ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELOS ESTADÍSTICOS PARA LA DETECCIÓN DE PATRONES EN MEDIO AMBIENTE Y ECONOMÍA

APLICACIÓN DE GRÁFICOS DE CONTROL FUNCIONAL PARA MONITORIZAR SERIES TEMPORALES METEOROLÓGICAS Y CLIMÁTICAS PARA LA DETECCIÓN DE ANOMALÍAS

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

JULIÁN ENRIQUE RODRÍGUEZ TOCAGÓN

julian.rodriguez@epn.edu.ec

DIRECTOR: PH. D. MIGUEL ALFONSO FLORES SÁNCHEZ

miguel.flores@epn.edu.ec

DMQ, FEBRERO 2022

CERTIFICACIONES

Yo, JULIÁN ENRIQUE RODRÍGUEZ TOCAGÓN, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

_____

Julián Enrique Rodríguez Tocagón

Certifico que el presente trabajo de integración curricular fue desarrollado por Julián Enrique Rodríguez Tocagón, bajo mi supervisión.

_____

Ph. D. Miguel Alfonso Flores Sánchez

DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

Julián Enrique Rodríguez Tocagón

Ph. D. Miguel Alfonso Flores Sánchez

AGRADECIMIENTOS

Agradezco a mi hermano, a mis hermanas, a mi madre y padre quienes desde un inicio me brindaron su apoyo constante en el trascurso de esta etapa motivándome, dado impulso y nunca dejándome caer día tras día. Gracias a ustedes pude dar un gran paso en mis metas.

A Ruth quién estuvo a mi lado a pesar de todo, con su apoyo y ayuda incondicional siempre pendiente de mí. Además, de darme un hogar y motivo para esforzarme más.

A mi tutor por ayudarme a conseguir un proyecto de titulación, pese a las dificultades que se presentaron, también le agradezco por su tiempo y dedicación al guiarme en el desarrollo de este proyecto.

A mis compañeros, que me brindaron su amistad y apoyo en los bueno y malos momento universitarios.

DEDICATORIA

A mi familia y hogar.

RESUMEN

Una de las herramientas del análisis y detección de problemas en el control estadístico es el gráfico de control, empleado para recopilar y analizar datos con el objetivo de monitorear, controlar e identificar las variaciones en el proceso que se deben a causas comunes y especiales.
Se propone la fase I del grafico de control mediante el uso de profundidades que se extenderá hacia el análisis de datos funcionales para monitorear las curvas suaves, llamando a este método como gráficos de control funcionales, en donde se considera la profundidad modal y procedimientos bootstrap para determinar el cuantil y realizar el grafico de control funcional de rangos para detectar funciones atípicas.
El método se aplicó a los datos meteorológicos de la temperatura ambiente de 7 estaciones de la provincia Chimborazo monitorizando y comparando el clima en el tiempo. Detectando que los datos atípicos son parte de una o varias estaciones producidos por fenómenos naturales presentados en la sección de resultados.

Palabras clave: Análisis Funcional, Gráficos de control, Profundidad Funcional, Series Meteorológicas , Atípicos, Bases de Fourier.

ABSTRACT

One of the tools of analysis and problem detection in statistical control is the control chart, used to collect and analyze data in order to monitor, control and identify variations in the process that are due to common and special causes.

Phase I of the control chart is proposed using depths that will be extended to the analysis of functional data to monitor smooth curves, calling this method as functional control charts, where the modal depth and bootstrap procedures are considered to determine the quantile and perform the functional control chart of ranges to detect outlier functions.

The method was applied to meteorological data of ambient temperature from 7 stations in the Chimborazo province monitoring and comparing the climate over time. Detecting that the atypical data are part of one or several stations produced by natural phenomena presented in the results section.

Keywords: Functional Analysis, Control Charts, Functional Depth, Meteorological Series, Outliers, Fourier basis.

Índice general

Índice de figuras

Capítulo 1

Descripción del componente desarrollado

Se propone aplicar gráficos de control a datos funcionales, realizando un monitoreo de series temporales meteorológicas y climáticas para detectar anomalías, la información será proporcionada por el Grupo de Energías Alternativas y Ambiente (GEAA), situada en la provincia de Chimborazo contando con un total de 14 estaciones climáticas. En el presente proyecto se hará uso únicamente de 7 estaciones tomado la información de la temperatura ambiente con el propósito de encontrar una variación individual y general, lo que nos lléva a un análisis univariante funcional respecto a cada estación y multivariante respecto a la agrupación de las estaciones respectivamente. Así, se podrá detectar si los cambios bruscos de temperatura se deben a un fenómeno natural climático de la región, localidad o se trata de un fallo del sensor de medición de una estación predeterminada. Debido a estas situaciones la información meteorológica tratada tendrá muchos comportamientos anómalos conocidos como valores atípicos, outliers, o simplemente anomalías, cuya definición intuitiva podría ser una observación que se desvía de otras despertando las sospechas que fue generado por diversas causas.
Para responder estas preguntas se implementará diversos conceptos y definiciones en el software estadístico R que se usarán para el desarrollar el proyecto, como la transformación de la información discreta a datos funcionales mediante el uso de bases de Fourier tomando en cuenta su rugosidad. Para esto se aplicará el concepto de validación cruzada gene-

ralizada usando el paquete *fda.usc* que nos indicara el número de bases optimo y su penalización rugosidad para tener un buen ajuste en las funciones indicando el comportamiento intrínseco de los datos. Seguido de la fase I del gráfico de control funcional, se realiza el cálculo de la profundidad funcional y el Bootstrap suavizado para la detección de datos atípicos por estación usando el paquete *qcr*, para el caso multivariante calculamos la profundidad multivariante seguido de un remuestreo y la fase 1 del gráfico control de las profundidades.

1.1. Objetivo general

Aplicar el cálculo de profundidad de datos funcionales a las series meteorológicas para monitorizar continuamente la detección de anomalías mediante gráficos de control.

1.2. Objetivos específicos

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

2. Aplicar a las series de datos funcionales el cálculo de profundidad funcional para detectar funciones atípicas.

3. Monitorizar mediante los gráficos de control la series de datos funcionales proporcionadas por el Grupo de Energías Alternativas y Ambiente GEAA. Para la detección de anomalías y poder detectar posibles fallas en los equipos, pronosticar futuras catástrofes naturales, o patrones del cambio climático.

1.3. Alcance

La estadística multivariante se usa frecuentemente para analizar este tipo de información logrando conseguir un análisis mas robusto y predicciones más precisas. Sin embargo, el método del Análisis de Datos

Funcionales (FDA) está empezando a tener buen desempeño en el análisis de grandes volúmenes de datos, obteniéndose resultados de manera más óptima, transformando los datos discretos mediante la suavización del comportamiento intrínseco del grupo de datos, calculando una base óptima y su penalización por rugosidad, obteniendo una función o curva caracterizada por la evolución de la variable en el transcurso del tiempo (*proceso estocástico*) en general, estas funciones toman uno o varios argumentos a diferencia del análisis multivariante clásico que toma vectores. Así, esta técnica es más eficiente que el enfoque estadístico multivariante ya que se puede estudiar sus características de mejor modo.
Además, para la detección de datos funcionales atípicos que influyan en la muestra, se usara la metodología de profundidad funcional modal y un algoritmo de detección bootstrap para el remuestreo y dectección de profundidades con valores bajos utilizando el método de *Bootstrap suavizado* mediante la covarianza de los datos. Finalmente obteniendo la información del cuantil para realizar la fase I del grafico de control de rangos funcionales y monitorizar la información para detectar anomalías y para el caso multivariante calculamos la profundidad de las estaciones mediante el uso de la distancia de *Mahalanobis*, seguido de un remuestreo y finalizando con la fase I del grafico de control, como este caso es el general se comparará las anomalías con los del caso funcional.
El proceso se lo indicará paso a paso con todos los conceptos a utilizar e implementar hasta cumplir nuestro objetivo.

1.4. Marco teórico

En esta sección se indicará diferentes nociones, conceptos y algunas definiciones teóricas para lograr tener un mejor entendimiento de la metodología empleada para realizar la aplicación de gráficos de control funcional y asi lograr monitorizar series temporales meteorológicas y climáticas para la detección de anomalías. Para esto se presentan el análisis de datos funcionales, profundidad funcional, algoritmo de detección, los gráficos de control y los gráficos de control funcional.

1.4.1. Análisis de Datos Funcionales

Introducción

En esta sección abordaremos *el análisis de datos funcionales* con sus siglas en inglés (FDA) la cual es parte de una rama de la estadística. Cuyo objetivo es de estudiar y analizar la información presentada en funciones, curvas o elementos que varíen sobre un espacio continuo. Estas técnicas y modelos empleados por el análisis de datos funcionales, tiene similitud con los modelos que ocupan el análisis de datos multivariados, complementándose con diferentes técnicas [13].
Los objetivos del análisis de datos funcionales en esencia son la obtención de una parte representativa de los datos de modo que ayude a entender y resaltar las distintas características, posibles patrones de variabilidad en los datos, la explicación del porqué de la variación de un resultado o la más común comparaciones de datos respecto a ciertas variables. Entre otras técnicas, las cuales se realizarán mediante **Software estadístico R**

Datos Funcionales

El análisis de datos funcionales, a pesar de ser un tema relativamente nuevo, tiene un gran número de publicaciones uno de los más referentes son los libros de [13] [5], en los cuales explican los problemas básicos de la estadística funcional. Cabe recalcar que la unidad básica para este tipo de análisis es el dato funcional o variable funcional. Ahora extendiéndonos al contexto multivariado, tenemos que los datos se los puede observar en tiempos distintos de un rango $\mathcal{T} = [t_1, t_m]$ con $m > 0$, de tal modo que se puede expresar en una familia aleatoria $X(t_j)_{j=1,\dots,m}$. En el (FDA) si tenemos instantes seguidos en un rango, estos por los general son pequeños, entonces se asume que las muestras son observaciones de una familia continua [5].

$$\mathcal{X} = \{X(t) : t \in T\}$$

Definiciones:

En el trabajo de *ferraty y Vieu* [5], se define que una variable aleatoria funcional es como una variable aleatoria que toma valores en un espacio de funciones.

