perspectiva, se propone reconstruir la forma funcional de las trayectorias muestrales tanto de la Temperatura como de la Velocidad del Viento en términos de bases funcionales de Fourier como se analizó en la Sección 2.5.1, pues estas variables tienen un comportamiento periódico.

4.1.1. SELECCIÓN DEL NÚMERO DE BASES FUNCIONALES

El análisis se centra en el promedio diario de la Temperatura (análogamente, se considera el promedio diario de la Velocidad del Viento) entre los años 2010 y 2020, de 25 lugares estratégicamente ubicados por su alta productividad en maíz duro seco del país. La idea es utilizar cada variable dada por 365 observaciones en 25 lugares para reconstruir la forma funcional de las muestras de cada variable a partir de bases de Fourier.

Se procede a seleccionar el número de bases κ_x de Fourier que permiten tener una mejor representación de las curvas muestrales como datos funcionales para cada variable. Para ello, se realiza Validación Cruzada Generaliza (GCV) haciendo variar $\kappa_x = 5, 11, 12, 15, 21, 35, 65$. Estas bases se visualizan en el Anexo 1.

En el caso del promedio diario de la Temperatura se ha obtenido los siguientes resultados:

N° Bases	GCV **Óptimo**	$R^2 ajustado$
5	1,001	0,826
11	0,950	0,835
12	0,509	0,933
15	0,549	0,905
21	0,470	0,939
35	0,430	0,925
65	0,321	0,944

TABLA 4.1. Criterio GCV para la selección del número de bases κ_x de Fourier en el caso del promedio diario de la Temperatura.

Se puede observar que el valor GCV óptimo decrece conforme aumenta el número de bases κ_x de Fourier mientras que el valor de $R^2 ajustado$ aumenta hasta

$\kappa_x = 12$. A partir de $\kappa_x = 15$ el $R^2ajustado$ cambia tanto que alcanza un $0,94$ considerando 65 bases lo que implica que se empieza tomar información que no aporta al suavizamiento. De manera que, la explicación de la variación del suavizamiento realizado se ve afectado una cantidad muy alta de bases seleccionadas. Para visualizar el suavizamiento del promedio diario de la Temperatura, mediante distintas bases $\kappa_x = \{5, 15, 21, 35, 41\}$ de Fourier, se muestra el caso del cantón Babahoyo ubicado en la provincia de Los Ríos (la más productiva de maíz duro seco) en la siguiente figura:

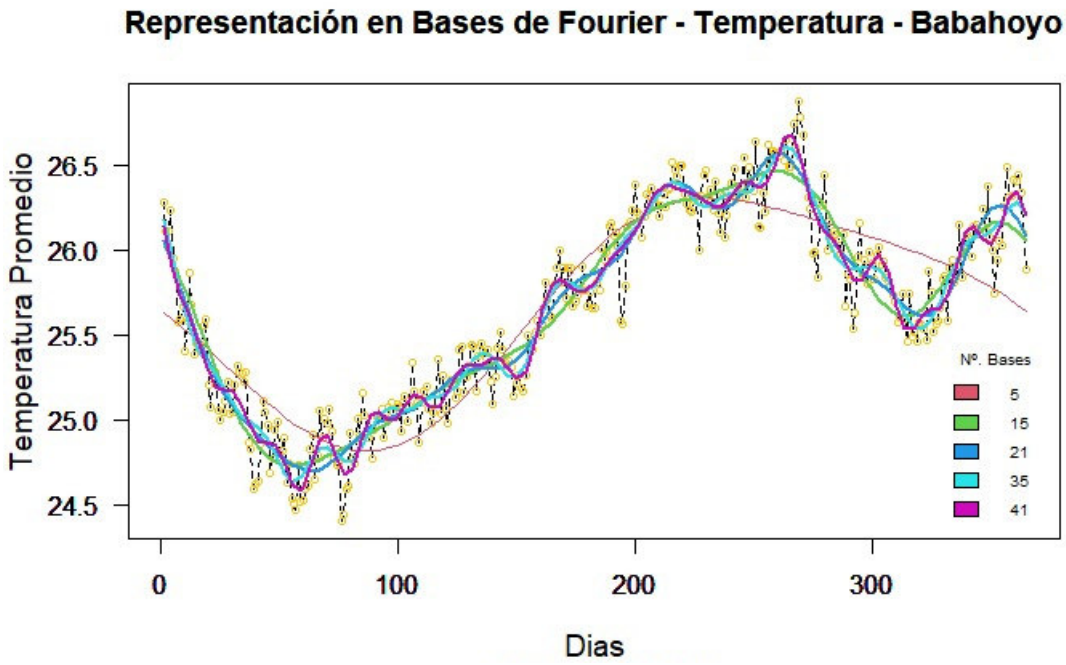

FIGURA 4.1. Suavizamiento del promedio diario de la Temperatura de Babahoyo usando diferentes números de bases κ_x de Fourier.

En la figura (4.1) se aprecia que mientras crece el número de bases los datos reales son más variantes que los valores predichos. Por lo que, se decide tomar el valor mínimo de bases que tenga mayor $R^2ajustado$, pues siempre la menor cantidad tendrá mayor facilidad en la interpretación y de los procesos computacionales.

En este caso, por los resultados de la tabla (4.1) se escogen las 12 primeras bases de Fourier, ya que a partir de dicho valor el $R^2ajustado$ tiene mayor explicación y se determina que es despreciable la disminución de GVC.

De manera análoga, se muestra la selección del número óptimo de bases κ_x para el promedio diario de la Velocidad del Viento en la siguiente tabla:

N° Bases	GCV **Óptimo**	$R^2 ajustado$
5	0,284	0,951
11	0,222	0,971
12	0,171	0,978
15	0,205	0,973
21	0,216	0,972
35	0,205	0,973
65	0,208	0,973

TABLA 4.2. Criterio GCV para la selección del número de bases κ_x de Fourier en el caso del promedio diario de la Velocidad del Viento.

Se puede observar que el valor GCV óptimo varía conforme aumenta el número de bases κ_x de Fourier y el valor de $R^2 ajustado$ aumenta hasta alcanzar un 0.97 considerando 65 bases. Se muestra el caso del suavizamiento del promedio diario de la Velocidad del Viento en el cantón de Babahoyo a continuación:

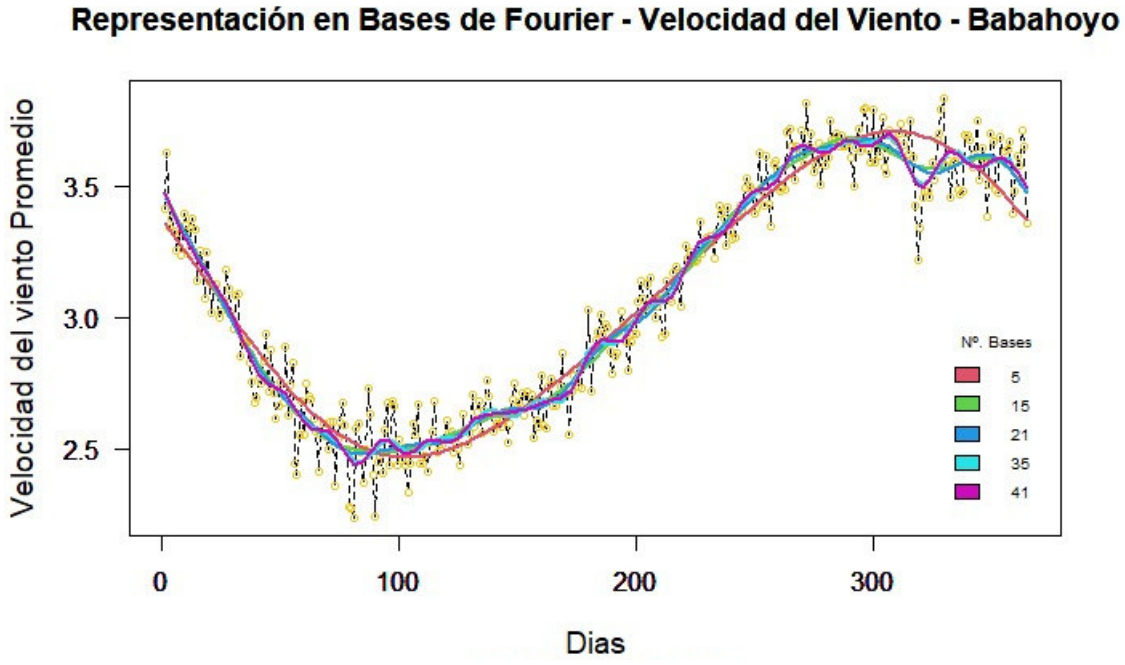

FIGURA 4.2. Suavizamiento del promedio diario de la Velocidad del Viento de Babahoyo usando representación con diferentes números de bases κ_x de Fourier.

En la figura (4.2) se puede ver la aproximación de los valores predichos hacia los datos reales tiene la misma variación cuando la cantidad de bases incrementa. En este caso, se escogen las 11 primeras bases de Fourier por los resultados de la tabla (4.2), ya que a partir de aquí el $R^2 ajustado$ no tienen mayor fluctuación y se determina que es despreciable la disminución de GVC.

4.1.2. SUAVIZAMIENTO PARA LA TEMPERATURA

Se procede a realizar el suavizamiento de la Temperatura mediante las 12 bases funcionales de Fourier determinadas anteriormente. En la siguiente figura se reflejan tanto los datos discretos como los funcionales del promedio diario de la Temperatura para los 25 cantones:

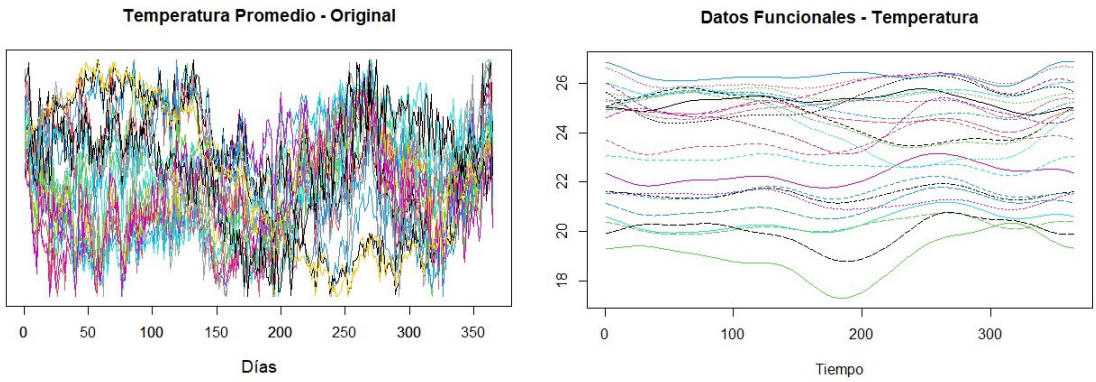

FIGURA 4.3. Datos Discretos (izquierda) y Funcionales (derecha) del promedio diario de la Temperatura de los 25 cantones seleccionados en el país.

Las curvas presentadas a la derecha de la figura (4.3) son los datos funcionales de la Temperatura que se utilizarán para continuar el desarrollo de este trabajo.

4.1.3. SUAVIZAMIENTO PARA LA VELOCIDAD DEL VIENTO

En este caso, la Velocidad del Viento se suaviza con 11 bases funcionales de Fourier del promedio diario de la Velocidad del Viento para los 25 cantones. Se exhiben los resultados a continuación:

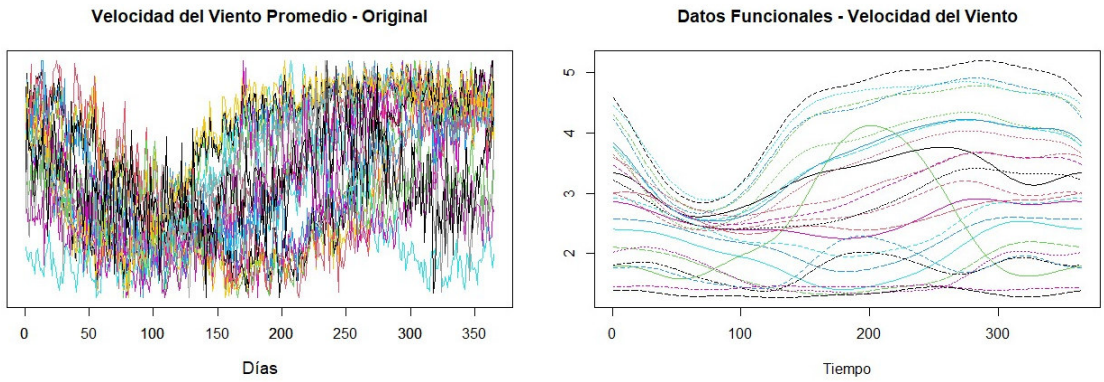

FIGURA 4.4. Datos Discretos (izquierda) y Funcionales (derecha) del promedio diario de la Velocidad del Viento de los 25 cantones seleccionados en el país.

Las curvas presentadas a la derecha de la figura son los datos funcionales de la Velocidad del Viento que también se utilizarán en este trabajo.

4.2. ANÁLISIS EXPLORATORIO DE DATOS FUNCIONALES

Esta sección empieza con el análisis descriptivo de los niveles de la Temperatura y de la Velocidad del Viento por medio de las medidas funcionales de centralidad (media, mediana), dispersión (varianza) y covarianza; además se usan variantes de estas medidas con distintas profundidades.

Luego, se buscan las bandas de confianza por donde oscila la media funcional y la media recortada con profundidad FM aplicando el método bootstrap que se encuentra en la función `fda.bootstrap` del paquete `fda.usc`. También se incluye el análisis de la mediana funcional como el dato más profundo mediante el uso de las medidas de profundidad de: Fraiman-Muniz (`depth.FM`), modal (`depth.mode`), proyecciones aleatorias (`depth.RP`) y proyecciones de Tukey (`depth.RT`).

Por último, se realiza un estudio sobre la presencia de datos atípicos ya que pueden influir en la estimación y desempeño de los modelos mediante las funciones `outliers.depth.trim` y `outliers.depth.pond`. Pues la profundidad es una medida de robustez, donde los puntos tengan mayor valor de profundidad se hallarán más cerca del comportamiento central de los datos; y los valores con menor profundidad son posibles candidatos a datos atípicos.

4.2.1. ANÁLISIS FUNCIONAL DE LA TEMPERATURA

Se procede a realizar un análisis de los datos funcionales o curvas que representan el promedio diario de la Temperatura. A continuación se presentan las medidas funcionales normales y recortadas al $15\,\%$ de la Temperatura con profundidad FM, $modal$, RP y RT:

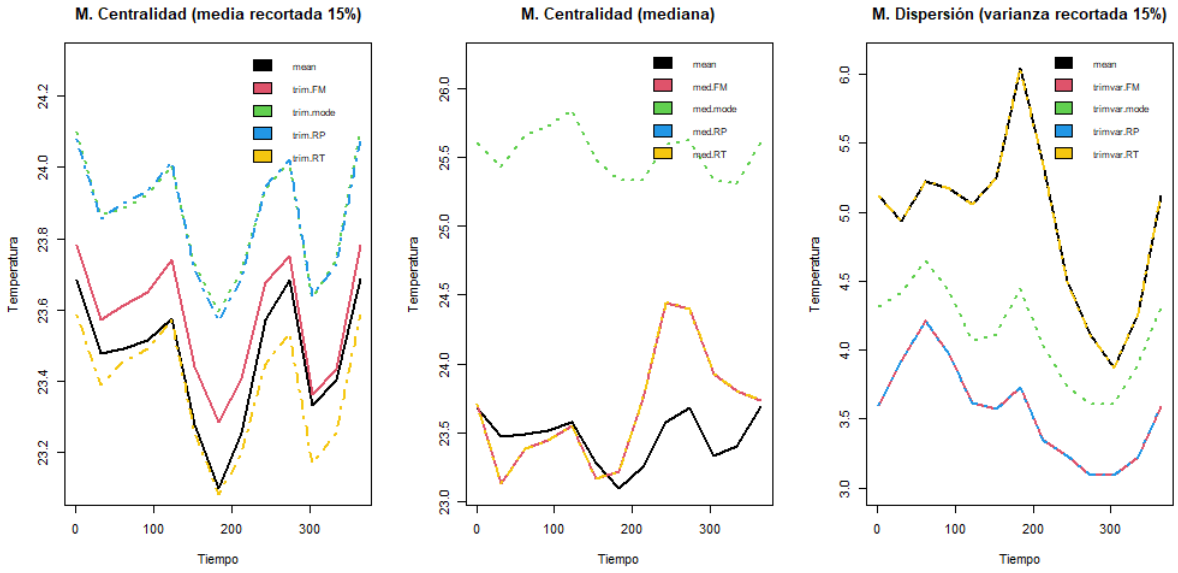

FIGURA 4.5. Medidas funcionales para de la Temperatura por profundidad: media recortada $15\,\%$ (izquierda), mediana (centro), varianza recortada $15\,\%$ (derecha).

En la figura (4.5) se observa que el comportamiento de las medias funcionales (normal y recortada) varía dependiendo de la profundidad. La media alcanza valores más altos cuando se usa la profundidad RP y $modal$ al ser recortadas el $15\,\%$. La mediana normal toma los mayores valores si se considera la profundidad $modal$. Mientras que la varianza normal y la recortada al $15\,\%$ con profundidad RT alcanzan altos valores de variación. Se prosigue con la visualización de la covarianza:

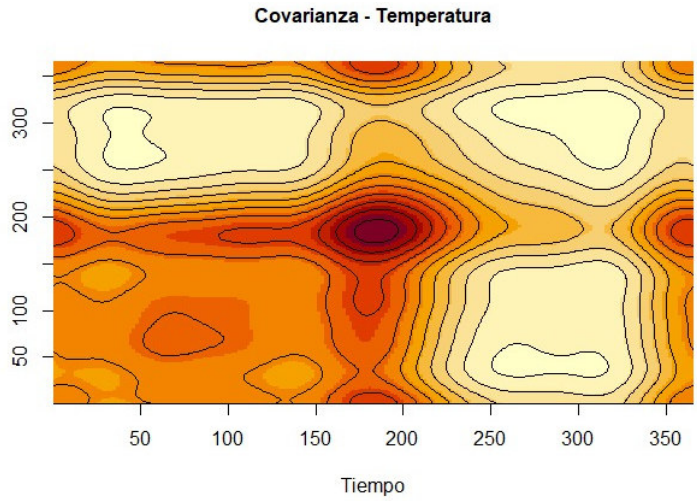

FIGURA 4.6. Covarianza del promedio diario de la Temperatura.

Se observa que la Temperatura promedio alcanza altos niveles en los 200 días y los promedios más bajos en los últimos 300 días del año en la figura (4.6). Ahora, el método bootstrap con 1000 remuestreos se refleja en la siguiente figura:

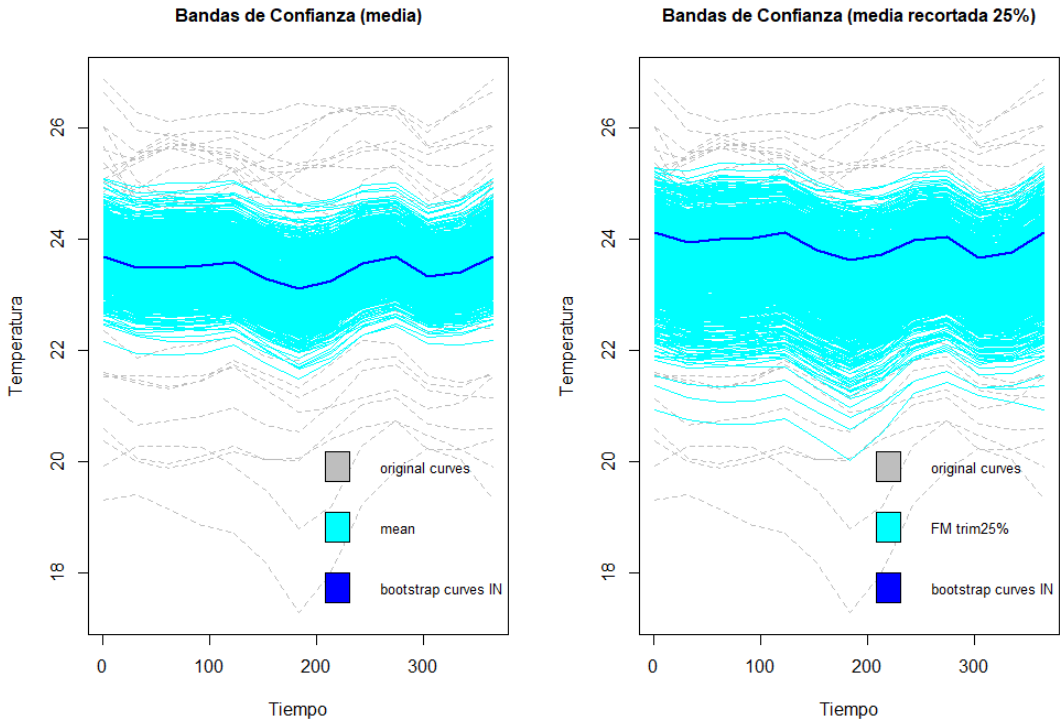

FIGURA 4.7. Bandas de confianza de la Temperatura mediante: media (izquierda) y media recortada al $25\,\%$ con profundidad FM (derecha).

La amplitud de las bandas se debe al número reducido de curvas consideradas en el remuestreo. En la figura (4.7) se observan las respuestas bootstrap de las curvas espectrométricas de la Temperatura donde las bandas de confianza más ajustadas se dan con la media funcional normal.

Como consecuencia del estudio de la tendencia central y la variabilidad de los datos se prosigue a detectar datos atípicos en la muestra. Para ello, se presentan las cuatro medidas de profundidad recortadas un $25\,\%$ respecto a la mediana:

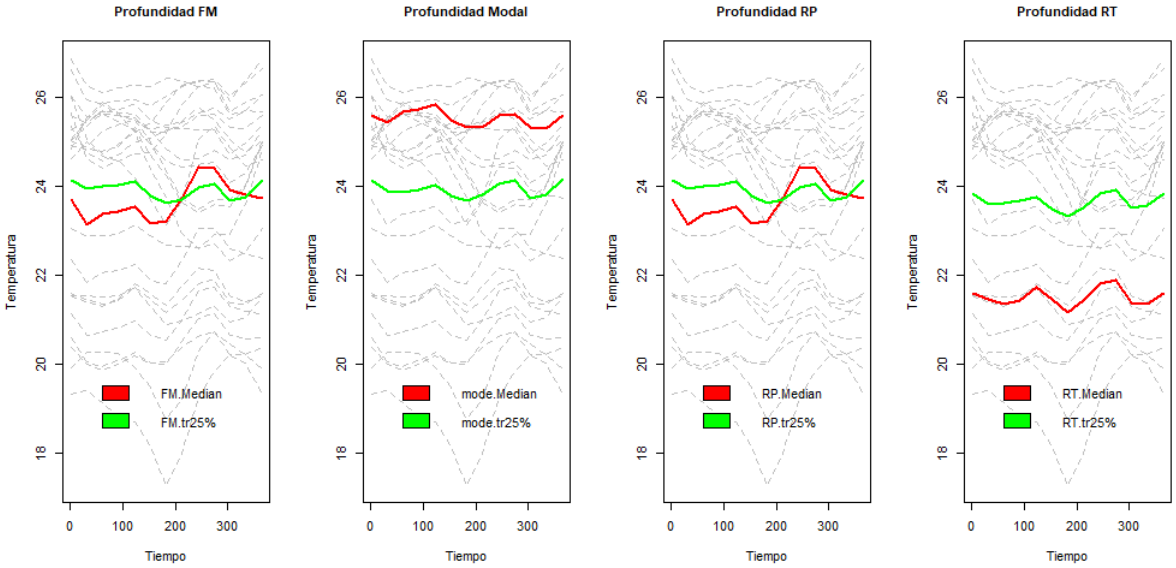

FIGURA 4.8. Representación de las medidas de profundidad FM, $modal$, RP y RT (de izquierda a derecha) recortadas al $25\,\%$ respecto a la mediana para la Temperatura.

En la figura (4.8) se ve que la mediana con profundidad FM y RP son similares, los valores que se alcanza con la profundidad $modal$ son más altos y al usar profundidad RT los valores de la mediana son los más bajos. Luego, los datos atípicos hallados en el caso de la Temperatura se pueden ver en la figura (4.9).

Las funciones indicadas permiten determinar las curvas atípicas en la muestra. En la figura (4.9) se obtiene que las curvas atípicas son: la curva 21 correspondiente al cantón Loja y Loreto (25).

Las curvas encontradas intervienen en la estimación de los modelos, por lo que,

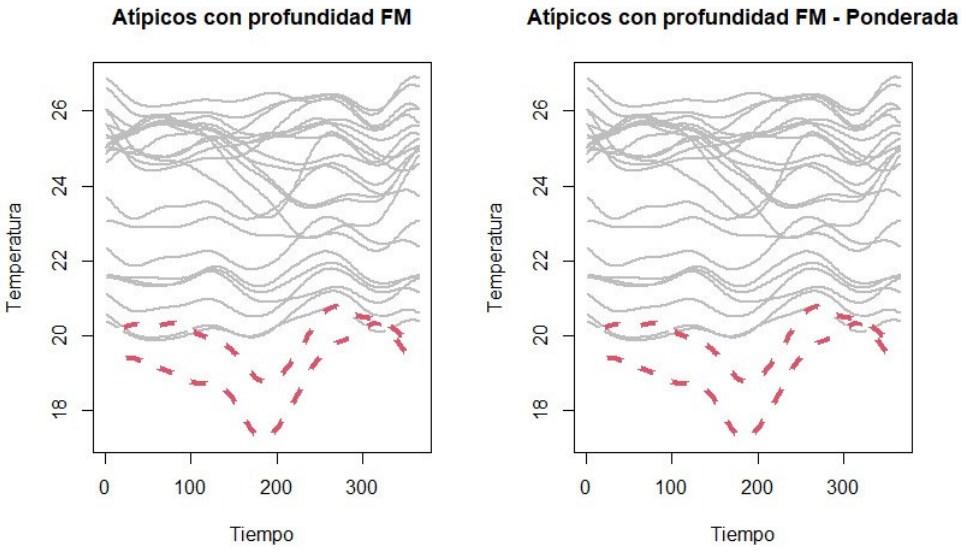

FIGURA 4.9. Datos Atípicos detectados con la función `outliers.depth.trim` (izquierda) y `outliers.depth.pond` (derecha) usando la medida de profundidad FM recortada al $25\,\%$ de la Temperatura.

se procederá a comparar los resultados obtenidos considerando y sin considerar dichas curvas para obtener el mejor ajuste en cada modelo.

4.2.2. ANÁLISIS FUNCIONAL DE LA VELOCIDAD DEL VIENTO

De manera análoga, se procede a visualizar las medidas funcionales normales y recortas al $15\,\%$ del promedio diario de la Velocidad del Viento en la figura 4.10.

En la figura (4.10) se ve que el comportamiento de la media funcional y la varianza no varía tanto respecto a las medidas de profundidad como en el caso de la mediana. La media alcanza los mayores valores con las profundidades $modal$, RP y RT; la mediana toma altos valores con profundidad $modal$. La varianza normal y la recortada al $15\,\%$ con profundidad RT alcanzan valores altos de variación. Ahora, se observa la covarianza en la figura (4.11).

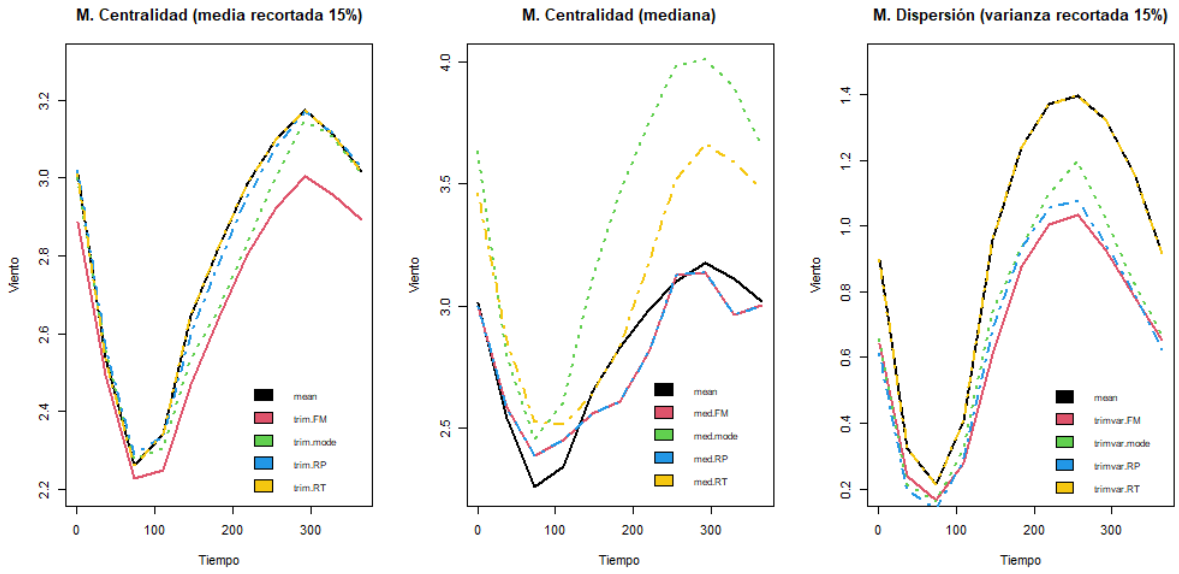

FIGURA 4.10. Medidas funcionales para de la Velocidad del Viento según la profundidad: media recortada $15\,\%$ (izquierda), mediana (centro), varianza recortada $15\,\%$ (derecha).

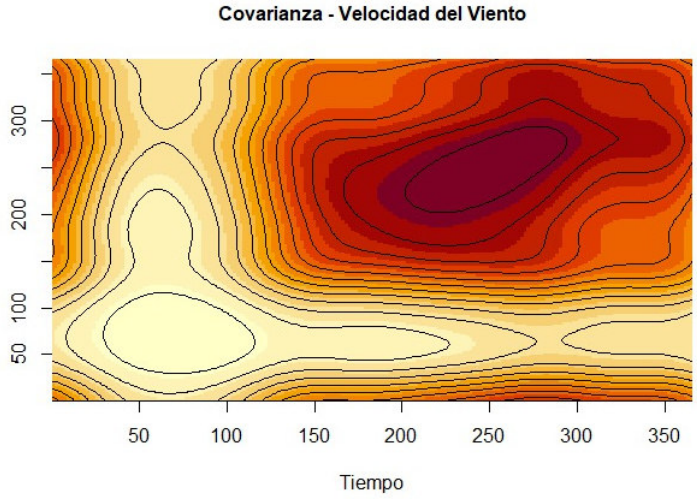

FIGURA 4.11. Covarianza del promedio diario de la Velocidad del Viento.