Definición 1: Sea $L^2(T)$,un espacio de Hilbert separable, la cual está definido por la función del cuadrado integrable en el siguiente intervalo $\mathcal{T} = [a, b] \subset \mathbb{R}$

$$L^2(T) = \{X : \mathcal{T} \to \mathbb{R}; \int_{\mathcal{T}} X(t)^2 dt < \infty\}$$

Y su producto interno se lo define de la siguiente manera:

$$\langle X(t), Y(t) \rangle = \int_{\mathcal{T}} X(t)Y(t)dt$$

Definición 2: Una variable aleatoria $\mathcal{X}$ se llama variable funcional (v.f), si toma valores en un espacio infinito dimensional (Espacio Funcional). Un dato funcional es una observación X $\in \mathcal{X}$

Definición 3: Si tenemos n variables funcionales $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_n$ idénticamente distribuidas, a ellos se los conoce como un conjunto de datos funcionales $\{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_n\}$.
Además, en la práctica se supone que los valores observados en un intervalo finito $\mathcal{T}$, son discretizaciones en instantes de tiempo $t_1, t_2, ..., t_m \in \mathcal{T}$ con $m > 0$. Así, el primer paso es la adaptación de la estructura funcional del conjunto mencionado [14].

Bases:

La base es un grupo de funciones conocidas $\phi_{k k \in N}$, con las cuales diferentes funciones pueden aproximarse tanto como se desee, haciendo uso de K (funciones base) de la siguiente manera:

$$X_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) = \mathbf{c}_i^T \phi(t)$$

Donde los coeficientes de la combinación lineal es $\mathbf{c}_i$ y Φ es la matriz de los valores de las K funciones base $\phi_k(t)$.

Esta elección de base y K es de suma importancia para la suavización de las funciones. Sin embargo, no existe una ley que indique un valor óptimo, ya que va a estar en función de la información que se esté tratando ya que se puede tomar la base de b-spline o la de Fourier en el presente proyecto se utilizará la *base de Fourier* debido a que contamos con datos periódicos.

Base de Fourier

En algunos casos existe repeticiones en las funciones durante un periodo τ de oscilación, entonces para mantener esto se utiliza el seno y coseno de frecuencia creciente

$$1, sen(\omega t), cos(\omega t), sen(2\omega t), cos(2\omega t), ...$$

donde $\omega = \frac{2/pi}{\tau}$.
Si el conjunto de datos son equiespaciados y se le asigna un periódo se utiliza la base de función ortonormal. Además, a partir del primer valor de $\phi_0(t) = 1$ todo se representa en seno y coseno con la respectiva multiplicación del argumento, siendo esta base la más utilizada en el caso periódico.

Ajuste a funciones suaves

Para realizar un ajuste al dato funcional, debemos considerar la existencia de una función suave que sea una representación de los datos. La técnica con más acogida para representar estos datos consiste en expresar los datos mediante expansiones o combinaciones lineales de **funciones base**, pero asumiendo la existencia de ruido en los datos suponemos que se observa a un dato funcional mediante este modelo:

$$\begin{aligned} x_{ij} &= X_i(t_j) + \epsilon_{ij} && i \in 1, ..., n, j \in 1, ..., m \\ \mathbf{x}_i &= X_i(\mathbf{t}) + \epsilon_i && (*) \end{aligned}$$

Donde (*) es su representación vectorial, ϵ_i son residuos independientes. Entonces nuestro objetivo en este caso es tratar de reducir el ruido o error para tener una mejor estimación de las funciones X_i, mediante el

uso de un conjunto de $\{\phi_k\}_{k\in N}$ funciones bases construyendo vectores de coeficientes aproximando mediante expanciones de los ϕ_k. Además, notemos que en la aplicacion el valor de $K \in \mathbb{N}$ es finito, con el cual se aproxima X_i

$$X_i(y) = \sum_{k=1}^{K} c_{ik}\phi(t)$$
$$X_i(y) = \mathbf{c}_i^T \phi(t)$$

con c_i coeficientes en $\mathbb{R}^K$ de $X_i(t)$ y el vector ϕ de las K bases.
Ahora seguimos con el método de los mínimos cuadrados, el modelo que se utilizara es el siguiente:

$$SSE_i = \sum_{j}^{m} [x_{ij} - X_i(t_j)]^2$$

$$SSE_i = \sum_{j}^{m} [x_{ij} - \mathbf{c}_i^T \phi(t_j)]^2 \tag{1.1}$$

donde x_i es el vector de los valores que serán ajustados y ϕ una matriz de tamaño m x k compuestas de las K bases $\phi_k(t_j)$ y el vector ϵ tiene un orden de m x 1, y remplazando en (1.1) se la expresa de la siguiente manera:

$$SSE_i = (x_i - \Phi\mathbf{c}_i)^T (x_i - \Phi\mathbf{c}_i)$$
$$SSE_i = ||x_i - \Phi\mathbf{c}_i||^2$$

calculando c se obtiene un estimador $\hat{c}$ el cual tiene como función minimizar la suma de cuadrados.

$$\hat{c}_i = (\Phi^t\Phi)^{-1}\Phi^t X_i$$

De esta manera el vector de los datos ajustados $\hat{X}(t_j)$ esta representado por:

$$\hat{X}_i = \Phi(\Phi^t\Phi)^{-1}\Phi^t X_i$$
$$\hat{X}_i = \Phi\hat{c}_i$$

Así, conseguimos que las funciones cuenten con sus derivadas y se verifica que son suaves. Sin embargo, no queremos que el ajuste sea demasiado exacto, ya que de ser el caso existiría una rugosidad muy excesiva, para que no ocurran estos casos se propone la penalización por rugosidad la cual se mide mediante la integral el cuadrado de la segunda derivada como en la siguente ecuación:

$$PEN_2(X) = \int [D^2X(t)]^2 dt \tag{1.2}$$

Donde $[D^2x(t)]^2$, en el tiempo t, se lo conoce como la curvatura de las funciones en el tiempo (t). Sin embargo, en las aplicaciones de (1.2) no puede ser la correcta, debido a que únicamente controla la curvatura de las funciones originales. (*Ramsey y Silverman*)[14] proponen una técnica de suavización mediante un factor que penaliza al realizar un mal suavización, donde c es la minimización de la siguiente expresión.

$$PENSEE_\lambda(x|X) = \sum_j [x_{ij} - X_j(t_j)]^2 + \lambda \int [D^2x(t)]^2 dt$$

λ es el parámetro que indica la penalización de la suavización en un ajuste, medida en la SEE_i y el D^m que viene a ser la curvatura de $x(t)$. Entonces si se remplaza $X(t)$ por Φc_i en la expresión anterior tenemos:

$$PENSEE_\lambda(x_i|c_i) = (x_i - \Phi\mathbf{c}_i)^T(x_i - \Phi\mathbf{c}_i) + \lambda x PEN_2(x) \tag{1.3}$$

Ahora tomamos $PEN_2(x)$ y lo representamos de forma matricial.

$$\begin{aligned} PEN_2(X) &= \int [D^2X(t)]^2 dt \\ &= \int [D^2\mathbf{c}^T\phi(t)]^2 dt \\ &= \int [\mathbf{c}^T D^2\phi(t)][D^2\phi(t)^T\mathbf{c}] dt \\ &= \mathbf{c}^T R\mathbf{c} \end{aligned}$$

con:

$$R = \int D^2\Phi(t)D^2\Phi^T(t)dt$$

Entonces sustituimos en la ecuación (1.3) obteniendo lo siguiente:

$$PENSEE_\lambda(x_i|c_i) = (x_i - \Phi\mathbf{c}_i)^T(x_i - \Phi\mathbf{c}_i) + \lambda x\mathbf{c}^T R\mathbf{c} \tag{1.4}$$

Calculamos la derivada de (1.4) de c e igualamos a $cero$:

$$-2\Phi^T\mathbf{x}_i + 2\Phi^T\Phi\mathbf{c}_i) + 2\lambda R\mathbf{c} = 0$$

De esta manera conseguimos el estimador de c.

$$\hat{c}_i = (\Phi^T\Phi + \lambda R)^{-1}\Phi^T\mathbf{x}_i)$$

Además, notamos los datos que se se ajusta de $\hat{X}_i$:

$$\hat{X}_i = \Phi(\Phi^T\Phi + \lambda R)^{-1}\Phi^T\mathbf{x}_i)$$

Entonces la matriz estará dada por $G = \Phi(\Phi^T\Phi + \lambda R)^{-1}\Phi^T$, por tanto:

$$\hat{X}_i = G\mathbf{x}_i$$

Para calcular el valor indicado del parámetro de la suavización λ, se utiliza la técnica de validación cruzada generalizada con sus siglas en ingles (GCV), esta técnica consiste en determinar el λ óptimo para minimizar la expresión.

Estadística descriptiva para los FDA

De *(Ramsay, J.O y Sylverman, B.W)* [14], obtenemos las siguientes definiciones las cuales son funciones muéstrales descriptiva y se calculan de un conjunto de datos funcionales $S_n = \{X_1, ..., X_n\}$ definimos en $t \in \mathcal{T} \subset \mathbb{R}$.