En la figura (4.11) se ve que el promedio diario de la Velocidad del Viento alcanza altos niveles a partir de los 150 días del año. El método bootstrap con 1000 remuestreos se muestra en la figura (4.12).

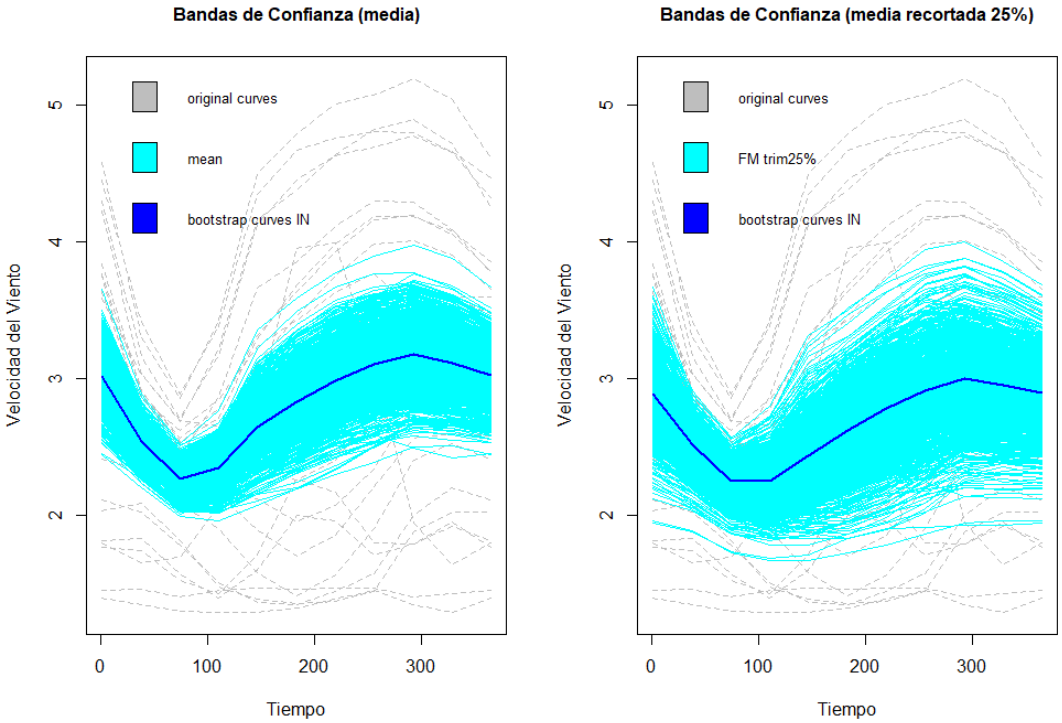

FIGURA 4.12. Bandas de confianza de la Velocidad del Viento mediante: media (izquierda) y media recortada al $25\,\%$ con profundidad FM (derecha).

Al igual que el caso anterior, estas bandas de confianza son más ajustadas para la media en comparación de la media recortada el $25\,\%$.

Se sigue con la detección datos atípicos, gracias al estudio de la tendencia y variabilidad de los datos, en la muestra. Se visualizan las cuatro medidas de profundidad recortadas un $25\,\%$ respecto a la mediana en la figura (4.13).

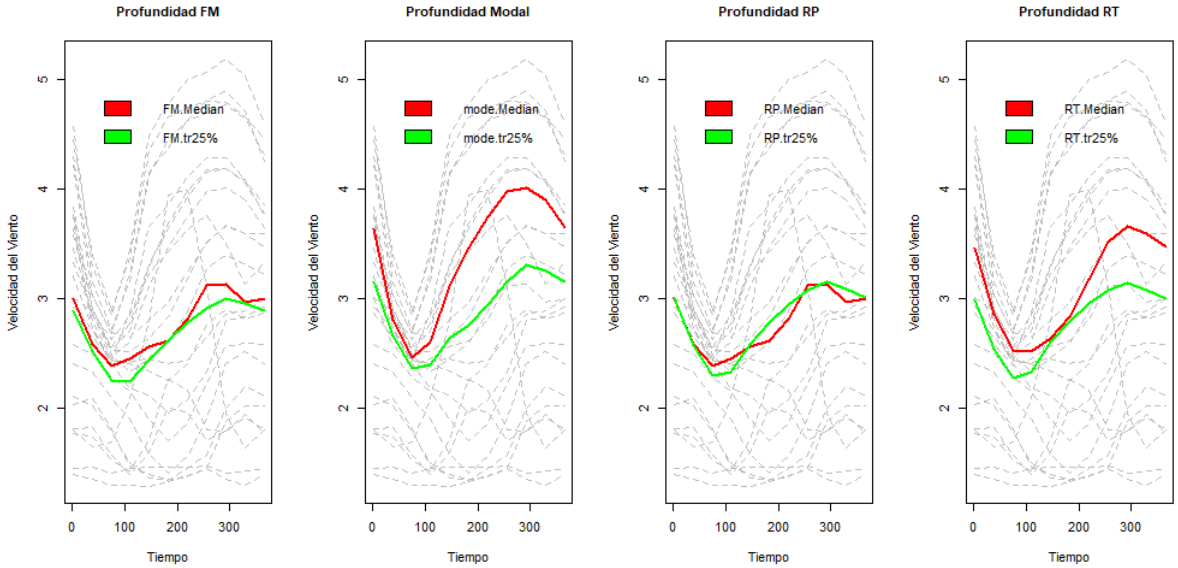

FIGURA 4.13. Representación de las medidas de profundidad FM, $modal$, RP y RT (de izquierda a derecha) recortadas al $25\,\%$ respecto a la mediana para la Velocidad del Viento.

Se aprecia en la figura (4.13) los diferentes casos de profundidades recortadas respecto a la mediana de estos datos funcionales. Los datos atípicos son:

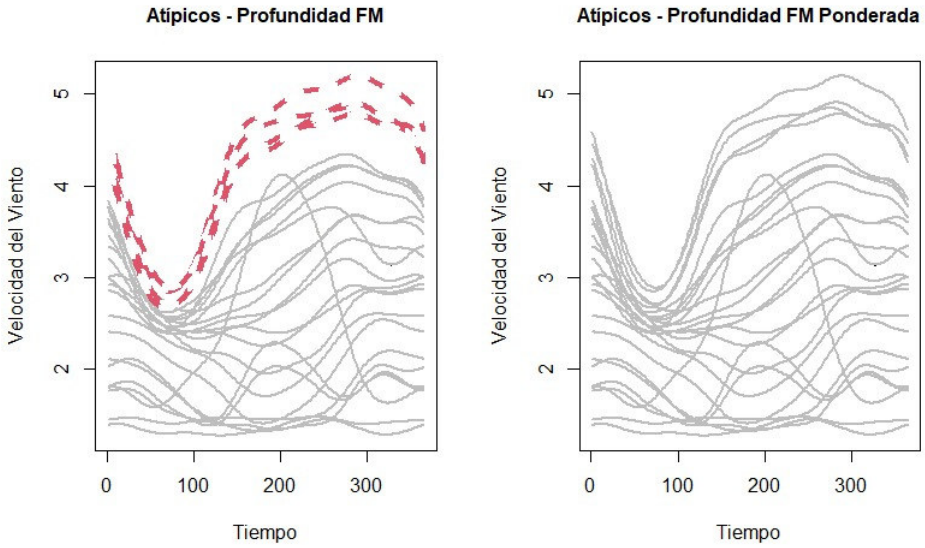

FIGURA 4.14. Datos Atípicos detectados con la función `outliers.depth.trim` (izquierda) y `outliers.depth.pond` (derecha) usando la medida de profundidad FM recortada al $25\,\%$ de la Velocidad del Viento.

Mediante la primera función se detecta cuatro curvas atípicas correspondientes a Manta (7), Santa Elena (23), Pedernales (4) y Jipijapa (9); mientras que la segunda función no encuentra datos atípicos. Por lo que, no se considera datos atípicos en este caso; sin embargo, existe una curva influyente correspondiente a Loja (21).

4.3. AJUSTE DE LOS MODELOS DE REGRESIÓN FUNCIONAL CON RESPUESTA ESCALAR

Se aplican los modelos estudiados para hallar su respectivo ajuste. Para ello, en cada modelo propuesto anteriormente se involucran todos los datos funcionales (25 curvas) correspondientes a la variable meteorológica, y por otra parte se eliminan las curvas atípicas e influyentes en la estimación de los modelos.

4.3.1. MODELOS CON LA TEMPERATURA COMO COVARIABLE FUNCIONAL

En primer lugar, se usan los modelos en los cuales se considera solamente a la Temperatura como covariable explicativa funcional para estimar la Precipitación.

En el Anexo 2 se presentan los resultados del primer modelo, propuesto en la ecuación (3.2), obtenidos mediante la función `fregre.basis`. A continuación, se muestra la tabla de resumen:

Modelo FLR con Bases Funcionales: Temperatura						
# Curvas	$R^2 ajustado$	**RSE**	**df**	**Curvas Influyentes**	**Atípicos**	**# Coeficientes Significativos**
25	0,27	0,75	19	21 y 25	21 y 25	0
23	0,67	0,52	17	0	0	4

TABLA 4.3. *FLR* con Representación en Bases usando la Temperatura.

La aplicación de este modelo indica que si no se involucran las curvas influyentes y atípicas (Loja y Shushufindi) el valor de $R^2 ajustado$ aumenta a $67\,\%$, es decir, se tiene una mejor explicación de la variación de la Precipitación a través de la Temperatura con un error bajo. En la tabla (4.3) se muestra que, existen 4 coeficientes

estimados $\hat{b}_j$ significativos en la regresión, ya que el $valor\ p$ de cada uno de ellos es menor que la significancia $\alpha = 0,05$.

Los resultados del segundo modelo, planteado en (3.3), obtenidos mediante la función `fregre.pc` se muestran en el Anexo 3. El resumen se ve a continuación:

Modelo FLR con Bases FPC: Temperatura						
# Curvas	$R^2ajustado$	**RSE**	**df**	**Curvas Influyentes**	**Atípicos**	**# Coeficientes Significativos**
25	0,29	0,73	21	21 y 25	21 y 25	1
23	0,76	0,44	17	0	0	3

TABLA 4.4. *FLR* con Bases *FPC* usando la Temperatura.

Al aplicar el modelo *FLR* con bases *FPC* se aprecia que al no tomar en cuenta las curvas atípicas el valor de $R^2ajustado$ alcanza un $76\,\%$, es decir, que la variación de la Precipitación es explicada a través de la Temperatura con un mayor porcentaje. En la tabla (4.4) se tiene que, aumenta a 3 la cantidad de coeficientes estimados $\hat{b}_{j_{FPC}}$ significativos en la regresión.

Es importante mencionar que para el desarrollo de este modelo se deben elegir las componentes de la regresión, por lo que se considera un subconjunto óptimo de 5 de componentes *FPC* por medio de validación cruzada MSC que se estudió en la Sección 3.3 con el criterio SIC. Se pueden ver los resultados en la siguiente tabla:

En la tabla (4.5) se tiene que la primera componente *FPC* no es significativa en ninguno de los casos, siendo la que mayor porcentaje de variación tiene en comparación al resto. También se ve que al no considerar los datos atípicos las componentes $2, 3$ y 5 tienen significancia explicando apenas el $5,34\,\%$.

Los resultados del tercer modelo, expuesto en la ecuación (3.5), obtenidos mediante la función `fregre.pls` se pueden apreciar en el Anexo 4. La tabla de resumen es la siguiente:

Validación Cruzada para FPC: Temperatura				
	25 Curvas		23 Curvas	
	% Variabilidad	$valor\ p$	% Variabilidad	$valor\ p$
FPC 1	91,94 %	0,09	90,76 %	0,17
FPC 2	3,39 %	0,00	4,74 %	0,00
FPC 3	0,73 %	0,33	0,55 %	0,00
FPC 4	2,06 %	0,11	2,33 %	0,58
FPC 5	0,29 %	0,29	0,05 %	0,00
Total	98.41 %		98.43 %	

TABLA 4.5. Porcentaje de variación de las componentes *FPC* y su respectivo $valor\ p$ en el caso de la Temperatura.

Modelo FLR con Bases FPLS: Temperatura						
# Curvas	$R^2 ajustado$	**RSE**	**df**	**Curvas Influyentes**	**Atípicos**	**# Coeficientes Significativos**
25	0,73	0,27	18	21 y 25	21 y 25	2
23	0,85	0,21	13	0	0	3

TABLA 4.6. *FLR* con Bases *FPLS* usando la Temperatura.

La aplicación del modelo *FLR* con bases *FPLS* presenta un $R^2 ajustado$ aún mejor cuando no se involucran las curvas atípicas, es decir, que la variación de la Precipitación es explicada en un $85\,\%$ a través de la Temperatura. En la tabla (4.6) se tiene que, existen 3 coeficientes estimados $\hat{b}_{j_{FPLS}}$ significativos en la regresión.

De manera análoga, para el desarrollo de este modelo se considera un subconjunto óptimo de 5 componentes *FPLS* por medio de la validación cruzada MSC con el criterio SIC. Los resultados obtenidos se pueden ver en la tabla (4.7).

Se observa que el porcentaje de variación total de las componentes llega hasta el $99,23\,\%$ al no tomar en cuenta las curvas atípicas. En la tabla (4.7) se ve que, las componentes $2, 3$ y 4 tienen mayor significancia, pero tan solo explican el $8,56\,\%$.

	Validación Cruzada para FPLS: Temperatura			
	25 Curvas		23 Curvas	
	% Variabilidad	$valor\ p$	% Variabilidad	$valor\ p$
FPLS 1	90,85 %	0,56	89,38 %	0,08
FPLS 2	4,21 %	0,02	5,61 %	0,00
FPLS 3	0,64 %	0,05	0,52 %	0,00
FPLS 4	1,94 %	0,08	2,43 %	0,00
FPLS 5	0,89 %	0,00	1,29 %	0,60
Total	98.53 %		99.23 %	

TABLA 4.7. Porcentaje de variación de las componentes *FPLS* y su respectivo $valor\ p$ en el caso de la Temperatura.

4.3.2. MODELOS CON LA VELOCIDAD DEL VIENTO COMO COVARIABLE FUNCIONAL

De manera similar, se utilizan los modelos propuestos pero considerando a la Velocidad del Viento como única covariable explicativa funcional.

En este caso, en el Anexo 5 se detallan los resultados obtenidos del primer modelo propuesto en (3.2). El resumen se aprecia en la siguiente tabla:

Modelo FLR con Bases Funcionales: Velocidad del Viento						
# Curvas	$R^2 ajustado$	**RSE**	**df**	**Curvas Influyentes**	**Atípicos**	**# Coeficientes Significativos**
25	0,79	0,40	19	21	0	3
24	0,80	0,38	18	0	0	5

TABLA 4.8. *FLR* con Representación en Bases usando la Velocidad del Viento.

En este modelo no se presentan curvas atípicas, pero si existe una curva (Loja) influyente en la estimación. En la tabla (4.8) se ve que, al no implicar dicha curva se obtienen mejores resultados, donde el $R^2 ajustado$ alcanza el $80\,\%$ de explicación de la variación de la Precipitación a través de la Velocidad del Viento, y existe una mayor cantidad de 5 coeficientes estimados $\hat{b}_j$ significativos en la regresión.

En el Anexo 6 se pueden apreciar los resultados obtenidos del segundo modelo planteado en (3.3). La tabla de resumen es la siguiente:

Modelo FLR con Bases FPC: Velocidad del Viento						
# Curvas	$R^2 ajustado$	**RSE**	**df**	**Curvas Influyentes**	**Atípicos**	**# Coeficientes Significativos**
25	0,62	0,54	19	21	0	3
24	0,65	0,52	18	0	0	3

TABLA 4.9. *FLR* con Bases *FPC* usando la Velocidad del Viento.

En este caso, en la tabla (4.9) se ve que al no involucrar la curva influyente la variación de la Precipitación es explicada en un $65\,\%$ a través de la Velocidad del Viento, y se mantiene en 3 la cantidad de coeficientes estimados $\hat{b}_{j_{FPC}}$ significativos.

Análogamente, se elige un subconjunto óptimo de 5 de componentes *FPC* por medio de la validación cruzada. Se puede ver los resultados en la siguiente tabla:

Validación Cruzada para FPC: Velocidad del Viento				
	25 Curvas		**24 Curvas**	
	% Variabilidad	$valor\ p$	**% Variabilidad**	$valor\ p$
FPC 1	89,07 %	0,00	92,52 %	0,00
FPC 2	6,48 %	0,23	3,28 %	0,13
FPC 3	0,68 %	0,47	0,53 %	0,85
FPC 4	0,31 %	0,00	0,34 %	0,00
FPC 5	0,25 %	0,02	0,20 %	0,02
Total	96,79 %		96,87 %	

TABLA 4.10. Porcentaje de variación de las componentes *FPC* y su respectivo $valor\ p$ en el caso de la Velocidad del Viento.

Se determina en la tabla (4.10) que la primera componente *FPC* es significativa en ambos casos. De modo que, a partir de las componentes $1, 4$ y 5 el modelo que no involucra la curva influyente alcanza el $93,06\,\%$ de representación de la variabilidad.

Finalmente, los resultados del tercer modelo expuesto en (3.5) se pueden observar

en el Anexo 7. El resumen se tiene en la siguiente tabla:

Modelo FLR con Bases FPLS: Velocidad del Viento						
# Curvas	$R^2 ajustado$	**RSE**	**df**	**Curvas Influyentes**	**Atípicos**	**# Coeficientes Significativos**
25	0,80	0,20	18	21	0	2
24	0,85	0,16	16	0	0	4

TABLA 4.11. *FLR* con Bases *FPLS* usando la Velocidad del Viento.

Al aplicar este modelo se determina un mejor $R^2 ajustado$ al no implicar la curva influyente. En la tabla (4.11) se tiene que, la variación de la precipitación es explicada en un $85\,\%$ a través de la Velocidad del Viento, y se tiene 4 coeficientes estimados $\hat{b}_{j_{FPLS}}$ significativos en la regresión.

Se considera un subconjunto óptimo de 5 de componentes *FPLS* por medio de la validación cruzada. Los resultados obtenidos se pueden ver en la siguiente tabla:

Validación Cruzada para FPLS: Velocidad del Viento				
	25 Curvas		**24 Curvas**	
	% Variabilidad	$valor\ p$	**% Variabilidad**	$valor\ p$
FPLS 1	90,13 %	0,01	91,68 %	0,00
FPLS 2	7,22 %	0,09	6,21 %	0,70
FPLS 3	0,78 %	0,00	0,62 %	0,00
FPLS 4	0,48 %	0,53	0,51 %	0,03
FPLS 5	0,36 %	0,11	0,29 %	0,01
Total	98.97 %		99,31 %	

TABLA 4.12. Porcentaje de variación de las componentes *FPLS* y su respectivo $valor\ p$ en el caso de la Velocidad del Viento.

S observa en la tabla (4.12) que en ambos casos la primera componente *FPLS* es significativa. Así, a partir de las componentes $1, 3, 4$ y 5 el modelo que no involucra la curva influyente alcanza el $93,1\,\%$ de representación de la variabilidad.

4.3.3. MODELO CON DOS COVARIABLES FUNCIONALES

Se realiza la adaptación de un modelo *FLR* con representación en Bases Funcionales considerando dos covariables explicativas funcionales para la estimación de la Precipitación; este modelo se ha planteado en la ecuación (3.7), y cuyos resultados se ven en el Anexo 8. A continuación, se tiene la siguiente tabla de resumen:

Modelo FLR con Bases Funcionales: Dos Covariables						
# Curvas	$R^2ajustado$	**RSE**	**df**	**Curvas Influyentes**	**Atípicos**	**# Coeficientes Significativos**
25	0,92	0,24	14	21 y 25	21 y 25	3
23	0,97	0,14	12	0	0	6

TABLA 4.13. *FLR* con Representación en Bases Funcionales usando la Temperatura y la Velocidad del Viento.

En este caso, se han encontrados 2 curvas influyentes que son las curvas atípicas halladas en el caso de la Temperatura, las cuales se han dejado de lado. En la tabla (4.13) se aprecia que existe una cantidad de 6 coeficientes estimados $\hat{b}_j$ significativos en la regresión, y el valor de $R^2ajustado$ más alto con el $97\,\%$ de variación de la Precipitación explicada a través de la Temperatura y Velocidad del Viento.

Se procede a realizar una comparación entre de todos los modelos mediante indicadores estadísticos que permitirán escoger el mejor se ajuste.

4.3.4. EVALUACIÓN Y VALIDACIÓN DE LOS MODELOS FLR

Los resultados obtenidos permiten comparar los ajustes de cada modelo, de esta forma podrá escoger el mejor de los ajustes para nuestros propósitos de predicción. Cabe recalcar que se compararán únicamente los modelos en los que no intervienen datos atípicos e influyentes en la estimación.

Se presentan las medidas estadísticas mostradas en la Sección 3.4 para evaluar el rendimiento y la capacidad predictiva de cada uno de los modelos:

Medidas de evaluación y comparación							
Variable	**Modelo FLR**	MAE	$PRESS$	$MAPE$	AIC	AIC_C	BIC
Temperatura	**Bases Funcionales**	0,66	15,89	22,28	57,82	54,69	65,77
	FPC	0,60	13,50	20,34	57,53	54.40	65,48
	FPLS	0,60	12,01	21,54	57,11	53,98	65,05
Velocidad del Viento	**Bases Funcionales**	1,11	45,57	27,50	57,69	54,36	63,84
	FPC	1,10	44,98	27,66	57,53	54,19	63,68
	FPLS	1,10	44,69	28,01	57,45	54,12	63,60
Temperatura y Velocidad del Viento	**Bases Funcionales**	0,23	2,08	8,31	65,56	66,13	79,19

TABLA 4.14. Medidas de evaluación y comparación de los modelos FLR

En primer lugar, al tomar en cuenta el error absoluto medio MAE para el pronostico se tiene que, en el caso de los modelos que consideran una sola covariable funcional, el menor valor viene dado por el modelo *FLR-FPC* para la Temperatura; si se considera la suma de cuadrados de error de predicción $PRESS$ el modelo *FLR-FPLS* para la Temperatura es el que menor valor presenta. Ahora, el error porcentual absoluto medio $MAPE$, que expresa la exactitud como un porcentaje del error, indica que el modelo *FLR-FPC* para la Temperatura tiene un $20,34\,\%$ de error, el cual es el menor entre todos los modelos.

Continuando con la comparación, los criterios de Akaike y Akaike corregido indican que el mejor ajuste se da por el modelo *FLR-FPLS* que utiliza la Temperatura como única covariable explicativa, pues tiene el menor valor de AIC y AIC_c. Finalmente, el criterio de información Bayesiano indica que el mejor ajuste se da con el modelo *FLR-FPLS* que considera como covariable explicativa a la Velocidad del Viento, ya que el valor de BIC es el menor.

Si de elegir un modelo con una sola covariable explicativa se tratase, gracias a la comparación realizada, se determina que el mejor ajuste está dado por los modelos *FLR-FPLS*. Basándose en estas medidas el mejor ajuste sobre la Precipitación se da por este modelo que usa a la Temperatura como única covariable funcional.

Evaluando en conjunto todos los modelos, se puede apreciar en la tabla 4.14 que

los valores de MAE y $PRESS$ son menores para el modelo que involucra simultáneamente a la Temperatura y Velocidad del Viento como covariables funcionales explicativas, y el cual tiene el $8,31\,\%$ como menor error porcentual absoluto medio $MAPE$; a pesar de que los criterios de información AIC, AIC_C y BIC son mayores al considerar dos covariables explicativas no son excesivamente mayores al resto, esto se debe al hecho de que se ha utilizado mayor información muestral (dos variables a la vez) para hallar sus resultados.

Se puede concluir que el mejor ajuste para la estimación de la Precipitación se da cuando se implican estas dos covariables funcionales sin la implicación de las curvas atípicas. Este es el modelo que se usará para las predicciones.

Ahora, se verifican los supuestos establecidos en la Sección 3.5 con el objetivo de validar el modelo seleccionado. Los resultados se muestran a continuación:

Supuesto de Normalidad en los Residuos Se usa el método gráfico como herramienta para intuir previamente el comportamiento de los residuos:

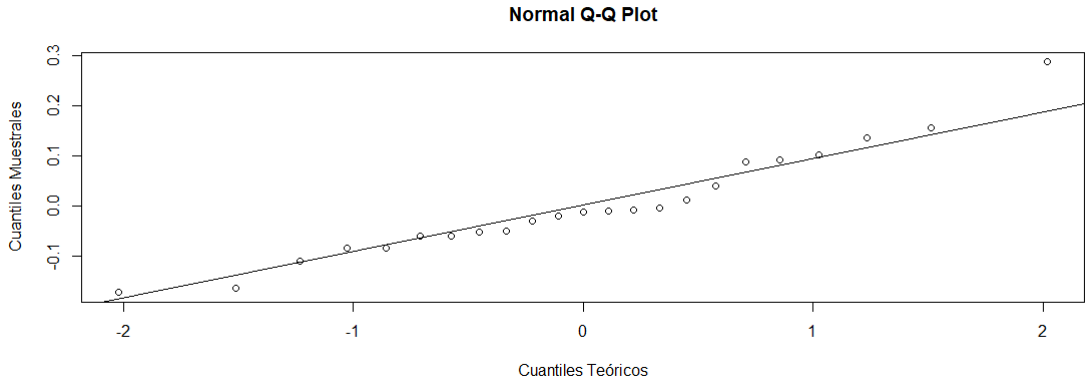

FIGURA 4.15. Comportamiento de los residuos mediante cuantiles.

La probabilidad acumulada teórica que corresponde a cada puntuación típica en una curva normal con media 0 y desviación típica 1 se observa en la figura (4.15), y se aprecia que los residuos siguen un comportamiento adecuado en torno a la línea normal, por lo que, parece indicar que los residuos se distribuyen normalmente.

Analíticamente se usa el test de Shapiro-Wilks mediante la función `shapiro.test`. El resultado arrojado da un $valor\ p = 0,28$ asociado al estadístico de prueba que es mayor al nivel de significancia $\alpha = 0,05$. Así, no se puede rechazar H_o asegurando que los residuos cumplen con el supuesto de normalidad.

Supuesto de Autocorrelación, Independencia en los Residuos Primero se analiza la gráfica de los residuos dada a continuación:

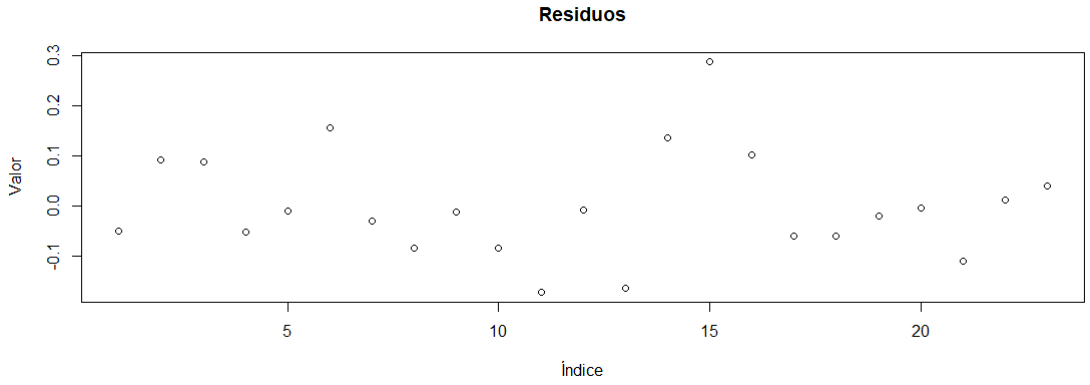

FIGURA 4.16. Residuos del modelo *FLR* con dos covariables funcionales.

En la figura (4.16) se trata de detectar algún patrón en el comportamiento de los residuos. En este caso no se visualiza la existencia de ningún patrón o tendencia evidente en los datos.

Este supuesto se verifica analíticamente mediante los test de Box-Pierce $\&$ Ljung-Box. Para esto, se utiliza la función `Box.test` donde se especifica el `type` como `"Box-Pierce"` y `"Ljung-Box"` respectivamente. En el caso de Box-Pierce se obtiene un $valor\ p = 0,19$ y para el caso de Ljung-Box un $valor\ p = 0,17$, los cuales están asociados a sus estadísticos de prueba respectivos, y en ambos casos son mayores a la significancia $\alpha = 0,05$. Por lo tanto, no se rechaza la hipótesis H_o determinando que los residuos son independientes.

Supuesto de Homocedasticidad de los Residuos Por lo general, la variabilidad de los residuos se encuentra en función de las variables explicativas, se analiza los

valores ajustados (pronósticos) vs. residuos a continuación:

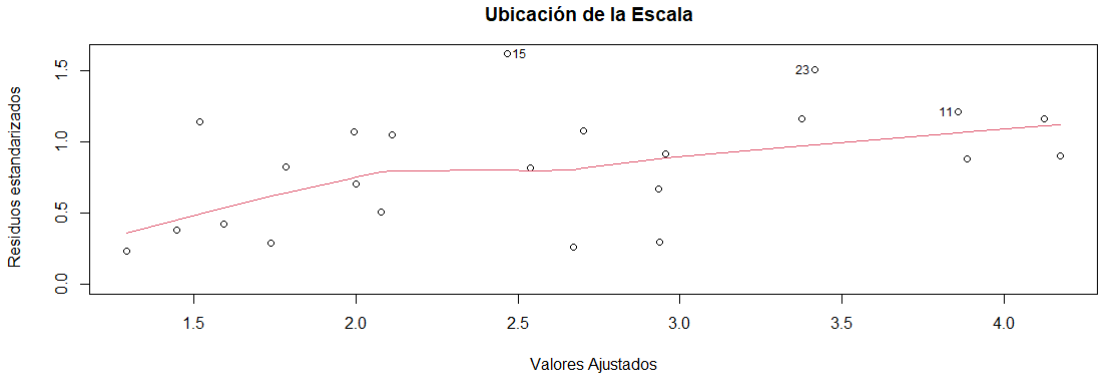

FIGURA 4.17. Residuos de los valores ajustados del modelo *FLR* con dos covariables funcionales.

Este supuesto se acepta si los residuos se distribuyen de forma aleatoria, tal y como se puede observar en la figura (4.17) una nube de puntos de forma similar en todo el rango de las observaciones.

Analíticamente se verifica por el test de Breusch-Pagan con la función `bptest`. El resultado proporcionado por esta prueba da un $valor\ p = 0,28$ asociado al estadístico de prueba mayor que la significancia $\alpha = 0,05$. De manera que, no se rechaza H_o concluyendo que los residuos tienen varianza constante.

Supuesto de Especificación o Linealidad del Modelo Para verificar este supuesto se realiza el test Ramsey que se usa para detectar los errores de especificación ocasionados por: omisión de covariables explicativas, existencia de correlación entre las covariables en explicativas (multicolinealidad) o bien, porque no es apropiada la forma funcional de las covariables explicativas.