- La media funcional muestral se define de la siguiente manera:

$$\hat{\mathcal{X}}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t)$$

- La varianza funcional muestral se calcula como:

$$s^2_{\mathcal{X}}(t) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i(t) - \hat{\mathcal{X}}(t))^2$$

- La desviación estándar se calcula como:

$$s_{\mathcal{X}}(t) = \sqrt{s^2_{\mathcal{X}}(t))}$$

En este caso la media funcional poblacional, es la minimización de las suma de la distancia de todos los puntos de un trayecto, debido a esto la media muestral funcional es un estimador del centro en la distribución funcional [4]. Por otra parte tenemos otra definición en la cual utilizamos una base $\{\phi_k\}_{k\in\mathbb{N}}$, y d la cual es una función de distancia que pertenece a L^2. Así, se tiene los siguientes estimadores funcionales.

- La Media:

$$\underset{a\in S_n}{\operatorname{argmax}} \sum_{i=1}^{n} d(X_i.a)^2$$

- La Mediana:

$$\underset{a\in S_n}{\operatorname{argmax}} \sum_{i=1}^{n} d(X_i.a)$$

- La Varianza:

$$\frac{1}{n}\sum_{i=1}^{n} d(X_i.a)^2$$

Además tenemos mas características de datos que pertenecen al análisis exploratorio los cuales son la forma de la suavidad y rugosidad. Sin

embargo, para este caso se debe utilizar la medida de proximidad, esta es una semi-metrica y es la mejor manera de abordar un caso si tenemos un espacio infinito [5]. Una alternativa es la consideración de la distancia que existe entre las derivadas, para medir cual es la proximidad entre las curvas. Siendo $\{X_i(t)\}_{i=1}^n$ realizaciones independientes e idénticamente distribuidas de la variable aleatoria funcional $X(t)$ y X_i y $X_i^{'}$, entonces se parametriza a estas semi-metricas de la siguiente manera:

$$d_q^{'}(X_i, X_{i'}) = \sqrt{\int (X_i^{(q)}(t) - X_{i'}^{(q)}(t))^2 dt}$$

Donde, $X_i^{(q)}$ es la q-esíma derivada de X.

Medidas de Profundidad

El concepto de las medidas de profundidad indica un forma de orden de los puntos en el espacio Euclidiano desde el centro hacia afuera de los datos tomados de la muestra, de tal forma que los puntos que se encuentran más cerca al centro, se dice que tiene mayor profundidad a diferencia de los puntos que se encuentran en los extremos los cuales tendrían una menor profundidad, posiblemente consideradas como outliers en los datos. Esta definición de profundidad llevada al caso funcional indica de manera similar, medir la centralidad de una curva X_i en las curvas $X_1, \ldots, X_n$, donde se generaron del proceso estocástico S, donde toman valores en el espacio de funciones continuas cuadrado integrables $L^2([a, b])$, la cual esta en el intervalo $[a, b] \subset \mathbb{R}$ [6].
La profundidad más utilizada es la modal y la de Fraiman y Muniz. En este proyectos se trabajará con la profundad modal.

Profundidad Modal

Definicion: Sea $\{X_i(t), t \in \mathcal{T}\}_{i=1}^n$ una muestra funcional de una (v.f) $\mathcal{X}$, consideramos una función kernel $K : \mathbb{R}^+ \to \mathbb{R}^+$ y un parámetro suavizado h, denominado como el ancho de la banda. De esta manera tenemos que la profundidad funcional de la Moda (MD) para el dato i, viene dado por:

$$MD_{X_i} = \sum_{k=1}^{n} K\left(\frac{||X_i - X_k||}{h}\right)$$

Normalmente h es tomado como el Quinceavo percentil de las distancias que existen entre las curvas de la distribución empírica. Además, este método es muy utilizado debido a su forma útil de realizar las ponderaciones locales; es decir, realizar la ponderación alrededor de X_i es atribuir un peso teniendo presente la norma de L^2 y mientras sea mas distante X_k a X_i, la ponderación será menor ([5].
Como una sugerencia se debería utilizar el Kernel de Gauss:

$$K(t) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \qquad t > 0$$

De esta manera notamos que la Moda Funcional es el dato más profundo:

$$\hat{\mu}_{MOD} = \text{argmax}\{MD_{X_i}, i = 1, ..., n\}$$

Bandas de Confianza Bootstrap

El cálculo de las bandas de confianza es de gran ayuda, como por ejemplo en la evaluación de la precisión de un estimador de localización, en este caso sería de la media funcional. Pero para la construcción de las bandas de confianza se utilizara el procedimiento de Bootstrap suavizado. Tomando en cuenta la covarianza de los datos [15]. Entonces procedemos a detallar el procedimiento de Bootstrap Suavizado en las construcciones de las bandas de confianza:

1. Primero obtenemos b muestras Bootstarp a partir de $\mathcal{S}_n$: $\mathcal{S}_n^{*j} = \{\{X_i^*\}_{i=1}^n\}$ donde $X_i^{*(j)}(t) = X_i^*(t) + Z(t)$, con $j = 1, 2, ..., b$ e $i = 1, ..., n$, en este caso $X_i^*(t)$ se seleccionara al azar, mientras que $Z(t)$ es una variable distribuida normalmente con media igual a cero y una matriz de covarianza $\gamma\Sigma_{S_n}$; donde Σ_{S_n} sera la matriz de varianza-covarianza de la muestra y γ viene a ser el parámetro de bootstrap suavizado que manejara la variación de las remuestras.

2. Como segundo paso, apartir de las muestras originales, $X_1(t), ..., X_n(t)$ y además de las sucesivas muestras bootstrap $X_1^{*(j)}, ..., X_n^{*(j)}$, se debe calcular las estimaciones requeridas $\hat{\theta}(\mathcal{S}_n)$ y $\hat{\theta}(\mathcal{S}_n^{*j})$, respectivamente.

3. Como tercer paso se obtendrá las distancias $\left\{d(\hat{\theta}(\mathcal{S}_n)), d(\hat{\theta}(\mathcal{S}_n^{*j}))\right\}_{j=1}^{b}$, utilizando la norma L^2.

4. En el ultimo paso, una banda de confianza bootstrap, que corresponde al nivel de confianza $(1-\alpha)$ es calculado a partir del cuantil $(1-\alpha)$ de
$\left\{d(\hat{\theta}(\mathcal{S}_n)), d(\hat{\theta}(\mathcal{S}_n^{*j}))\right\}_{j=1}^{b}$

1.4.2. Método de Detección de Atípicos para Datos Funcionales

La detección de datos atípicos funcionales mediante la profundidad de los datos, se resume en la curva menos profunda se tomaría como una curva o función atípica la cual fue propuesto por (Febrero-Bande 2007) [2], en el paquete $fda.usc$ que encuentra los datos atípicos funcionales mediante la función $outliers.deph.trim$, utilizando el método de profundidad funcional (MD), Presentamos el algoritmo de estas funciones para la detección de los datos atípicos
Sea $\mathcal{X}$ una variable funcional, con una muestra de n datos funcionales, definimos los siguiente:

1. Primero definimos una medida a utilizar luego calculamos las profundidades funcionales mediante, $\{D(X_i)_{i=1}^n\}$

2. Como segundo paso debemos conseguir Θ muestras de n funciones. sin tomar en cuenta el $\alpha\,\%$ de las curvas que se encuentre menos profundas. Entonces se puede escribir a estas muestras de la siguiente manera: X_i^θ con $i^* = 1, ..., n$ y $\theta = 1, ..., \Theta$.

3. Para la tercera parte debemos determinar muestras de bootstrap suavizado $Y_{i^*}^\theta = X_{i^*}^\theta + \Gamma_{i^*}^\theta$, con $\Gamma_{i^*}^\theta$, donde $\Gamma_{i^*}^\theta(t_1), ..., \Gamma_{i^*}^\theta(t_m)$, tales que estas se distribuyan normalmente con una media 0 y con una matriz de covarianza $\gamma\Sigma_{\mathcal{X}}$, donde $\Sigma_{\mathcal{X}}$ será la matriz de covarianza de

la forma $X(t_1), ..., X(t_m)$ y γ viene a ser el parámetro de bootstrap suavizado que manejara la variación de las remuestras.

4. En el cuarto paso, se debe obtener un valor K^θ para cada una de las remuestras $\theta = 1, .., \Theta$ la cual esta relacionado al k-ésimo percentil de la distribución $\{D(Y_{i^*}^\theta)_{i^*=1}^n\}$ empírica.

5. En el quinto paso se debe tomar el límite K como la media de los Θ valores K^θ.

6. En el último paso, si se hallan curvas de tal manera que $D(X_j(t)) \leq K$, se les considerara como curvas atípicas o mejor dicho en este caso un dato funcional atípico y se procederá a eliminarlo de las muestra.

1.4.3. Gráficos de Control

Introducción

En la actualidad tenemos que la calidad es un factor muy importante en la decisión del consumidor al adquirir algún tipo de servicio o producto, sea este una persona en particular, una empresa o industria. Entonces de esta manera decimos que la calidad es un factor muy importante para el desarrollo y exíto de la empresas. Por este motivo la estadística tiene una gran acogida por las empresas para definir una estrategia y el mejoramiento de sus procesos, donde existe un grupo de procedimientos los cuales ayudan a mejorar y controlar los procesos y se conoce a este procedimiento como Control estadístico de Procesos (CEP). En este grupo de procedimientos se encuentra herramientas gráficas y mediante a esto se puede monitorizar los dichos procesos. Ya que se puede aplicar gráficos de control a secuencias o patrones no aleatorios como ayuda para que detecte variaciones fuera de control [11].

Variación

En todo proceso es común que exista variación, esto se debe a que intervienen algunos factores para la creación de uno nuevo y ocurre mediante las $6M$, la cuales son medio ambiente, método, mano de obra, me-

dición, maquinaria y material, en condiciones normales. Estas $6M$ contribuyen a las variables variación en la salida y esta aportación puede ser de 2 tipos o grupos, ya que en el transcurso del tiempo estas M pueden tener ciertos problemas como: cambios, errores, desajustes, desgaste, fallas, etc [16].

La variabilidad en los diferentes procesos o análisis se los puede clasificar en dos grupos:

- **Variación por Motivo Aleatorio:** Esta variación es independiente a los detalles de un proceso, donde este resulta de la combinación y suma de diferentes motivos, los cuales son complicados de hallarlos y eliminarlos. Sin embargo, en un tiempo prolongado existe la oportunidad de mejorarlos.
- **La Variación por Motivo Especial o Asignables:** La variabilidad en este caso puede provenir de 3 fuentes: fallas en maquinarias, errores humanos o material defectuoso. Esto ayuda a tener un mal proceso, por lo cual es de suma importancia detectar esta fuente y corregirlo o eliminarlo

(6 sigma) [16] Cuando el procedimiento trabaja con la variación por motivo aleatorio se dice que el procedimiento esta estable debido a que esta variación en el transcurso del tiempo se la puede predecir en un futuro cercano. Además en la figura (1.1), se indica un grafico bajo control. Sin embargo, si se trabaja con las variaciones por motivo especial el procedimiento se encuentra fuera de control estadístico, estos procesos tienen un comportamiento impredecible en un futuro cercano, ya que en un momento inesperado puede suceder una de estas variaciones dando un cambio a la línea central y en la variabilidad. Al no distinguir estos tipos de variabilidades pueden caer en uno de los dos $errores$ los cuales son conocidos como $error$ 1 y el $error$ 2

- **Error 1:** Realizar alguna acción ante una variación, suponiendo que es causada por un motivo especial, aunque en verdad proviene de un motivo aleatorio.
- **Error 2:** Tratar una variación, suponiendo que es causada por un motivo aleatorio, cuando este proviene de un motivo especial.

Se conoce que estos errores generán pérdidas. Sin embargo, se puede evitar caer en el primero o el segundo error, pero no en las dos. Lo que se puede hacer es evitar cometer rara vez las dos, para esto el Dr. Walter Shewhart propuso los gráficos de control en el año 1924.

Gráficos de Control

El gráfico de control es representado por un esquema como se tiene en la *figura (1.1)*, donde se tiene el eje y que representa la medición de interés en el estudio, contra el eje x la cual puede representar el número de la muestra analizada o el tiempo transcurrido. Además, los puntos representan los valores de la muestra de estudio y es común que estos puntos sean unidos con líneas rectas en los gráficos de control para poder apreciar de mejor manera el comportamiento de los puntos en el tiempo[7]. Los gráficos de control es compuesto por la línea central, límite de control superior y control inferior, que se calculan mediante la variación de los datos:

- **Línea central:** Esta línea indica que el estado del proceso estudiado se encuentra controlado y representa el valor promedio del eje y
- **Límite de Control Superior e Inferior, (LCS) y (LCI):** Los dos límites indican que si el proceso de estudio se encuentra controlado, probablemente todos los puntos de la muestra se encuentran entre estos dos límites, así representando la variabilidad del proceso y en el caso que exista un punto fuera de este rango se lo puede tomar como que el proceso se encuentra fuera de control con respecto a la tendencia central. El siguiente paso que se debe tomar es empezar una búsqueda de los motivos de la variación encontrada[7].

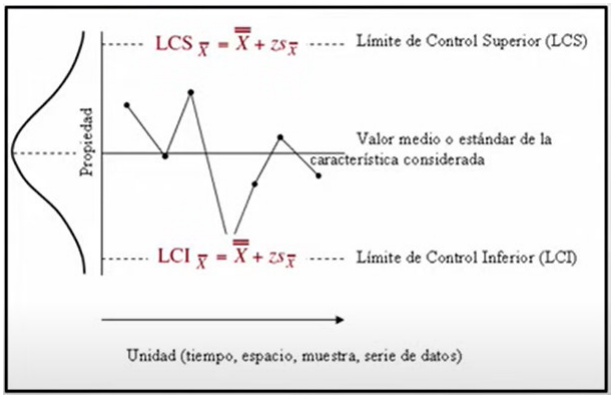


Figura 1.1: Elementos del gráfico de control

Prueba de hipótesis

La idea de los contrastes de hipótesis es muy similar a la función de un gráfico de control para que analice su rendimiento, son las siguientes decisiones.

Se rechaza la hipótesis nula cuando esta es verdadera, a este caso se lo conoce como *Error Tipo I* e indica que el proceso está fuera de control cuando en verdad no lo está.

No se rechaza la hipótesis nula cuando esta es falsa, a este caso se lo conoce como *Error Tipo II* e indica que el proceso está en control cuando en verdad no lo está.

Se tiene los siguientes escenarios.

Situación Real		
	H_0 Cierta	H_0 Falsa
No se rechaza H_0	Decisión Correcta	Error Tipo II
Se rechaza H_0	Error Tipo I	Decisión Correcta

Cuadro 1.1: Errores Estadísticos

El diagrama de control es similar al trazado de la región de aceptación de una secuencia de pruebas de hipótesis en el transcurso del tiempo. Para la monitorización de alguna característica de calidad X, se analiza una sub muestra $\{X_1, X_2, \ldots.\}$ en cada intervalo de tiempo, siguiendo una distribución F con media μ_0 y desviación estándar σ. Se prueba la hipótesis $H_0 : \mu = \mu_0$ vs. $H_\alpha : \mu \neq \mu_0$ en cada instante de tiempo para verificar si el proceso se encuentra en control[12].

Si no hay pruebas suficientes para rechazar H_0 se concluye que el proceso se encuentra bajo control; caso contrario, el proceso está fuera de control.

Limites de Control

Los límites de control se calculan con la variación del estadístico del grafico de control, para determinar estos límites se debe ser cuidadoso al momento de cubrir un porcentaje de la variación del proceso, ya que puede ser muy alto el porcentaje y no se detecten las variaciones deseadas o el porcentaje muy bajo y presentar el *error tipo 1*. Este método es conocido como *Limites de Probabilidad*, se calcula con la distribución de probabilidad del estadístico obteniendo un porcentaje definido (Duncan 1989). Otra manera es utilizando los parámetros de control los cuales son la media y la desviación estándar de y, siendo un caso particular que y siga una distribución normal con media μ_y y desviación estándar σ_y, se define los límites de la forma $\mu_y - 3\sigma_y$ y $\mu_y + 3\sigma_y$, donde el $99{,}73\,\%$ de los posibles datos de y se encuentren entre estos intervalos. Ahora en el caso de no tener una distribución normal, pero si una distribución unimodal con forma similar a una normal se puede aplicar el *Teorema Chebyshev*. Así generalizando el modelo de la forma:

$$\begin{aligned} LCI &= \mu_y - 3\sigma_y \\ LC &= \mu_y \\ LCI &= \mu_y + 3\sigma_y \end{aligned}$$

Se debe tener en cuenta que para el diseño de un diagrama de control se debe detallar el tamaño de la muestra y la frecuencia de muestreo, debido al coste para abordar estas decisiones se lo hace a través de la *longitud media de corrida* (ARL), el cual indica que el promedio de puntos que se debe graficar antes que se señale una condición fuera de control. Si no están correlacionadas las observaciones, el ARL se calcula para el grafico de control de la siguiente manera: $ARL = \frac{1}{p}$, p es el valor de la probabilidad de que un punto sobrepase un límite de control, esta ecuación nos ayuda a revisar el rendimiento, en ciertos casos una mejor manera de expresar el rendimiento es mediante el termino de

Tiempo Promedio a la Señal (ATS), su ecuación es $ATS = ARL \cdot h$, h es el intervalo fijo en horas, en el cual se toman muestras. Mientras más pequeño sea el ARL en el diagrama de control el desempeño es mejor ya que detecta con mayor rapidez un desajuste.

Fases de un Gráficos de Control

Para la construcción de un gráfico de control normalmente se utilizan dos fases, las cuales tienen como objetivo diferentes escenarios. La fase I, Consiste en que se recopile y analice un conjunto de datos del proceso asumiendo que este proceso se encuentra en control, son conocida como datos históricos o muestra de calibrado del proceso, siendo de utilidad para estimar los parámetros. Además, se recomienda que los *límite de control inferior* y *límite de control superior* naturales se estime mediante una muestra de calibrado de 20 o más datos [12]. Los límites de control, nos ayude a indicar si el proceso ha estado en control, durante el momento que se tomó los datos históricos y verificar si es una buena decisión ocupar estos datos para controlar los procesos futuros, caso contrario se identificará los principales motivos de la variación y se aplicaran medidas para corregirlo, para obtener una estimación de los límites de control para la variable que corresponde al proceso en control, esto se va a repetir hasta tener un conjunto de datos históricos en control. La fase II, toma inicio cuando se tenga los datos históricos tratados o $limpios$ del proceso, que sea estable y además representativo monitorizando el proceso, verificando los nuevos datos con los límites de control del previo estudio.

Tipos de Gráficos de Control

A los gráficos de control se los puede dividir en 2 grupo generales, el primero para variables y el segundo para atributos.
Para las variables son todas aquellas que puede medirse o ser representado por un número en una escala continua, para realizar estas mediciones se necesita diferentes tipos de instrumentos.
Los gráficos de control para las variables más comunes son las siguientes:

- $\overline{X}$ (Promedios)
- R (Rangos)
- S (Desviación Estándar)
- X (Medidas individuales)
- T^2 de Hotelling (Multivariante)

$\overline{X}$ **y** R

El gráfico de $\overline{x}$ es utilizado para controlar la media y para la variabilidad se monitorea mediante un diagrama de control para el rango R. Para los límites de control, primero se calcula la media y el rango de las muestras $\overline{x}_i$ y R_i , para $i = 1, 2, \ldots, n$, de la siguiente manera:

$$R = \frac{\sum_i R_i}{k} \qquad\qquad \overline{x} = \frac{\sum_i \overline{x}_i}{k}$$

Además, tenemos que R_i/d_2 es un estimador insesgado y d_2 viene a ser un estadístico de rango, representando la media del rango relativo dado por $y = \frac{R}{\sigma}, \frac{R}{d_2} = \frac{\sum_i R_i}{d_2 k}$ viniendo a ser un estimador centrado de la desv. estándar. Luego se contrasta si para cada valor de $\overline{x}$ esta en este intervalo.

$$\left[\overline{x} - \frac{3R}{\sqrt{n}d_2}, \overline{x} + \frac{3R}{\sqrt{n}d_2}\right]$$

teniendo una probabilidad de $97{,}73\,\%$ aproximadamente. Además tenemos que:

$$\frac{3}{\sqrt{n}d_2} = A_2$$

Entonces los límites de control quedarían expresados de la siguiente manera.

$$\begin{aligned} LCI &= \overline{x} - A_2 R \\ LC &= \overline{x} \\ LCs &= \overline{x} + A_2 R \end{aligned}$$

A_2 viene a ser una constante de tabulación para diferentes tamaños de muestras.