La función `resettest` proporciona un $valor\ p = 0,41$ que es mayor a la significancia $\alpha = 0,05$. Por lo que, no se rechaza H_o concluyendo que el modelo tiene una correcta especificación, es decir, el modelo es lineal.

4.4. DIAGNÓSTICO DEL BALANCE HÍDRICO PARA ANALIZAR LA PRODUCTIVIDAD DEL MAÍZ

En esta sección se realiza la predicción deseada sobre la Precipitación para determinar el balance hídrico y los requerimientos de riego en las zonas de mayor producción de maíz duro seco en Ecuador con el fin de comprender la productividad de este cultivo. Para ello, se utiliza el modelo *FLR* que considera dos covariables explicativas funcionales, el cual arrojó la mejor estimación sobre la variable de respuesta escalar y cumple los supuestos de validación.

4.4.1. PREDICCIÓN DE LA PRECIPITACIÓN

Se aplica el modelo mostrado en la Sección 4.3.3 en cantones de las provincias con productividad de maíz duro seco. Los cantones basados en la tabla (1.1) son:

Provincia	**Cantones**	**Coordenadas**	
		Latitud	**Longitud**
Santo Domingo de los Tsáchilas	Santo Domingo	-0,25	-79,17
Los Ríos	Valencia	-0,95	-79,35
	Palenque	-1,43	-79,74
Esmeraldas	Eloy Alfaro	1,25	-79,00
Guayas	Salitre	-1,82	-79,81
Manabí	San Vicente	-0,58	-80,40

TABLA 4.15. Cantones seleccionados en las provincias con productividad de maíz duro seco del país para predecir la Precipitación.

Se recolecta la información meteorológica de la Temperatura y Velocidad del Viento, entre los años 2010 y 2020, en los lugares de la tabla (4.15). Luego, de estos datos se realiza un promedio diario obteniendo una estructura matricial de 365 observaciones (días del año) y 6 columnas (cantones) de cada variable. Antes de usar el modelo se construye una estructura funcional de las variables recogidas, por lo que, se realiza el suavizamiento con 12 y 11 bases de Fourier de la Temperatura y la Velocidad del

Viento respectivamente y se muestran a continuación:

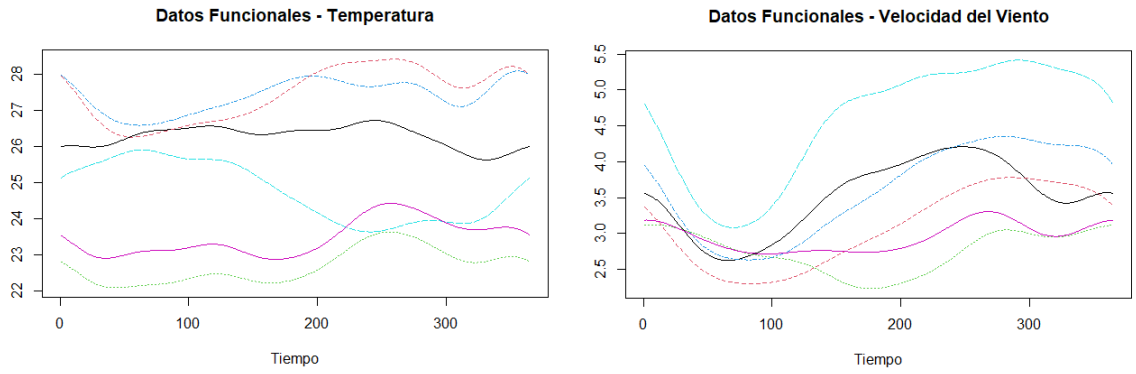

FIGURA 4.18. Datos suavizados del promedio diario de la Temperatura (Izquierda) y Velocidad del Viento (derecha) de 6 cantones para predecir la Precipitación.

Por último, se utiliza el modelo para obtener la predicción de la Precipitación. Los resultados se reflejan de acuerdo con su ubicación geográfica en el siguiente mapa:

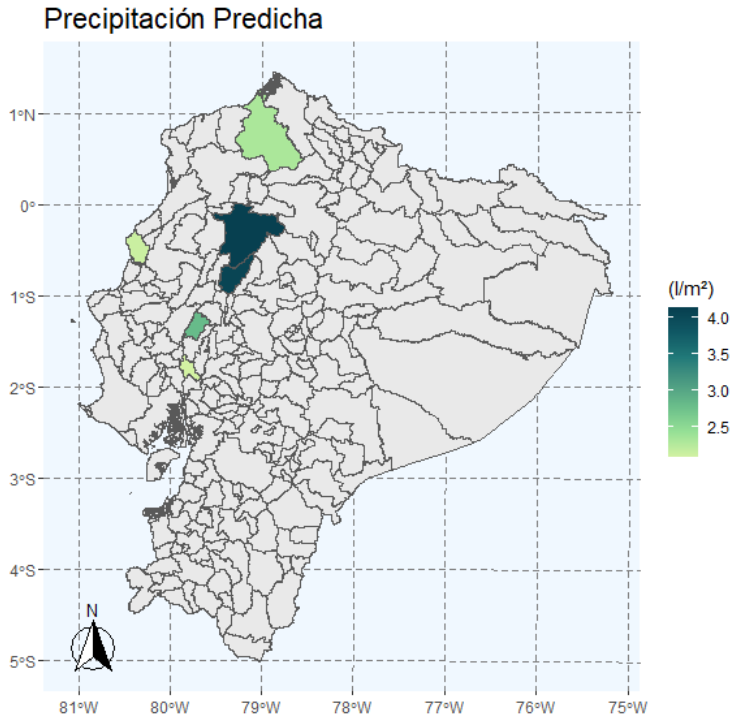

FIGURA 4.19. Predicción de la Precipitación en los 6 cantones seleccionados.

Se puede apreciar que los niveles de Precipitación varían respecto a la zona en la que se encuentran. Los resultados de la predicción se ven a continuación:

Predicción de la Precipitación usando dos covariables	
Cantón	**Precipitación** (mm/d)
Santo Domingo	4,14
Valencia	4,07
Palenque	2,82
Eloy Alfaro	2,32
Salitre	2,09
San Vicente	2,15

TABLA 4.16. Valores de predicción de la Precipitación.

En la Sección 4.3.4 se verificó que los errores obtenidos por el modelo *FLR* con Bases Funcionales usando dos covariables explicativas funcionales son muy bajos, por lo que, los niveles predichos de la precipitación mantienen la misma naturaleza dentro de cada provincia.

En Santo Domingo se tiene una predicción de $4,14\ mm$ diarios, que resulta estar conforme al clima de esta provincia. En Los Ríos donde se hallan los cantones Valencia con $4,07\ mm$ y Palenque con $2,82\ mm$ al día son acordes a los niveles de precipitación expuestos en la figura (1.5). En el caso de Eloy Alfaro la precipitación predicha de $2,32\ mm$ está dentro de la variación de los niveles visibles en la figura (1.3) de la izquierda que corresponde a Esmeraldas. En Salitre la predicción de $2,09\ mm$ diarios se halla entre los límites presentes en la provincia del Guayas como se evidencia en la figura (1.4) de la derecha. Para San Vicente, ubicado en Manabí, la predicción de $2,15\ mm$ por día se encuentra entre los valores de precipitación como se observa en la figura (1.3) de la derecha.

Para continuar el análisis, se procede a encontrar los indicadores del balance hídrico y los requerimientos de riego en los cantones indicados ya que están ubicados en las provincias idóneas para entender la productividad de maíz.

4.4.2. PRODUCTIVIDAD DEL MAÍZ

Es claro que los niveles de producción de una planta dependen de la zona agro-meteorológica donde se cultiva. En Segura & Andrade (2011) se indica que una mayor longitud, peso y diámetro de la mazorca influyen directamente para obtener los más altos rendimientos, es decir, si las magnitudes indicadas aumentan entonces incrementa el número y tamaño de granos por mazorca por unidad de superficie, esto significa que el rendimiento de la producción es más eficiente. Bajo este contexto, un adecuado desarrollo de las plantaciones de maíz, en promedio, puede conseguir los siguientes resultados: longitud de $21,5\ cm$, peso de $257,3\ g$ y diámetro de $4,7\ cm$.

La productividad del cultivo puede verse reflejada en la tasa de crecimiento del cultivo, por lo que, las plantas de maíz que alcanzan alturas más elevadas tienen rendimientos potenciales más altos. Cabe señalara también que la productividad se ve estrechamente relacionada con el balance hídrico que se da en las zonas de interés. (Segura & Andrade, 2011). En este sentido, las precipitaciones durante al desarrollo del cultivo son de gran importancia, por la cantidad de agua de lluvia almacenada en el suelo, pues tienen influencia directa en el correcto o erróneo crecimiento de las plantaciones. Esta relación se puede comprender a través de ciertas condiciones agro-meteorológicas que deben cumplirse para que la productividad del cultivo sea adecuada.

En el caso del maíz duro seco el requerimiento mínimo de agua durante todo el ciclo de cultivo, el cual puede durar entre 115 y 125 días, requiere de al menos $500\ mm$, pero lo óptimo para evitar cualquier eventualidad debería rondar los $750\ mm$. A continuación, se indica que la distribución del requerimiento de agua dado por Segura & Andrade (2011):

- Fase de germinación: A partir de la siembra, en el período de $0-5$ días, se necesita del $5\,\%$ del requerimiento total del agua.
- Fase de desarrollo vegetativo: Entre los días $5-35$ el requerimiento del agua debe ser el $23\,\%$ de la necesidad total de agua.

- Fase de prefloración: Entre los días $35 - 55$ el requerimiento del agua también es del $23\,\%$ de la necesidad total de agua del cultivo. En esta etapa se determinan los rendimientos de maíz en la zona de estudio.
- Fase de la floración: Se da entre los días $56 - 70$ se hace necesario el $14\,\%$ de la necesidad total del agua del cultivo.
- Fase de llenado de grano: Entre los días $71 - 95$ el requerimiento del agua es del $34\,\%$ del total necesitado por el cultivo.
- Maduración fisiológica: Entre los $96 - 125$ días los requerimientos de agua son mínimos ya que se considera el secado del grano, este no supera el $1\,\%$ del total requerido por el cultivo, y se da por que el suelo tiene suficiente humedad.

Estas fases están implicadas en las etapas mostradas en la Sección 1.4.2 por las que pasa el ciclo del cultivo de maíz, y se muestran a continuación:

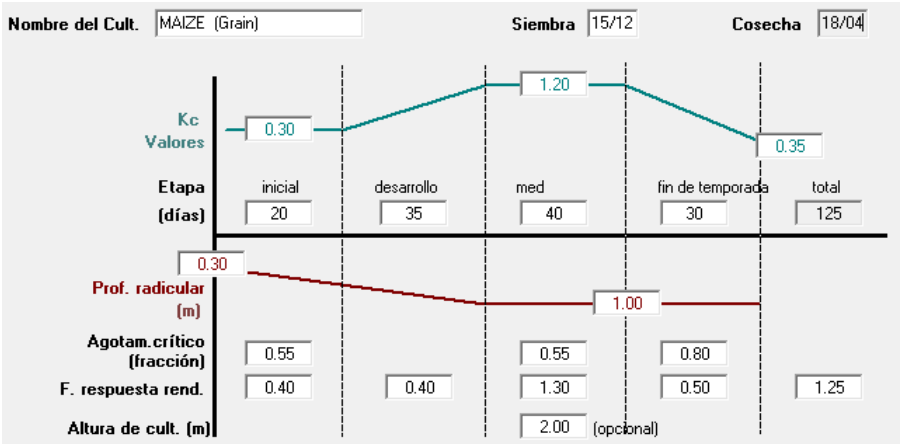

FIGURA 4.20. Etapas del ciclo de cultivo para el maíz duro seco.

La época de siembra del maíz varía respecto al período de lluvias. En general, comienza a finales de diciembre y avanza hasta mediados de abril coincidiendo con periodos de lluvia en las zonas de cultivo. En la figura (4.20) se muestra que el ciclo de 125 días empieza el 15 de diciembre y termina el 18 de abril, período en el cual se involucran los coeficientes K_c de cada etapa.

Ahora, se hallan las necesidades hídricas en las zonas de estudio a través de las precipitaciones predichas, cuyos valores se muestran en la tabla (4.16), esto se

realiza con ayuda del programa `Cropwat`. Para ello, se inicia con el cálculo de la evapotranspiración de referencia ET_o, los resultados se muestran en el Anexo 9. Luego, se usa la precipitación predicha en los meses establecidos para determinar la precipitación efectiva del cultivo, que es la fracción de la precipitación predicha, la cual aprovecha realmente la planta para su desarrollo.

En Santo Domingo y Valencia se tienen los siguientes resultados:

Meses	**Santo Domingo**			**Valencia**		
	ET_o mm/d	**Precipitación** mm	**Prec. Efectiva** mm	ET_o mm/d	**Precipitación** mm	**Prec. Efectiva** mm
Diciembre	4,31	128,34	78,60	4,67	126,20	100,70
Enero	3,58	128,34	78,60	4,22	126,20	100,70
Febrero	3,48	115,92	68,70	3,78	114,00	93,20
Marzo	3,40	128,34	78,60	3,73	126,20	100,70
Abril	3,33	124,20	75,40	3,53	122,10	98,20
Mayo	3,19	128,30	78,60	3,69	126,20	100,70
Junio	3,20	124,20	75,40	3,43	122,10	98,20
Julio	3,60	37,10	12,30	4,12	26,40	25,30
Agosto	4,45	19,40	1,60	4,63	15,20	14,80
Septiembre	5,00	29,10	7,50	5,27	18,90	18,30
Octubre	4,89	59,20	25,50	4,91	40,30	37,70
Noviembre	4,67	64,00	28,40	4,84	51,00	46,80
Promedio	3,93	90,53	50,78	4,24	84,57	69,61

TABLA 4.17. Balance hídrico en los cantones Santo Domingo y Valencia.

El índice ET_o dado en la ecuación (1.1) representa el primer acercamiento para encontrar los requerimientos hídricos requeridos. Se puede evidenciar en la primera y cuarta columna de la tabla (4.17) respectivamente, tanto para Santo Domingo como para Valencia, que los niveles hídricos ideales de ET_o son suficientes en los meses, en los que se procede con la siembra hasta la cosecha del maíz, comprendidos entre la segunda quincena de diciembre hasta mediados de abril. Esto se debe a que, los valores predichos de precipitación satisfacen las condiciones hídricas ideales la mayoría del tiempo.

En Santo Domingo se predijo un valor de $4,14\ mm$ que satisface un máximo de $3,58\ mm$ de precipitación por día entre enero y abril, pero en diciembre hay una pe-

queña diferencia de $0,17\ mm$. Por su parte, en Valencia se obtuvo una predicción de $4,07\ mm$ suficiente para cubrir un máximo de $3,58\ mm$ por día entre febrero y abril, mientras que en diciembre y enero idealmente se necesitaría cubrir la diferencia de $0,6\ mm$ y $0,15\ mm$ respectivamente.

Es esencial comprender que el valor de precipitación mostrado en la tabla (4.17) es el valor predicho transformado en un valor mensual. Esta transformación se realizó solo en los meses en que se cultiva el maíz, ya que en los meses posteriores no se realiza actividad agrícola y el resto de los valores son datos reales del promedio de la precipitación mensual de una década.

En el caso de Santo Domingo la precipitación efectiva que se necesita mientras se cultiva el maíz varía entre $64,7\ mm$ y $78,6\ mm$ de precipitación por mes. En Valencia se tiene una oscilación de precipitación entre $93,2\ mm$ y $100,7\ mm$ por mes.