$\overline{X}$ **y** S

En este caso para la variabilidad se monitorea mediante un diagrama de control directamente de la desviación estándar. Para los límites de control, primero se calcula la media y la desviación estándar de la muestras $\overline{x}_i$ y s_i , para $i = 1, 2, \ldots, n$, de la siguiente manera:

$$\hat{\sigma} = \frac{\overline{S}}{C_2} = \frac{\sum_i \overline{S}_i}{C_2 k}$$

Además, tenemos un estimador centrado de la desviación estándar teórica. c_2 representa una constante y dependerá del tamaño de la muestra.

$$\left[\overline{x} - \frac{3\overline{S}}{\sqrt{n}C_2}, \overline{x} + \frac{3\overline{S}}{\sqrt{n}C_2}\right]$$

teniendo una probabilidad de $97{,}73\,\%$ aproximadamente. Además tenemos que:

$$\frac{3}{\sqrt{n}C_2} = A_1$$

Entonces los límites de control quedaráan expresados de la siguiente manera.

$$\begin{aligned} LCI &= \overline{x} - A_1 R \\ LC &= \overline{x} \\ LCs &= \overline{x} + A_1 R \end{aligned}$$

A_1 viene a ser una constante de tabulación para diferentes tamaños de muestras.

Los gráficos de control para atributos no son medidas sobre una escala continua. En este caso son cada unidad generada por el proceso como conforme o no conforme, en base a que presente algún tipo de atributo o se enumeren los defectos que se aprecie en cada unidad. Los gráficos de control para atributos más comunes son:

- p (Proporción)
- np (Número de unidades defectuosos)
- c (Número de defectos)
- u (Número de defectos por unidad)

De las gráficas de control presentadas para el caso univariante, es decir, se intentaba monitorizar únicamente 1 variable. Además, aún existen más graficas de control de las mencionadas como referencia pueden revisar el libro (montgomery 1991) [11].

Gráfico de Control Multivariante

El gráfico de control Multivariante hace presencia cuando se desea monitorizar varias características de calidad en el mismo instante, tal que su respuesta viene dada por un vector de medida p características de calidad, se lo tiene de la siguiente manera $x^T = (x_1, x_2, \ldots, x_p)$. Así, de esta manera tenemos un proceso multivariado. El primer enfoque del análisis fue tomar una gráfica de control para cada característica y tuvieron varios inconvenientes, ya que no se dieron cuenta de que tienen correlación entre las características de calidad. Debido a este problema se introdujo los gráficos de control T^2 de Hotelling para el monitoreo del proceso multivariado.

T^2 de Hotelling

En la propuesta de Hotelling, se empieza con la suposición fuerte, la cual es que las medidas sucesivas de las características sigan una distribución normal multivariada, pero la normalidad no se puede suponer siempre.
En el caso que se desee controlar la media del proceso, el método a seguir es calcular la distancia del vector bivariado medio de cada muestra hacia la media del proceso, para este caso se toma la distancia de Mahalanobis.
Distancia de Mahalanobis. Esta distancia entre dos vectores aleatorios

con la misma distribución de probabilidad $\vec{X}$ y $\vec{Y}$ con matriz de covarianzas $\sum$ se define como:

$$d_p(\vec{X},\vec{Y}) = \sqrt{(\vec{X}-\vec{Y})^T \cdot \Sigma^{-1} \cdot (\vec{X}-\vec{Y})}$$

T^2 de Hotelling Límite de Control

Por lo mencionado anteriormente tenemos que en estos gráficos de control las observaciones son $p-variados$ y siguen una distribución $N_p(\mu,\Sigma)$, Normal p-variada con vector de media μ y la matriz de varianzas Σ. Ahora tomamos $x_1, x_2, \ldots, x_n$ v.a.i con distribución $x_1, x_2, \ldots, x_n \sim N_p(\mu,\Sigma)$ se define la media muestral como:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

se tiene la siguiente estimación:

$$T^2 = n(\overline{x}-\mu)^T \Sigma^{-1} (\overline{x}-\mu) \sim \chi^2_p$$

De la demostración mencionada se tiene que el límite de control superior que se debe tomar es el cuantil de esta distribución, de la siguiente forma.

$$LCS = \chi^2_{\alpha \cdot p}$$
$$LCI = 0$$

Donde $\chi^2_{\alpha\cdot p}$ es $100 \cdot (1-\alpha)$ de la distribución Ji-cuadrado con p-grados de libertad.

Si μ y Σ Son desconocidas Se estimaría de los m conjuntos de observaciones que son tomadas cuando suponemos que el proceso esta en control, es decir de los datos históricos del proceso. Si en el caso de que cada uno de los m conjuntos tenga n observaciones. Sus estimaciones serian de la siguiente manera:

$$\overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}, k = 1, ..., m.$$
$$S^2_k = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ik}-\overline{x}_k)(x_{ik}-\overline{x}_k)^T, k = 1, ..., m.$$

Ahora tomamos la media tomando en consideracion todos los grupos, obtenemos lo siguiente:

$$\overline{\overline{x}} = \frac{1}{m}\sum_{i=1}^{m}, \overline{x}_k \qquad\qquad S = \frac{1}{m}\sum_{i=1}^{m} s_k^2$$

Teniendo la siguiente estimacion:

$$T^2 = n(\overline{x} - \overline{\overline{x}})^T S^{-1}(\overline{x} - \overline{\overline{x}})$$

Como se mencionó anteriormente existen dos fases para realizar la construcción de los gráficos de control. Sin embargo, en este trabajo se centrara únicamente en la Fase I, la cual utiliza los gráficos de control para confirmar que los datos históricos que fueron usados para estimas $\overline{\overline{x}}$ y S, se obtuvieron de un proceso de control estadístico.
De (Montgomery)[12] se tiene que la Fase I los límites de control son los siguientes:

$$LCS = \frac{p(m-1)(n-1)}{mn - m - p + 1} F_{\alpha,p.mn-m-p+1}$$
$$LCI = 0.$$

Cuando μ y Σ sean estimadas mediante un número de datos históricos muy grandes, es común utilizar el límite de control superior $LCS = \chi^2_{\alpha,p}$.

Capítulo 2

Metodología

2.1. Aplicación de Gráficos de Control Funcional Para Monitorizar Series Temporales Meteorológicas y Climáticas

El marco teórico presentado en el capítulo 1, sobre la metodología del *Análisis de Datos Funcionales* y *Gráficos de Control* será de gran utilidad, en el presente capitulo el cual se desarrollara en la aplicación de gráficos de control funcional para series temporales meteorológicas y climáticas, detallando los paso importantes al momento de tratar los datos hasta obtener las conclusiones, cabe recalcar que en el presente proyecto se utilizara el software estadístico R.

2.1.1. Descripción de los datos

La base de datos se obtuvo del Grupo de Energías Alternativas y Ambiente (GEAA). Estos datos están organizados en distintos libros de Excel $(.csv)$, donde consta su respectiva fecha y estación. [1]

La información es recolectada de manera automática y descargada

[1] https://liveespochedu-my.sharepoint.com/:f:/g/personal/estaciones_espoch_espoch_edu_ec/EmR4sgWPDvNGshM9JBqpuj4B_lR9KH1eV1J-imO8uV3M_A?e=7RH77c

de las estaciones en archivos con formato dat, que mediante el software Veisala transforma a UTM en archivos en formato csv, las cuales tienen 29 variables, para el presente proyecto únicamente se tomara en cuenta la información de los años 2015 a 2016 y seleccionaremos 7 de las 14 estaciones con las variables $X1$, $X2$ y $StatTA$, no tomamos las otras estaciones debido a un gran número de datos faltantes, las variables en la base representan:

- *X1:* La fecha en que fue recolectada la información
- *X2:* La hora en que fue recolectada la información.
- *Stat TA 1m:* Temperatura Ambiente por minuto

Estructura de los datos

Una vez ya seleccionados los datos y variables, crearemos una base de datos con la información de la estación de $ALAO$, notamos en la tabla 2.1 que los datos se encuentran en una sola columna y como se explicó, la información está ingresada por minutos por lo cual se debe realizar un tratamiento de los datos para las 7 estaciones.

date,time,Avg
1/1/14,12:00:08 AM,17.575
1/1/14,12:01:08 AM,16.231
1/1/14,12:02:08 AM,17.075
1/1/14,12:03:08 AM,15.575
1/1/14,12:04:08 AM,16.991

Cuadro 2.1: Información Original de la estación ALAO

Debido a que la información se encuentra en minutos procedemos, agruparlos calculando el promedio por horas. Así, discretizamos los datos por horas de cada estación obteniendo 7 matrices de 730 días por 24 horas que cuenta con la información desde el año 2016 hasta el 2017 de la *Temperatura Ambiente*.

Las estaciones no tienen valores perdidos, por lo cual no se realiza una imputación previa, ahora en la *figura 2.1* se tiene la *Temperatura Ambiente*. de $ALAO$ durante los 2 años que se esta evaluando en este proyecto, el grafico esta representando los valores discretos, es decir por puntos

en la cual se puede apreciar el comportamiento de la temperatura en el transcurso del día, también vemos la existencia de datos atípicos en la estación, esta información es similar en las demás estaciones, por tanto tomamos uno para representar la idea general.

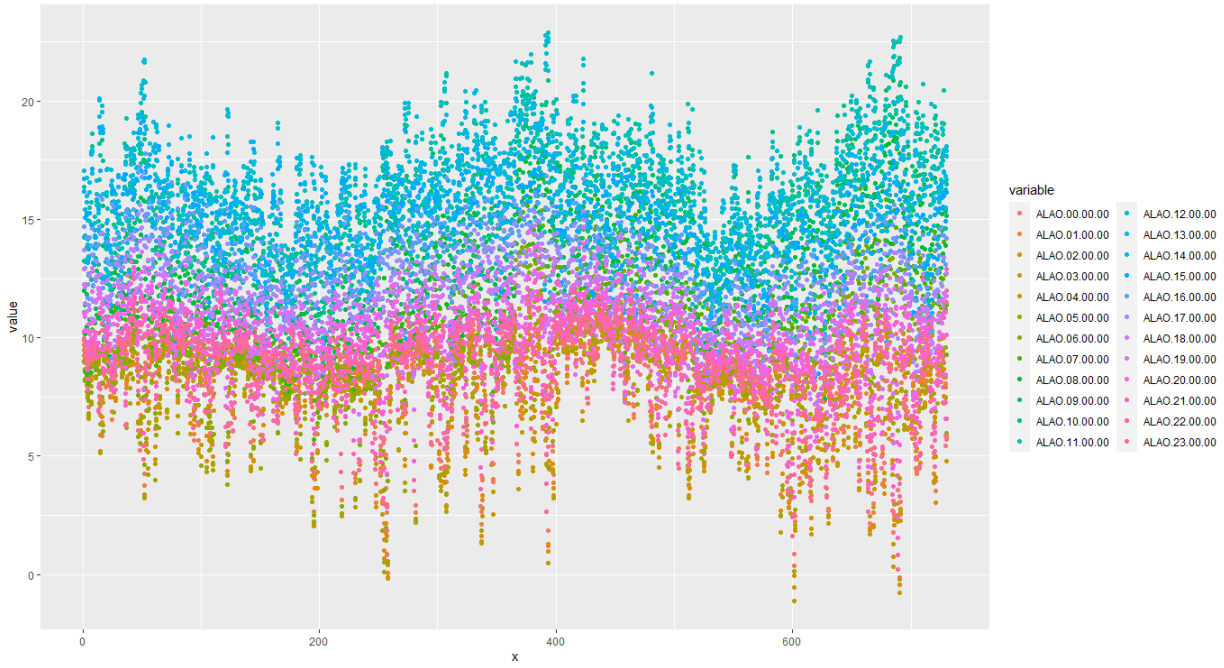


Figura 2.1: Temperatura Ambiente de ALAO por puntos

2.1.2. Tratamiento Funcional

Mediante los datos discretos construiremos datos funcionales con uso de la función $fdata$ del paquete fda teniendo un total de 730 curvas en un intervalo de 24 horas esto se cumple para las 7 estaciones, donde estas curvas o funciones deben cumplir estas características: Evaluar la tasa de cambio, reducir ruido y deben tener una escala de tiempo en común. Al conjunto de estas características se le conoce como el *método de suavizado* y su funcionamiento es reducir la cancelación del efecto de la variación aleatoria, cuando se realiza una buena aplicación del suavizado se nota con mayor claridad diferentes características como los componentes cíclicos, estacionales y tendencia subyacentes. Esté método es de suma importancia para realizar un análisis funcional.

Una función o curva se la puede representar mediante una base cuando asumimos que los datos están en L^2. Una base es un grupo de fun-

ciones conocidas $\{\phi_k\}_{k\in\mathbb{N}}$, donde cualquier función podría aproximarse arbitrariamente con una combinacion lineal de un numero k suficientemente grande [14]. Aproximando una función $X(t)$ usando la expansión de una base truncada en términos de K funciones bases conocidas.

Existe varios tipos de bases para realizar la aproximación. En el presente proyecto se hace uso de la *base de fourier* para aproximar la forma funcional de los datos, debido a que los datos presentan periocidad o la posibilidad de tener un patrón fijo.
Para determinar de manera correcta el K(Numero de bases), no existe un método universal que permita una elección óptima. Se utilizara el criterio de validación cruzada generalizada para minimizar el criterio de GCV, la cual esta implementada en el software estadístico R, mediante la función optim.basis. Además, de calcular el k optimo también determina la penalización de rugosidad optimo λ, minimizando el GCV. Este criterio es representado por el calculo de los promedios de GCV del dato funcional. Entonces, ingresamos 2 vectores para k y λ respectivamente Tabla(2.2), en la función optim.basis el cual retorna el par óptimo para el suvisamiento de cada estación tabla(2.3) notando que no existe mucha variación en los valores esto se debe a que las estaciones tienen un comportamiento similar.

k	5	8	11	15	18	22	25	29
λ	0.0312	0.0625	0.125	0.25	0.5	1	2	4

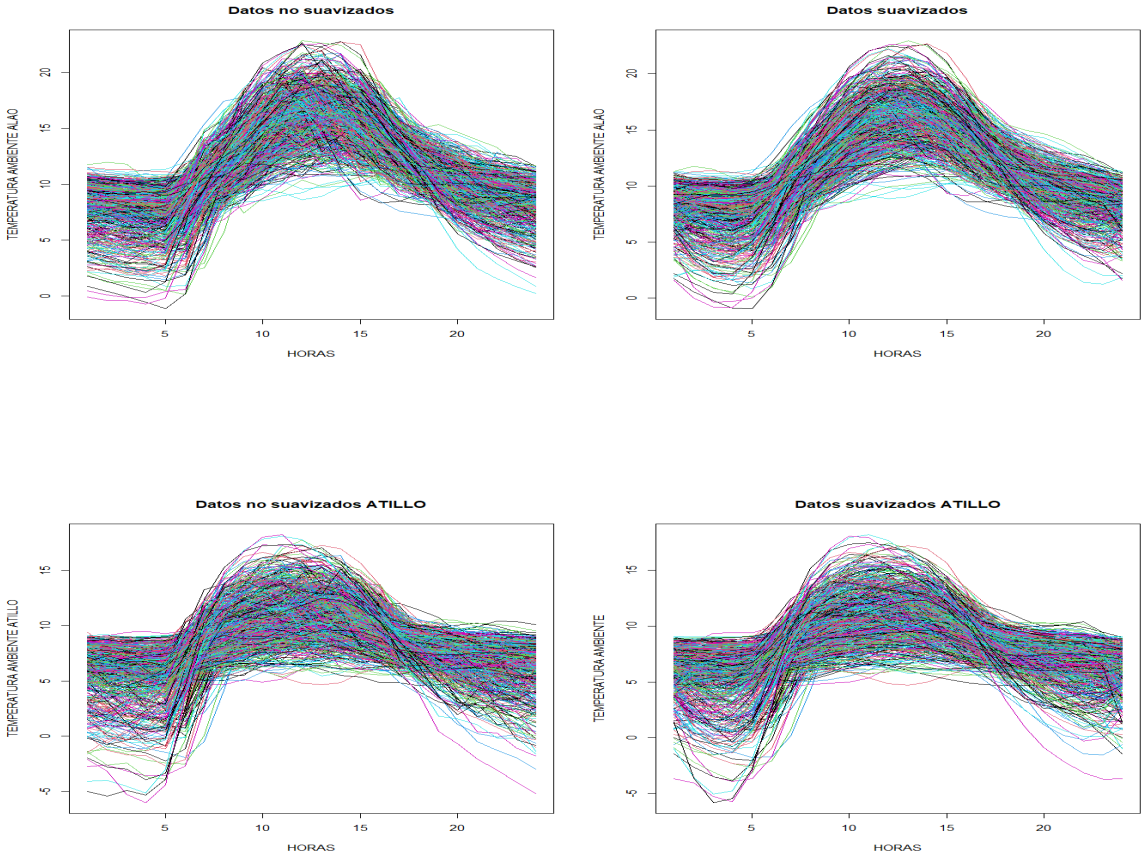
Cuadro 2.2: Posibles Valores de k y λ

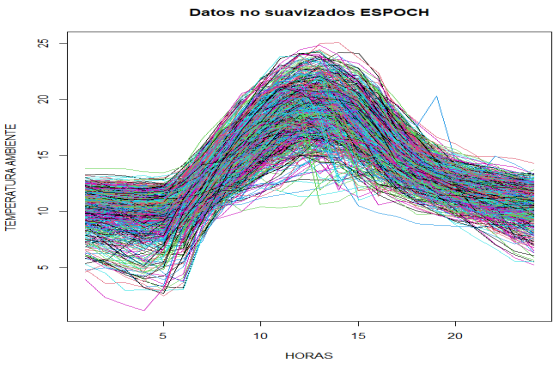
Estación	k optimo	λ optimo
Alao	15	0.25
Atillo	11	0.25
Espoch	15	0.25
Quimiag	15	0.25
San Juan	15	0.25
Tixan	15	0.25
Urbina	15	0.125

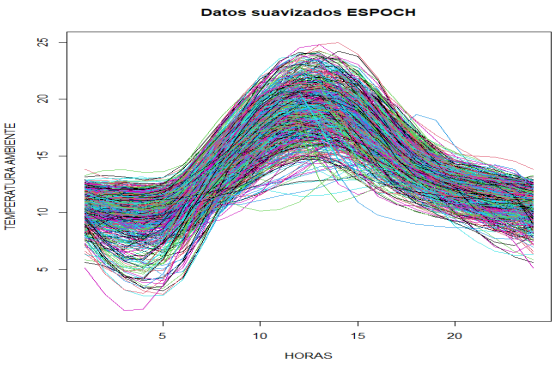
Cuadro 2.3: k y λ óptimo de las 7 estaciones