Los resultados para los cantones de Palenque y Eloy Alfaro se ven a continuación:

Meses	**Palenque**			**Eloy Alfaro**		
	ET_o	**Precipitación**	**Prec. Efectiva**	ET_o	**Precipitación**	**Prec. Efectiva**
	mm/d	mm	mm	mm/d	mm	mm
Diciembre	3,67	87,40	75,20	3,47	71,90	63,60
Enero	3,68	87,40	75,20	3,54	71,90	63,60
Febrero	3,47	79,00	69,00	3,68	65,00	58,20
Marzo	3,31	87,40	75,20	3,74	71,90	63,60
Abril	3,43	84,60	73,10	3,75	69,60	61,80
Mayo	3,19	68,60	61,10	3,40	107,90	89,20
Junio	3,07	22,10	21,30	3,35	62,70	56,40
Julio	3,37	5,70	5,60	3,76	40,90	38,30
Agosto	3,73	2,70	2,70	4,22	21,90	21,10
Septiembre	3,83	2,80	2,80	4,33	29,60	28,20
Octubre	3,79	8,70	8,60	4,17	71,00	63,00
Noviembre	3,66	12,00	11,80	3,76	104,40	87,00
Promedio	3,52	45,70	40,13	3,76	65,73	57,83

TABLA 4.18. Balance hídrico en los cantones Palenque y Eloy Alfaro.

En Palenque se tuvo una predicción baja de $2,82\ mm$ que no satisface las condiciones ideales con un valor mínimo de $3,31\ mm$ por día entre diciembre y abril. En Eloy Alfaro también se encontró una perdición de $2,32\ mm$ por día que no es suficiente

para cubrir un mínimo de $3,47\ mm$ por día diciembre y abril. De manera que, en ambos casos no se halla un balance hídrico ideal a través de ET_o.

En la tabla (4.18) se tiene que, la precipitación efectiva en Palenque necesaria mientras se cultiva el maíz oscila entre $69\ mm$ y $75,2\ mm$ de precipitación por mes. Por su lado, en Eloy Alfaro hay una fluctuación entre $58,2\ mm$ y $63,6\ mm$ por mes.

En el caso de Salitre y San Vicente se presentan los siguientes resultados:

Meses	**Salitre**			**San Vicente**		
	ET_o mm/d	**Precipitación** mm	**Prec. Efectiva** mm	ET_o mm/d	**Precipitación** mm	**Prec. Efectiva** mm
Diciembre	6,09	64,80	58,10	3,58	66,70	59,50
Enero	5,09	64,80	58,10	3,39	66,70	59,50
Febrero	4,11	58,50	53,00	3,36	60,20	54,40
Marzo	3,86	64,80	58,10	3,43	66,70	59,50
Abril	4,18	62,70	56,40	3,32	64,50	57,80
Mayo	4,40	53,90	49,30	3,17	68,40	60,90
Junio	4,98	15,20	14,80	3,10	21,70	21,00
Julio	5,80	4,10	4,10	3,44	6,10	6,00
Agosto	6,47	3,00	3,00	3,36	2,60	2,60
Septiembre	6,75	2,30	2,30	3,46	3,20	3,20
Octubre	6,34	6,90	6,90	3,61	6,10	6,10
Noviembre	6,30	8,20	8,10	3,63	6,90	6,80
Promedio	5,36	34,10	31,02	3,40	36,65	33,11

TABLA 4.19. Balance hídrico en los cantones Salitre y San Vicente.

En Salitre se predijo una precipitación de $2,09\ mm$ por día, la cual no cubre el valor de ET_o entre diciembre y abril para tener un balance hídrico ideal. En San Vicente tampoco se satisfacen las condiciones ideales pues se obtuvo una predicción de $2,15\ mm$ que no cubre el valor mínimo de $3,31\ mm$ por día entre diciembre y abril.

La precipitación efectiva mostrada en la tabla (4.19) en el caso de Salitre necesaria mientras se cultiva el maíz fluctúa entre $58,2\ mm$ y $63,6\ mm$ de precipitación por mes. En el caso de San Vicente la oscilación de precipitación efectiva se encuentra entre $54,4\ mm$ y $59,5\ mm$ por mes.

A partir de los valores encontrados anteriormente el programa `Cropwat` permite hallar la evapotranspiración real ET_c y los requerimientos de riego en cada etapa de crecimiento del cultivo. De modo que, se tiene un acercamiento más fino para comprender las necesidades hídricas del cultivo.

Los resultados referentes a los requerimientos hídricos están agrupados por periodos de 10 días en cada uno en los meses de cultivo del maíz. De aquí, se dividen en 4 etapas de crecimiento: inicial, desarrollo, media y final. Cada etapa tiene una constante K_c que depende estrictamente del tipo del producto a sembrar, en este caso se cuentan los valores específicos para el cultivo del maíz como se muestra en la Sección 1.4.2.

En el caso de Santo Domingo se tiene la siguiente tabla de resumen:

Mes	**Período** $p = 10$ *días*	**Etapa**	**Kc**	ET_c mm/d	ET_c mm/p	**Prec. Efec** mm/p	**Req.Riego** mm/d
Dic	2	Inic	0,30	1,40	8,40	21,60	0,00
Dic	3	Inic	0,30	1,36	14,90	35,20	0,00
Ene	1	Des	0,37	1,61	16,10	33,80	0,00
Ene	2	Des	0,60	2,55	25,50	33,90	0,00
Ene	3	Des	0,86	3,50	38,50	32,90	0,80
Feb	1	Med	1,10	4,31	43,10	31,40	1,50
Feb	2	Med	1,15	4,34	43,40	30,40	1,80
Feb	3	Med	1,15	4,32	34,60	31,40	0,80
Mar	1	Med	1,15	4,31	43,10	33,00	1,40
Mar	2	Fin	1,15	4,28	42,80	34,00	1,20
Mar	3	Fin	0,96	3,53	38,80	33,60	0,90
Abr	1	Fin	0,68	2,46	24,60	32,90	0,00
Abr	2	Fin	0,44	1,56	12,50	26,00	0,00

TABLA 4.20. Requerimiento de riego para la productividad de maíz duro seco en el cantón Santo Domingo.

En la tabla (4.20) se puede apreciar el cálculo de ET_c dada en la ecuación (1.2) por día y por período. Estos valores representan la cantidad de precipitación real que necesita el cultivo en cada etapa.

En Santo Domingo se tiene que, a partir del tercer período de enero hasta el tercer período de marzo, la precipitación predicha de $4,14\ mm$ no satisface el requerimiento de agua mínimo para efectuar de manera apropiada los procesos fisiológicos de la planta.

Además, se tiene el valor de la precipitación efectiva por período y el requerimiento de riego en mm por día que indican los valores reales de agua que deben ser cubiertos para que el maíz tenga una productividad óptima. Por ejemplo, entre los días 11 y 20 correspondientes al segundo período de la etapa media del mes de febrero se debería suplir con $1,8\ mm$ de agua por día para que la planta tenga buen desarrollo fisiológico y con ello alcanzar mejores niveles de capacidad productiva.

Para el cantón Valencia se tiene los siguientes resultados:

Mes	**Período** $p = 10$ *días*	**Etapa**	**Kc**	ET_c mm/d	ET_c mm/p	**Prec. Efec** mm/p	**Req.Riego** mm/d
Dic	2	Inic	0,30	1,40	8,40	21,60	0,00
Dic	3	Inic	0,30	1,36	14,90	35,20	0,00
Ene	1	Des	0,37	1,61	16,10	33,80	0,00
Ene	2	Des	0,60	2,55	25,50	33,90	0,00
Ene	3	Des	0,86	3,50	38,50	32,90	0,60
Feb	1	Med	1,10	4,31	43,10	31,40	1,20
Feb	2	Med	1,15	4,34	43,40	30,40	1,30
Feb	3	Med	1,15	4,32	34,60	31,40	0,30
Mar	1	Med	1,15	4,31	43,10	33,00	1,00
Mar	2	Fin	1,15	4,28	42,80	34,00	0,90
Mar	3	Fin	0,96	3,53	38,80	33,60	0,50
Abr	1	Fin	0,68	2,46	24,60	32,90	0,00
Abr	2	Fin	0,44	1,56	12,50	26,00	0,00

TABLA 4.21. Requerimiento de riego para la productividad de maíz duro seco en el cantón Valencia.

Se observa en la tabla (4.21) que, en la etapa inicial, en los dos primeros periodos de la fase de desarrollo y los dos últimos periodos de la fase final se satisfacen las necesidades hídricas en Valencia. Mientras que, en los periodos restantes se deben suplir estos requerimientos; por ejemplo, el tercer período de enero entre los días

21 y 31 en la etapa de desarrollo se debe cubrir con $0,6\ mm$ de agua extra al día para una mejor productividad.

Para el cantón Palenque se tiene que, en las etapas iniciales y para los dos primeros periodos de la etapa de desarrollo las necesidades hídricas están satisfechas. Esto se muestra a continuación:

Mes	**Período** $p = 10$ *días*	**Etapa**	**Kc**	ET_c mm/d	ET_c mm/p	**Prec, Efec** mm/p	**Req,Riego** mm/d
Dic	2	Inic	0,30	1,10	6,60	16,80	0,00
Dic	3	Inic	0,30	1,10	12,10	27,10	0,00
Ene	1	Des	0,37	1,35	13,50	25,20	0,00
Ene	2	Des	0,60	2,19	21,90	25,30	0,00
Ene	3	Des	0,84	3,05	33,50	24,50	0,90
Feb	1	Med	1,08	3,82	38,20	23,30	1,49
Feb	2	Med	1,13	3,92	39,20	22,40	1,67
Feb	3	Med	1,13	3,85	30,80	23,30	0,75
Mar	1	Med	1,13	3,79	37,90	24,60	1,33
Mar	2	Fin	1,13	3,72	37,20	25,40	1,18
Mar	3	Fin	0,95	3,17	34,90	25,10	0,98
Abr	1	Fin	0,67	2,29	22,90	24,90	0,00
Abr	2	Fin	0,44	1,51	12,10	19,90	0,00

TABLA 4.22. Requerimiento de riego para la productividad de maíz duro seco en el cantón Palenque.

En la tabla (4.22) se observa que, a partir del tercer período de la etapa de desarrollo hasta el tercer período de la etapa final se requiere ciertos niveles extras de agua. Por ejemplo, en la etapa media entre los días 11 y 20 del mes de febrero se necesitaría la mayor cantidad adicional de agua, particularmente se debe suplir $1,67\ mm$ de agua por día durante dicha etapa.

En el cantón Eloy Alfaro se tiene el siguiente resumen:

Mes	**Período** $p = 10$ *días*	**Etapa**	**Kc**	ET_c mm/d	ET_c mm/p	**Prec, Efec** mm/p	**Req,Riego** mm/d
Dic	2	Inic	0,30	1,04	6,20	12,10	0,00
Dic	3	Inic	0,30	1,05	11,50	20,50	0,00
Ene	1	Des	0,37	1,29	12,90	21,30	0,00
Ene	2	Des	0,60	2,11	21,10	21,40	0,00
Ene	3	Des	0,85	3,03	33,30	20,70	1,26
Feb	1	Med	1,08	3,92	39,20	19,70	1,95
Feb	2	Med	1,13	4,15	41,50	18,90	2,12
Feb	3	Med	1,13	4,18	33,40	19,70	1,37
Mar	1	Med	1,13	4,20	42,00	20,80	2,26
Mar	2	Fin	1,13	4,21	42,10	21,50	2,06
Mar	3	Fin	0,95	3,55	39,00	21,20	1,78
Abr	1	Fin	0,68	2,53	25,30	19,90	0,54
Abr	2	Fin	0,44	1,65	13,20	15,40	0,00

TABLA 4.23. Requerimiento de riego para la productividad de maíz duro seco en el cantón Eloy Alfaro.

Los requerimientos hídricos para Eloy Alfaro correspondientes a la etapa inicial y de desarrollo están cubiertas; sin embargo, para la etapa media y final se requerirá agua adicional como se observa en la tabla (4.23). Por ejemplo, entre los días 1 y 10 del mes de marzo de la etapa media se necesitarían de al menos $2,26\ mm$ de agua extra por día para suplir las condiciones de esta etapa.

Los resultados obtenidos para el cantón Salitre, a diferencia de los casos anteriores, muestran que es un lugar geográfico con mayores requerimientos hídricos. Esto se evidencia en la siguiente tabla:

Mes	**Período** $p = 10$ *días*	**Etapa**	**Kc**	ET_c mm/d	ET_c mm/p	**Prec, Efec** mm/p	**Req,Riego** mm/d
Dic	2	Inic	0,30	1,83	11,00	13,00	0,01
Dic	3	Inic	0,30	1,73	19,00	20,90	0,00
Ene	1	Des	0,37	2,01	20,10	19,50	0,06
Ene	2	Des	0,61	3,12	31,20	19,60	1,17
Ene	3	Des	0,88	4,17	45,90	18,90	2,70
Feb	1	Med	1,12	4,98	49,80	17,90	3,19
Feb	2	Med	1,18	4,83	48,30	17,20	3,11
Feb	3	Med	1,18	4,73	37,90	17,90	1,99
Mar	1	Med	1,18	4,63	46,30	19,00	2,74
Mar	2	Fin	1,17	4,53	45,30	19,70	2,56
Mar	3	Fin	0,98	3,90	42,90	19,40	2,35
Abr	1	Fin	0,69	2,83	28,30	19,10	0,92
Abr	2	Fin	0,45	1,87	14,90	15,20	0,00

TABLA 4.24. Requerimiento de riego para la productividad de maíz duro seco en el cantón Salitre.

Las necesidades hídricas mostradas en la tabla (4.24) son las más altas, siendo en el primer período del mes de febrero la etapa que más agua necesita, pues se requieren diariamente $3,19\ mm$ extras por día seguido de $3,11\ mm$ por día en el segundo período. Esto se debe realizar con el fin de cubrir la falta de agua en cada una de las etapas y obtener mejores resultados de productividad.

Finalmente, para San Vicente las cantidades de riego extras no alcanzan niveles tan altos como en el caso del cantón Salitre. A continuación, se muestran los resultados:

Mes	**Período** $p = 10$ *días*	**Etapa**	**Kc**	ET_c mm/d	ET_c mm/p	**Prec, Efec** mm/p	**Req,Riego** mm/d
Dic	2	Inic	0,30	1,07	6,40	13,40	0,00
Dic	3	Inic	0,30	1,05	11,60	21,50	0,00
Ene	1	Des	0,36	1,26	12,60	20,00	0,00
Ene	2	Des	0,59	1,99	19,90	20,00	0,00
Ene	3	Des	0,83	2,80	30,80	19,40	1,14
Feb	1	Med	1,06	3,56	35,60	18,40	1,94
Feb	2	Med	1,10	3,70	37,00	17,70	1,71
Feb	3	Med	1,10	3,73	29,90	18,40	1,15
Mar	1	Med	1,10	3,76	37,60	19,50	1,82
Mar	2	Fin	1,10	3,78	37,80	20,10	1,77
Mar	3	Fin	0,93	3,15	34,70	19,90	1,48
Abr	1	Fin	0,66	2,23	22,30	19,30	0,30
Abr	2	Fin	0,44	1,45	11,60	15,30	0,00

TABLA 4.25. Requerimiento de riego para la productividad de maíz duro seco en el cantón San Vicente.

Los resultados indican que en las etapas media y final se satisfacen las condiciones. En la tabla (4.25) se muestra que, los requerimientos en el cantón San Vicente son bajos; por ejemplo, entre los días 1 y 10 correspondientes al primer período del mes de febrero se requiere de $1,94\ mm$ adicionales por día, siendo la mayor cantidad de agua que se necesita.