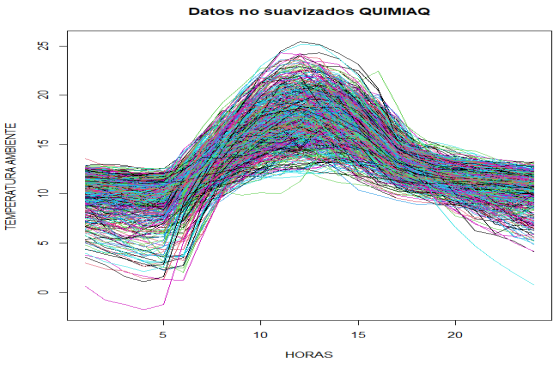
Suavizado

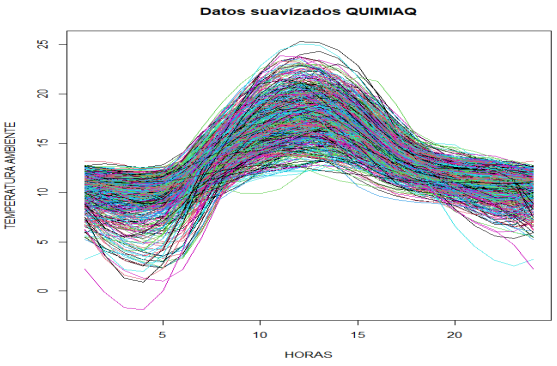
Como primera parte del análisis de datos funcionales consiste en definir la función $X_n(t)$ empezando desde los datos reales o originales, a esta aplicación se la conoce como suavizado de la curva o función. Así, de esta manera se realizó el ajuste de las estaciones, en esencia lo que se desea realizar es eliminar los movimientos pequeños en la información para así poder conservar la forma correcta y que el ajuste logre explicar el comportamiento de los datos.
Como se puede notar en la figura $(2{,}2)$ tenemos en lado izquierda las estaciones sin ajustar y en el derecho las ajustadas.

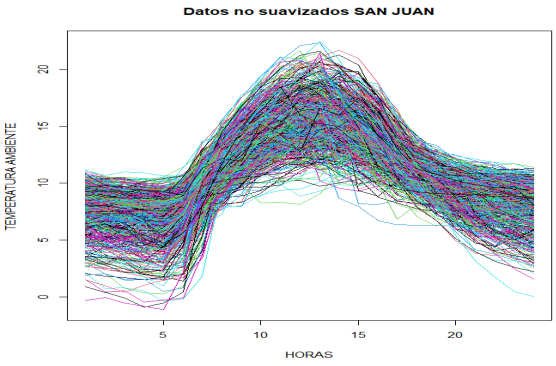


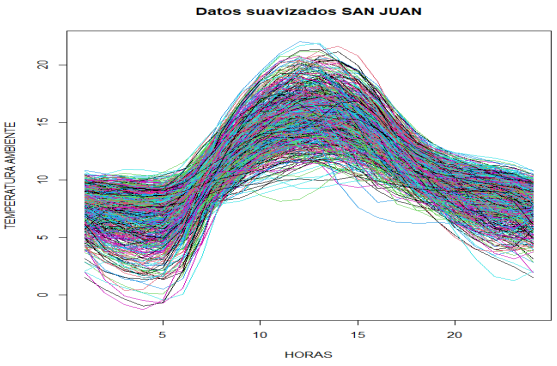

Datos no suavizados ESPOCH
TEMPERATURA AMBIENTE
HORAS


Datos suavizados ESPOCH
TEMPERATURA AMBIENTE
HORAS


Datos no suavizados QUIMIAQ
TEMPERATURA AMBIENTE
HORAS


Datos suavizados QUIMIAQ
TEMPERATURA AMBIENTE
HORAS


Datos no suavizados SAN JUAN
TEMPERATURA AMBIENTE
HORAS


Datos suavizados SAN JUAN
TEMPERATURA AMBIENTE
HORAS

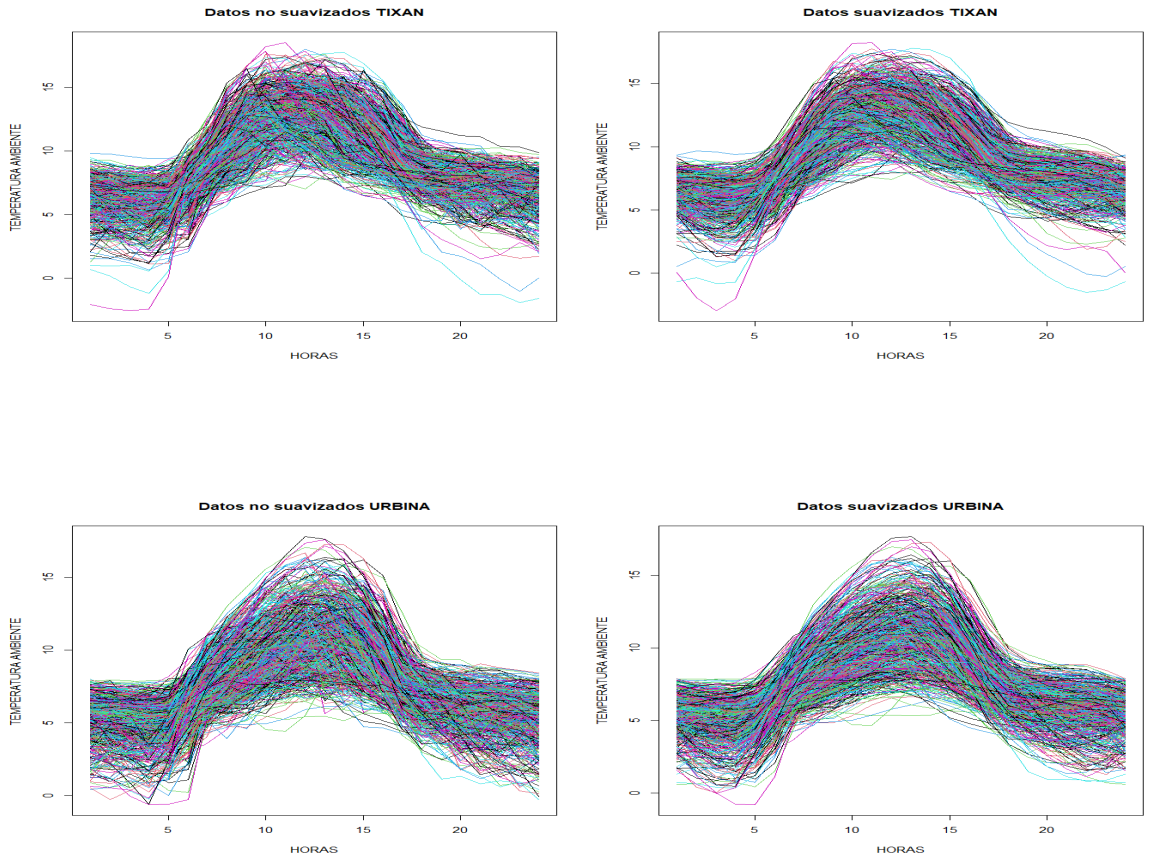


Figura 2.2: Grafico funcional de las curvas no suavizadas y suavizadas por estación

2.1.3. Análisis Descriptivo Funcional

Una vez que los datos ya pasaron por el proceso de trasformación, seguimos con el siguiente paso el cual es un análisis descriptivo y exploratorio para verificar su valides y lo más importante, explicar la información real de la mejor manera mediante las curvas o funciones obtenidas. Además, algunas características descriptivas se propuso por [14]:

- La media funcional.

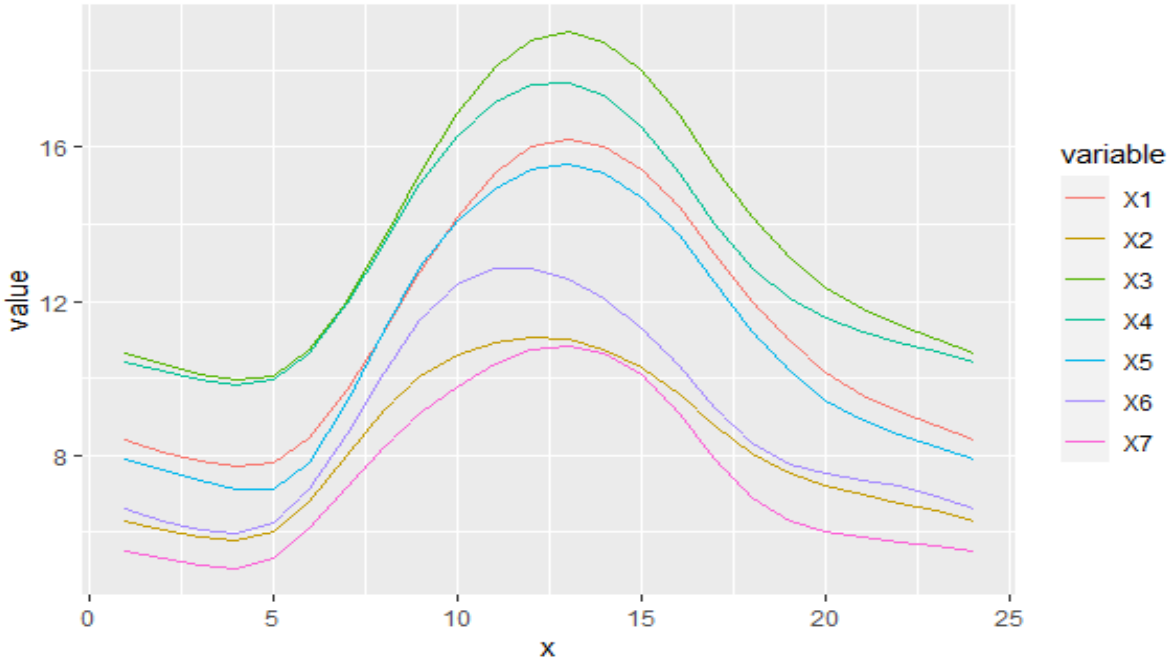


Figura 2.3: Representación de la media funcional de las 7 estaciones

En el gráfico (2.3) se representa la media funcional de las estaciones donde el orden esta dado por X1=Alao, X2=Atillo, X3=Espoch, X4=Quimiag, X5=San Juan, X6=Tixan, X7=Urbina'. Podemos notar que las estaciones tienen un comportamiento similar en el transcurso del tiempo y el cambio de temperatura ambiente. Sin embargo, la estación de Espoch, Quimiaq tiene una diferencia aproximadamente de 4 a 5 grados con las estaciones de Urbina, Atillo, y Tixan.