En base a estos resultados se describen las realidades de cada cantón referentes a la producción del maíz. Bajo este contexto, en la provincia de Santo Domingo durante el año 2020 se obtuvo una producción anual de $1.974\ Tm$ de maíz duro seco como se muestra en la tabla (1.1), siendo la provincia que menos produce en comparación al resto. El cantón Santo Domingo tiene registros bajos en productividad del maíz a pesar de producir dos tipos de maíz.

En la provincia de Los Ríos durante el 2020 se determinó $642.761\ Tm$ de maíz duro seco como se ve en la tabla (1.1). El cantón Valencia y principalmente en Palenque los niveles de productividad se encuentran en un nivel medio pese a encontrarse en la provincia con mayor producción de dos tipos diferentes de maíz.

En Esmeraldas la producción anual fue de $3.873\ Tm$ de maíz duro seco durante el 2020 como se evidencia en la tabla (1.1). En el cantón Eloy Alfaro los niveles de productividad pueden incrementar y ayudar en la contribución de producción de maíz a una de las provincias que menos aporta a nivel nacional.

En provincia del Guayas se determinó una producción $247.712\ Tm$ de maíz duro seco durante el 2020 como se indica en la tabla (1.1). Sin embargo, en el cantón Salitre se necesitan satisfacer los requerimientos hídricos más altos para incrementar aún más la productividad en esta provincia que es la tercera con mayor producción con tres distintos tipos de maíz.

En Manabí durante el 2020 la producción anual fue $280.757\ Tm$ de maíz duro seco como se observa en la tabla (1.1). El cantón San Vicente pertenece a la segunda provincia de mayor productividad con dos tipos de maíz.

En cada uno de estos cantones se necesitan sistemas de riego en ciertas etapas del cultivo que no se han implementado de forma óptima, por lo que se puede apreciar que la productividad en estas zonas mejoraría sustancialmente si se realiza una inversión para implementar proyectos adecuados. De manera que, al cultivar el maíz se obtengan las medidas idóneas de un correcto desarrollo de la plantación con el fin de que la productividad del maíz aumente considerablemente.

Por último, es importante mencionar que a nivel nacional la producción de maíz se genera por pequeños y medianos productores, quienes necesitan apoyo por parte del gobierno, con la finalidad de contar con proyectos en las zonas con alto potencial para el cultivo del maíz que beneficie la productividad.

CAPÍTULO 5

CONCLUSIONES Y RECOMENDACIONES

Este trabajo inicia detallando la producción del maíz en Ecuador, donde se observa que el maíz duro seco es el segundo cultivo transitorio con mayor producción por toneladas métricas, y cuyas superficies para la siembra y cosecha se encuentran en su mayoría en las provincias costeras del país. Para el caso de estudio se seleccionaron 25 cantones.

Se ha elegido la Precipitación como variable de respuesta escalar en base a los estudios desarrollados sobre el índice de potencialidad agrícola de Turc. De donde, se ha concluido que la Precipitación es la variable que se relaciona directamente los requerimientos hídricos necesarios para la productividad agrícola del maíz duro seco en el país.

Por otra parte, se ha escogido a la Temperatura y a la Velocidad del Viento como covariables explicativas funcionales. Estas variables meteorológicas se han elegido a partir del estudio, propuesto por Penman-Monteith en zonas andinas, que considera el balance hídrico mediante los índices de evapotranspiración de referencia ET_o y real ET_c; los cuales se calculan a través de factores meteorológicos como: temperatura, velocidad del viento, radiación solar, etc., por lo que, se concluye que las variables meteorológicas que mayor influencia tienen sobre el cálculo de estos índices son la Temperatura y la Velocidad del Viento. Estos índices permiten encontrar el requerimiento de agua que necesita una planta durante todo su proceso fisiológico, para alcanzar una productividad ideal.

El estudio de las variables orientado al *FDA* permite reducir la dimensionalidad de los datos recolectados, de forma que se ha construido una estructura funcional de la temperatura y de la velocidad del viento a partir del suavizamiento por bases funcionales de Fourier. A pesar de que no existe un método específico para obtener el número ideal de bases funcionales que se deben usar en dicho proceso, se concluye que, la técnica de validación cruzada solventa este problema. Por lo tanto, para

construir la mejor representación funcional de la Temperatura se ha tomado $\kappa_x = 12$ bases funcionales de Fourier de donde se obtuvo el mejor $R^2 ajustado$ que explica el $93,0\,\%$ de la variabilidad con la que se representa cada dato funcional y además tiene un GCV óptimo de $0,51$ donde se alcanza la estabilidad del suavizamiento. Mientras que, la estructura funcional de la Velocidad del Viento se ha obtenido con $\kappa_x = 11$ bases funcionales de Fourier que permiten explicar el $97,0\,\%$ del variabilidad con la que se representa cada dato funcional y la estabilidad del suavizamiento se da con un GCV óptimo de $0,22$.

Se recomienda realizar un análisis exploratorio de los datos funcionales ya que es necesario evidenciar la existencia de datos atípicos. Gracias al estudio de las medidas de centralidad y dispersión de los datos funcionales, se emplea el proceso bootstrap que proporciona las bandas de confianza para detectar datos atípicos en la muestra mediante las medidas de profundidad recortadas un $25\,\%$. Se hallaron dos curvas atípicas en la temperatura y una en la velocidad del viento.

En la aplicación de los modelos *FLR* se abordó la detección y tratamiento de curvas atípicas, en esta etapa se decidió eliminarlas del conjunto de observaciones. Luego de ello, se llevó a cabo un nuevo ajuste de cada modelo, procediendo a comparar los modelos en base a los valores del $R^2 ajustado$, la cantidad de parámetros significativos e índices estadísticos. Concluyendo que, los modelos en los cuales no se involucraron las curvas atípicas tienen una mejor descripción de la variabilidad de los datos.

En el caso de los modelos *FLR* que consideran una sola covariable explicativa, se recomienda usar aquellos modelos que consideran bases *FPLS* para la predicción. Específicamente, se debería seleccionar el modelo que considera a la Temperatura, pues presenta los mejores resultados en cuanto a la comparación de medidas de bondad de ajuste se refiere.

En general, se concluye que el mejor modelo *FLR* es aquel que considera dos covariables explicativas funcionales para obtener la predicción de la Precipitación. Además, este modelo fue validado mediante los contrastes de hipótesis que verifican el

cumplimiento de los supuestos de normalidad, independencia y homocedasticidad de los residuos, y el supuesto de especificación o linealidad del modelo a través de pruebas estadísticas. De la misma manera se verifica que el $MAPE = 8,31\,\%$ presenta el menor error porcentual absoluto medio; las medidas de bondad de ajuste AIC, AIC_C y BIC no son excesivamente mayores en comparación al resto, esto se debe a que se usa mayor información muestral (dos variables a la vez) para determinar la estimación. También se obtuvo un $97\,\%$ de la explicación de la variabilidad de los datos, siendo el modelo que cuenta con mayor cantidad de parámetros significativos (6).

Finalmente, en base a los resultados de la predicción de la Precipitación se ha encontrado un balance hídrico y los requerimientos de riego que se necesitan durante el ciclo de desarrollo fisiológico del maíz duro seco, en los cantones de Santo Domingo, Valencia, Palenque, Eloy Alfaro, Salitre y San Vicente. Se concluye que, entre los últimos días de enero y los primeros 20 días de marzo se precisa riego extra debido a que no se satisfacen los requerimientos de agua que proporciona la precipitación en las etapas de desarrollo y media del cultivo. Esto se debe a que, en las fases de desarrollo vegetal, prefloración y floración se necesita la cantidad óptima de agua para su adecuado proceso, por lo que, se recomienda implementar sistemas de riego que puedan suplir estas necesidades hídricas para mejorar la productividad de maíz en cada zona de estudio.

REFERENCIAS BIBLIOGRÁFICAS

Aguilera, A. M., Aguilera Morillo, M. C., & Preda, C. (2016). Penalized versions of functional pls regression. *Chemometrics and Intelligent Laboratory Systems*, 154:80–92.

Aparicio, J., Martínez, A., & Morales, J. (2004). *Modelos Lineales Aplicados en R*. PhD thesis, Universidad Miguel Hernández.

Apostol, C. & Pedra, C. (2010). PLS Methods for Functional Data. *REV. ROUMAINE MATH. PURES APPL*, 55(6):431–445.

BCE (2021). Información estadística mensual: Producto interno bruto por industria. *Banco Central del Ecuador*, 1(2029).

Cardot, H. & Sarda, P. (2006). Linear Regression Models for Functional Data. *Statistics and Probability Letters*, pages 49–66.

Carrillo Ramirez, A., Garatejo Escobar, O., & Pineda Ríos, W. (2017). Análisis multivariado de datos funcionales aplicado a curvas de encefalogramas. *Comunicaciones en Estadística*, 10(1):129–144.

Cuesta-Albertos, J. A. & Nieto-Reyes, A. (2008). The random tukey depth. *Computational Statistics and Data Analysis*, 52(11):4979–4988.

Cuevas, A., Febrero, M., & Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

De Boor, C. (1977). Package for Calculating with B-Splines. *SIAM Journal on Numerical Analysis*, 14(3):441–472.

ESPAC (2020). Encuesta de superficie y producción agropecuaria continua. Technical report, Instituto Nacional de Estadísticas y Censos.

Estrada, R. D. (2011). Ajustes al índice de potencialidad agrícola de turc para lograr mejores diseños de los mecanismos para compartir beneficios en los andes. Technical report, RIMISP.

Estrada, R. D., Burbano, J., Tapia, X., & Gavilanes, C. (2013). Identificación espacial del impacto del cambio climático en la productividad y competitividad de los pequeños productores. *Ministerio de Agricultura Ganadería, Acuacultura y Pesca (MAGAP)- GIZ*, 1:4–10.

FAO (1977). Evapotranspiracion. *Estudio FAO Riego y Drenaje*, 56:15–26.

Faraway, J. J. (2014). *Linear Models with R*. Chapman and Hall/CRC.

Febrero-Bande, M., Galeano, P., & González-Manteiga, W. (2010). Measures of influence for the functional linear model with scalar response. *Journal of Multivariate Analysis*, 101(2):327–339.

Febrero Bande, M. & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):3–20.

Fraiman, R. & Muniz, G. (2001). Trimmed means for functional data. *Sociedad de Estadistica e Investigacidn Operativa Test*, 10(2):419–440.

García de Pedraza, L. (1963). Los vientos en agricultura. Technical report, Ministerio de Agricultura de Madrid.

Haar, A. (1910). Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69(3):331–371.

Hastie, T. & Mallows, C. (1993). A Statistical View of Some Chemometrics Regression Tools: Discussion. *Technometrics*, 35(2):140–143.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Jiménez, S., Castro, L., Yépez, J., & Wittmer, C. (2012). Impacto del cambio climático en la agricultura de subsistencia en el ecuador. *Serie Avances de Investigación*, 66:15–22.

Márquez, J. (2021). Boletín técnico y presentación de la espac 2020. Technical report, Instituto Nacional de Estadísticas y Censos.

Martens, H. & Naes, T. (1991). Multivariate Calibration. *Biometrical Journal*, 33(4):418–418.

Marx, B. D. & Eilers, P. H. (1999). Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach. *Technometrics*, 41(1):1–13.

Oviedo de la Fuente, M. (2018). *Advances in Functional Regression and Classification Models*. PhD thesis, Universidade de Santiago de Compostela.

Preda, C. & Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 48(1):149–158.

Ramsay, J., Hooker, G., & Graves, S. (2009). Functional Linear Models for Scalar Responses. In *Functional Data Analysis with R and MATLAB*, pages 131–146. Springer, New York, NY.

Ramsay, J. O. & Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):539–572.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, volume 40. Springer, New York, NY.

Segura, M. & Andrade, L. (2011). *Efecto de las condiciones agrometeorológicas sobre un cultivar criollo y dos híbridos de maíz en cuatro fechas de siembra*. PhD thesis, Escuela Politécnica del Ejército.

Stone, M. & Brooks, R. J. (1990). Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(2):237–258.

Tenenhaus, M. (1998). *La régression PLS : Théorie et Pratique*. Éditions Technip.

ANEXOS

ANEXO 1

BASES κ_X DE FOURIER

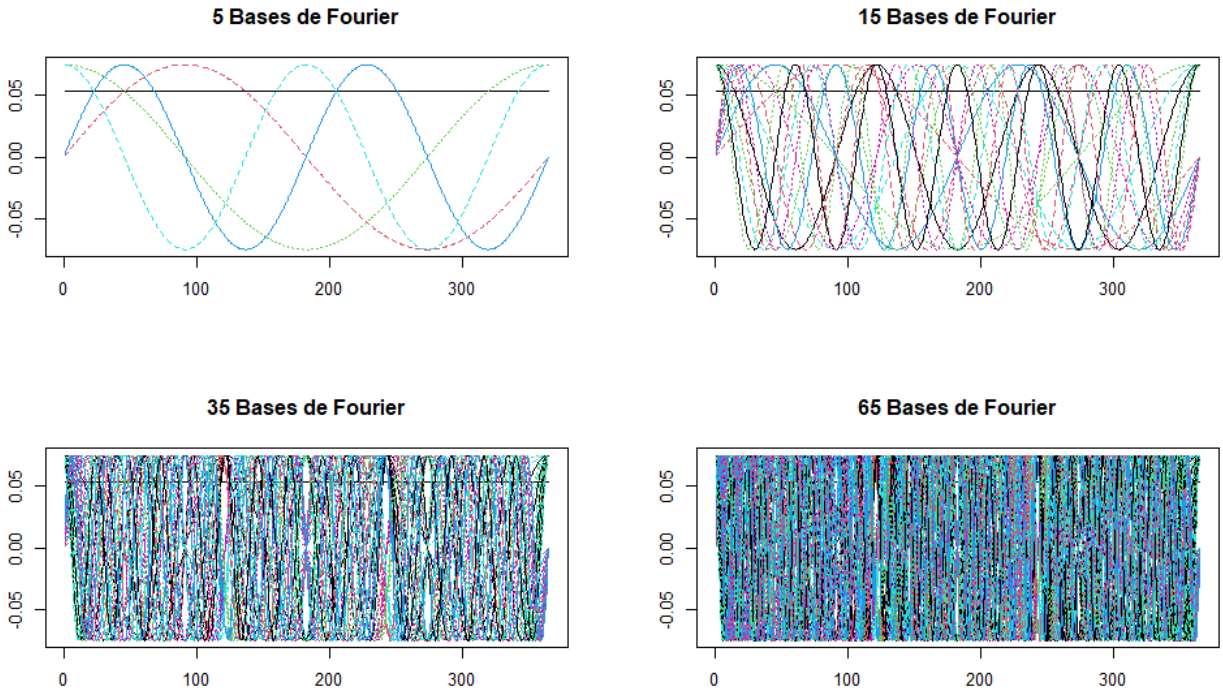

FIGURA 5.1. Bases Funcionales de Fourier para el suavizamiento de los datos discretos.

ANEXO 2

MODELO FLR CON REPRESENTACIÓN EN BASES PARA LA TEMPERATURA CON 23 CURVAS

```
*** Summary Functional Data Regression with representation in Basis ***
Call:
fregre.basis(fdataobj = datafd_x1_23, y = y1_23)
Residuals:
     Min       1Q   Median       3Q      Max
-0.91844 -0.24667 0.06822 0.35913 0.62143
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.59149    0.10811 23.971 1.52e-14 ***
datafd_x1_23.bspl4.1 -0.19513 0.05487 -3.556 0.00243 **
datafd_x1_23.bspl4.2 0.15043  0.03827  3.931  0.00108 **
datafd_x1_23.bspl4.3 -0.13689 0.02857 -4.792 0.00017 ***
datafd_x1_23.bspl4.4 0.02806  0.01925  1.458  0.16312
datafd_x1_23.bspl4.5 0.11448  0.04312  2.655  0.01666 *

Residual standard error: 0.5185 on 17 degrees of freedom
Multiple R-squared: 0.743,   Adjusted R-squared: 0.6674
F-statistic: 9.827 on 5 and 17 DF, p-value: 0.0001495

-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

ANEXO 3

MODELO FLR CON BASES FPC PARA LA TEMPERATURA CON 23 CURVAS

```
 *** Summary Functional Data Regression with Principal Components ***
Call:
fregre.pc(fdataobj = datafd_x1_23, y = y1_23, l = c(1:5))
Residuals:
    Min      1Q  Median      3Q     Max
-0.7818 -0.3000 0.0631 0.2863 0.5981
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.59149    0.09172 28.254 9.93e-16 ***
PC1          0.01637    0.01153  1.419 0.174030
PC2          0.06358    0.01170  5.432 4.48e-05 ***
PC3          0.15619    0.03431  4.553 0.000282 ***
PC4          0.03997    0.07137  0.560 0.582745
PC5         -0.62225    0.12568 -4.951 0.000121 ***

Residual standard error: 0.4399 on 17 degrees of freedom
Multiple R-squared: 0.815,   Adjusted R-squared: 0.7606
F-statistic: 14.98 on 5 and 17 DF, p-value: 1.03e-05

-With 5 Principal Components is explained 98.43 %
 of the variability of explicative variables.

-Variability for each principal components -PC- (%):
  PC1   PC2   PC3   PC4   PC5
90.76  4.74  0.55  2.33  0.05
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

ANEXO 4

MODELO FLR CON BASES FPLS PARA LA TEMPERATURA CON 23 CURVAS

```
*** Summary Functional Regression with Partial Least Squares***
-Call: fregre.pls(fdataobj = fdata(datafd_x1_23), y = y1_23, l = c(1:5))

             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 2.59148797 0.09524757 27.2079171 1.569012e-12
PLS1        0.05695725 0.02934672 1.9408389 7.508992e-02
PLS2        0.11509495 0.02865153 4.0170610 1.568600e-03
PLS3        0.32367784 0.07455391 4.3415274 8.671520e-04
PLS4        1.24722214 0.27413435 4.5496748 5.966204e-04
PLS5        0.09448459 0.17706406 0.5336181 6.029350e-01

-R squared: 0.8528071
-Residual variance: 0.2086583 on 12.54168 degrees of freedom
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

ANEXO 5

MODELO FLR CON REPRESENTACIÓN EN BASES PARA LA VELOCIDAD DEL VIENTO DE LAS 24 CURVAS

```
 *** Summary Functional Data Regression with representation in Basis ***
Call:
fregre.basis(fdataobj = datafd_x2_24, y = y1_24)
Residuals:
     Min       1Q   Median       3Q      Max
-0.85849 -0.19781 -0.02235 0.25240 0.65027

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.58587    0.07963 32.474 < 2e-16 ***
datafd_x2_24.bspl4.1 -0.28761 0.03983 -7.221 1.02e-06 ***
datafd_x2_24.bspl4.2 0.32332  0.03645  8.871 5.46e-08 ***
datafd_x2_24.bspl4.3 -0.38077 0.04001 -9.517 1.90e-08 ***
datafd_x2_24.bspl4.4 0.26866  0.03658  7.344 8.11e-07 ***
datafd_x2_24.bspl4.5 -0.13267 0.03348 -3.963 0.000912 ***

Residual standard error: 0.3901 on 18 degrees of freedom
Multiple R-squared: 0.8461,  Adjusted R-squared: 0.8033
F-statistic: 19.79 on 5 and 18 DF, p-value: 9.568e-07

-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

ANEXO 6

MODELO FLR CON BASES FPC PARA LA VELOCIDAD DEL VIENTO CON 24 CURVAS

```
*** Summary Functional Data Regression with Principal Components ***
Call:
fregre.pc(fdataobj = datafd_x2_24, y = y1_24, l = c(1:5))
Residuals:
     Min       1Q   Median       3Q      Max
-1.12581 -0.21697  0.00063  0.30072  0.91029
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.58587    0.10638 24.309 3.24e-15 ***
PC1         0.03161    0.00911  3.469  0.00274 **
PC2         0.05943    0.03839  1.548  0.13904
PC3         0.01512    0.08443  0.179  0.85987
PC4         0.71493    0.12146  5.886 1.42e-05 ***
PC5         0.34917    0.14692  2.377  0.02878 *

Residual standard error: 0.5211 on 18 degrees of freedom
Multiple R-squared: 0.7253,  Adjusted R-squared: 0.649
F-statistic: 9.506 on 5 and 18 DF, p-value: 0.0001435

-With 5 Principal Components is explained 96.86 %
 of the variability of explicative variables.

-Variability for each principal components -PC- (%):
  PC1   PC2   PC3   PC4   PC5
92.52  3.28  0.53  0.34  0.20
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

ANEXO 7

MODELO FLR CON BASES FPLS PARA LA VELOCIDAD DEL VIENTO CON 24 CURVAS

```
*** Summary Functional Regression with Partial Least Squares***

-Call: fregre.pls(fdataobj = fdata(datafd_x2_24), y = y1_24, l = c(1:5))

             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 2.58587323 0.08189938 31.5737817 5.493135e-16
PLS1        0.18127373 0.04551164 3.9830189 1.045554e-03
PLS2        0.01743354 0.04463378 0.3905907 7.011860e-01
PLS3        0.83936535 0.10093001 8.3163111 3.042269e-07
PLS4       -0.27304324 0.11392861 -2.3966168 2.894129e-02
PLS5        0.88524120 0.29703023 2.9803068 8.739525e-03

-R squared: 0.8533938
-Residual variance: 0.1609802 on 16.20725 degrees of freedom
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

ANEXO 8
MODELO FLR CON REPRESENTACIÓN EN BASES PARA LA TEMPERATURA Y LA VELOCIDAD DEL VIENTO CON 23 CURVAS

```
Call:
lm(formula = pf, data = XX, x = TRUE, control = ..1)
Residuals:
     Min       1Q   Median       3Q      Max
-0.17241 -0.05972 -0.01078 0.06475 0.28731

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.5914880 0.0301559 85.936 < 2e-16 ***
fdTemp.bspl4.1 -0.0168741 0.0206170 -0.818 0.42905
fdTemp.bspl4.2  0.0081211 0.0292109  0.278 0.78573
fdTemp.bspl4.3 -0.0204468 0.0465994 -0.439 0.66862
fdTemp.bspl4.4  0.0009009 0.0462488  0.019 0.98478
fdTemp.bspl4.5  0.0353754 0.0488884  2.724 0.04318 *
fdViento.bspl4.1 -0.2360398 0.0847826 -2.784 0.01653 *
fdViento.bspl4.2 0.2724088 0.0820781 3.319 0.00612 **
fdViento.bspl4.3 -0.3041279 0.0810516 -3.752 0.00276 **
fdViento.bspl4.4 0.2245548 0.0675803 3.323 0.00608 **
fdViento.bspl4.5 -0.1347893 0.0544854 -2.474 0.02929 *

Residual standard error: 0.1446 on 12 degrees of freedom
Multiple R-squared: 0.9859,  Adjusted R-squared: 0.9741
F-statistic: 83.8 on 10 and 12 DF, p-value: 1.583e-09
```

ANEXO 9

RESULTADOS DE CROPWAT PARA EL CÁLCULO DE LA ET_O

Mes	Temp Min $^\circ C$	Temp Max $^\circ C$	Humedad $\%$	Viento m/s	Insolación $horas$	Radiación $MJm^{-2}d^{-1}$	ET_o mm/d
Enero	21,0	28,8	85,0	3,1	4,6	16,0	3,6
Febrero	21,2	28,0	87,0	2,7	4,6	16,5	3,5
Marzo	21,4	28,1	87,0	2,6	4,1	15,9	3,4
Abril	21,3	28,4	87,0	2,7	4,1	15,4	3,3
Mayo	21,1	28,3	87,0	2,8	4,2	14,7	3,2
Junio	20,5	28,5	85,0	3,0	4,2	14,2	3,2
Julio	20,1	29,7	80,0	3,1	4,3	14,5	3,6
Agosto	20,0	32,0	72,0	3,3	4,3	15,3	4,5
Septiembre	20,4	32,9	69,0	3,5	4,4	16,1	5,0
Octubre	20,4	32,1	70,0	3,4	4,7	16,6	4,9
Noviembre	20,2	31,6	71,0	3,1	5,0	16,7	4,7
Diciembre	20,7	30,8	76,0	3,3	5,0	16,4	4,3
Promedio	20,7	29,9	80,0	3,0	4,5	15,7	3,9

TABLA 5.1. Evapotranspiración calculada en el cantón Santo Domingo por mes.

Mes	Temp Min $^\circ C$	Temp Max $^\circ C$	Humedad $\%$	Viento m/s	Insolación $horas$	Radiación $MJm^{-2}d^{-1}$	ET_o mm/d
Enero	19,1	27,1	82,0	3,1	8,3	21,7	4,2
Febrero	19,4	25,9	87,0	2,9	7,3	20,8	3,8
Marzo	19,5	25,9	88,0	2,8	7,2	20,8	3,7
Abril	19,4	26,5	87,0	2,6	6,3	18,7	3,5
Mayo	19,1	26,8	85,0	2,5	7,7	19,7	3,7
Junio	18,2	27,6	81,0	2,2	6,0	16,6	3,4
Julio	17,7	28,6	75,0	2,3	8,0	19,7	4,1
Agosto	17,8	30,2	67,0	2,5	6,7	18,8	4,6
Septiembre	18,3	30,6	65,0	2,9	7,5	20,9	5,3
Octubre	18,5	29,3	68,0	3,0	6,6	19,6	4,9
Noviembre	18,4	28,5	69,0	2,9	7,7	20,9	4,8
Diciembre	18,8	28,4	73,0	3,1	8,0	21,0	4,7
Promedio	18,7	27,9	77,0	2,7	7,3	19,9	4,2

TABLA 5.2. Evapotranspiración calculada en el cantón Valencia por mes.

Mes	Temp Min $^{\circ}C$	Temp Max $^{\circ}C$	Humedad $\%$	Viento m/s	Insolación $horas$	Radiación $MJm^{-2}d^{-1}$	ET_o mm/d
Enero	22,8	34,3	69,0	1,3	2,6	13,1	3,7
Febrero	22,8	32,0	79,0	1,4	2,9	14,0	3,5
Marzo	22,7	31,4	81,0	1,4	2,7	13,7	3,3
Abril	22,6	32,4	78,0	1,3	3,3	14,1	3,4
Mayo	22,2	33,6	73,0	1,2	2,4	12,0	3,2
Junio	21,7	35,0	66,0	1,2	1,2	9,9	3,1
Julio	21,5	36,5	59,0	1,3	0,8	9,5	3,4
Agosto	21,5	37,7	55,0	1,3	1,0	10,3	3,7
Septiembre	21,7	37,8	55,0	1,2	1,2	11,1	3,8
Octubre	21,8	36,9	56,0	1,3	0,9	10,8	3,8
Noviembre	21,8	36,4	56,0	1,3	0,8	10,4	3,7
Diciembre	22,7	36,1	59,0	1,3	1,3	11,0	3,7
Promedio	22,2	35,0	65,5	1,3	1,8	11,7	3,5

TABLA 5.3. Evapotranspiración calculada en el cantón Palenque por mes.

Mes	Temp Min $^{\circ}C$	Temp Max $^{\circ}C$	Humedad $\%$	Viento m/s	Insolación $horas$	Radiación $MJm^{-2}d^{-1}$	ET_o mm/d
Enero	23,7	29,1	83,0	3,4	3,8	14,6	3,5
Febrero	23,9	29,3	83,0	2,9	4,2	15,8	3,7
Marzo	24,0	29,4	83,0	2,7	4,3	16,2	3,7
Abril	24,2	29,5	84,0	3,0	4,7	16,4	3,8
Mayo	24,1	29,3	84,0	3,5	3,5	13,9	3,4
Junio	23,7	29,5	83,0	3,8	3,1	12,8	3,4
Julio	23,6	30,1	79,0	4,0	3,2	13,1	3,8
Agosto	23,7	30,8	76,0	4,2	3,2	13,8	4,2
Septiembre	23,7	30,8	76,0	4,3	3,2	14,3	4,3
Octubre	23,5	30,3	77,0	4,0	3,4	14,5	4,2
Noviembre	23,3	29,8	79,0	3,6	3,1	13,6	3,8
Diciembre	23,5	29,4	82,0	3,6	3,1	13,4	3,5
Promedio	23,7	29,8	80,8	3,6	3,6	14,4	3,8

TABLA 5.4. Evapotranspiración calculada en el cantón Eloy Alfaro por mes.

Mes	Temp Min $^{\circ}C$	Temp Max $^{\circ}C$	Humedad $\%$	Viento m/s	Insolación $horas$	Radiación $MJm^{-2}d^{-1}$	ET_o mm/d
Enero	23,0	34,5	68,0	3,6	3,2	14,1	5,1
Febrero	23,2	32,2	78,0	3,0	3,4	14,8	4,1
Marzo	23,1	31,4	81,0	2,7	3,7	15,3	3,9
Abril	22,9	32,6	77,0	2,8	4,3	15,6	4,2
Mayo	22,4	34,0	72,0	3,2	3,5	13,5	4,4
Junio	21,9	35,4	64,0	3,5	2,7	11,9	5,0
Julio	21,5	36,6	59,0	3,9	2,9	12,4	5,8
Agosto	21,1	37,2	57,0	4,2	3,3	13,7	6,5
Septiembre	21,3	37,2	57,0	4,4	3,4	14,5	6,8
Octubre	21,5	36,4	59,0	4,4	2,5	13,3	6,3
Noviembre	21,7	36,0	59,0	4,4	3,1	14,0	6,3
Diciembre	22,7	35,9	60,0	4,2	3,1	13,8	6,1
Promedio	22,2	35,0	65,9	3,7	3,3	13,9	5,4

TABLA 5.5. Evapotranspiración calculada en el cantón Salitre por mes.

Mes	Temp Min $^{\circ}C$	Temp Max $^{\circ}C$	Humedad $\%$	Viento m/s	Insolación $horas$	Radiación $MJm^{-2}d^{-1}$	ET_o mm/d
Enero	24,0	27,1	83,0	4,5	3,0	13,6	3,4
Febrero	24,5	27,1	85,0	3,5	3,5	14,8	3,4
Marzo	24,5	27,1	85,0	3,2	3,8	15,4	3,4
Abril	24,3	27,2	85,0	3,6	3,5	14,5	3,3
Mayo	24,0	27,0	84,0	4,6	2,6	12,4	3,2
Junio	23,2	26,5	83,0	5,0	2,2	11,4	3,1
Julio	22,7	26,2	82,0	5,2	3,8	13,8	3,4
Agosto	22,3	26,1	81,0	5,3	2,2	12,2	3,4
Septiembre	22,2	26,1	81,0	5,4	2,3	12,9	3,5
Octubre	22,6	26,4	80,0	5,6	2,3	12,9	3,6
Noviembre	22,6	26,4	80,0	5,5	2,8	13,3	3,6
Diciembre	23,3	26,9	81,0	5,2	3,1	13,5	3,6
Promedio	23,4	26,7	82,5	4,7	2,9	13,4	3,4

TABLA 5.6. Evapotranspiración calculada en el cantón San Vicente por mes.