- La varianza funcional.

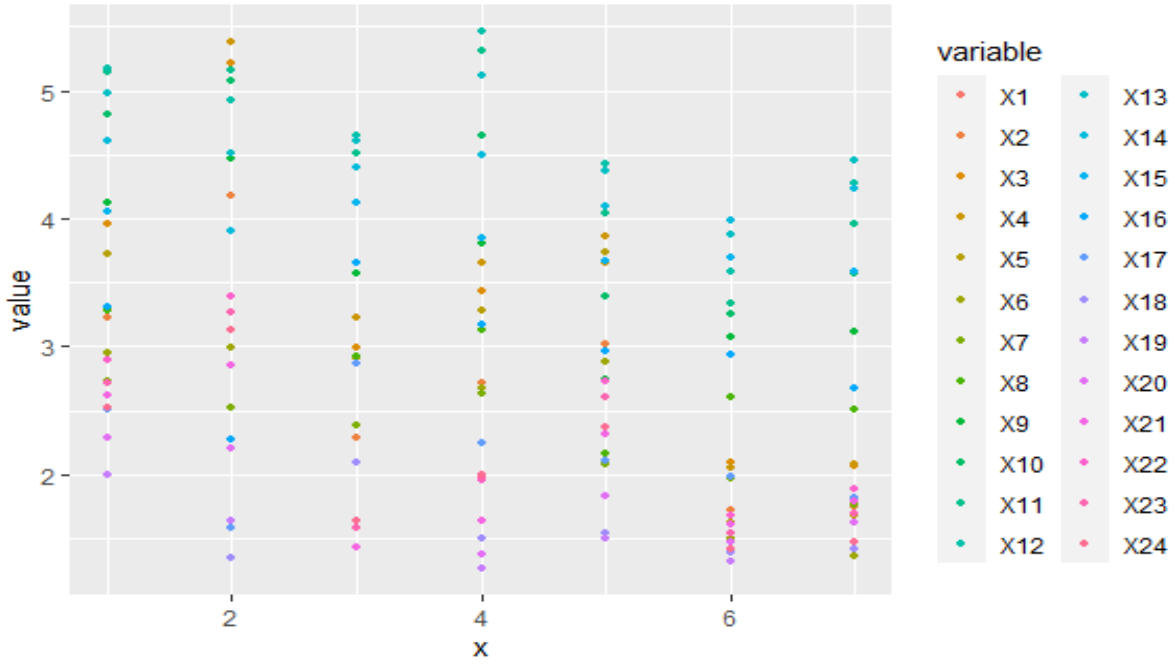


Figura 2.4: Representación de la variación funcional de las 7 estaciones

Notamos la variación de la temperatura ambiente de cada estación durante el día, representado por cada hora. donde la variación mas grande esta dada entre la 13 y 16 horas en las estación de Quimiag.

Derivadas

En esta sección mediante las funciones suaves, existe la posibilidad de usar la información que ocurre respecto a la variación, por normalidad para la derivada se le ocupa a la primera y segunda, pues recordando resultados fundamentales del cálculo:
La primera derivada, que representa una expresión de la pendiente por punto de la función, se la relaciona con el comportamiento de la función real, es decir, si es creciente, decreciente o constante en los intervalos.
la segunda derivada tiene una similitud con la curvatura de la función real, lo que en cierto modo se lo puede ver como un ritmo del crecimiento.
Para realizar esta operación se hará uso de la función $fdata.deriv$, el cual no ayudara a entender el comportamiento de los datos funcionales.
Calcularemos la primera derivada de cada estación.

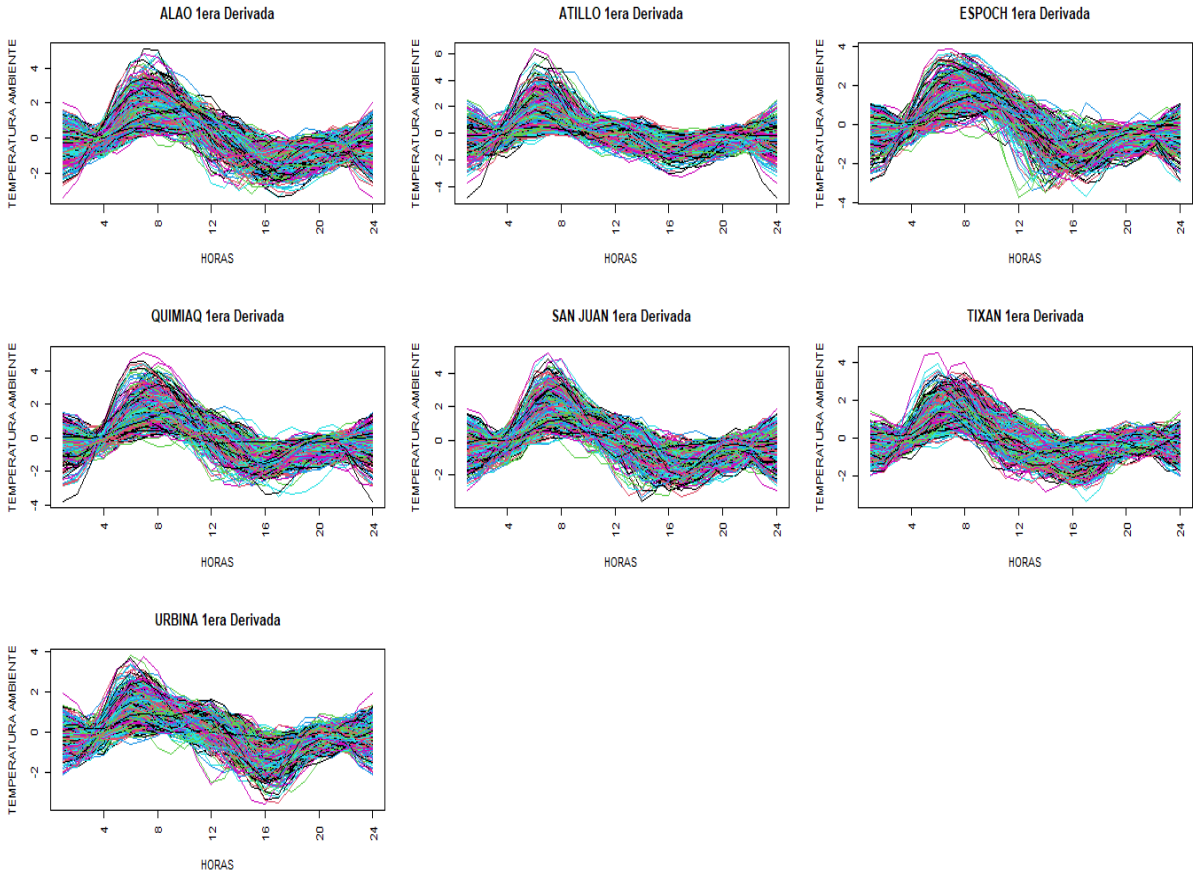


Figura 2.5: Derivadas funcionales de las 7 estaciones

Notamos en la figura $(2{,}5)$, que las funciones ajustadas están creciendo y decreciendo en el caso de las 7 estaciones teniendo estos cambios alrededor de las 7 de mañana donde el sol va tomando más fuerza e incrementando la temperatura y alrededor de las 4 de la tarde cuando esta bajando la temperatura lo cual concuerda con la realidad por tanto el ajuste funcional de los datos discretos se la hizo correctamente. Además, el análisis diferencial se lo ocupa para buscar variabilidad [14].

2.1.4. Gráficos de control funcionales

En esta sección realizaremos la monitorización de las estaciones con datos funcionales, donde se toma en cuenta la profundidad de la función la cual si cae más allá del límite inferior (LCI) se los detecta como datos atípicos de la carta de control de Fase I y se consideraría que el proceso se encuentra fuera de control. El (LCI) se obtiene un remuestreo bootstrap

suavizado.

Construcción de un gráfico de control Fase I

En estos días el manejo de grandes volúmenes de datos conocida como $BigData$, a establecido que no hay diferencia entre el grafico de control de Fase I con la Fase II.
Por tanto, en el presente proyecto únicamente se trabajará con la Fase I, que tiene como objetivo detectar anomalías y la estimación de los parámetros para la carta de control, (Febrero) [3] propone un método de detección de anomalías para construir un grafico de Control Fase I, con el propósito de tener una muestra de calibrado de un proceso en control y que se pueda monitorizar con un gráfico de control de rangos.
Entonces procedemos a realizar los gráficos de control funcionales para cada estación que esta compuestas por n variables aleatorias funcionales $\mathbf{X}$, con $\mathbf{X} \in L^2(T)$, $T \subset \mathbb{R}$
Para la fase I el análisis de las observaciones históricas que corresponden a la muestra de calibrado de dimensión n, con el propósito de probar la estabilidad y estimar el parámetro de la carta de control [1]. Una carta de control nos lleva a un contraste de hipótesis donde no hay cambios de distribución de los valores de las variables en t $\mathcal{X}_1(t), \mathcal{X}_2(t), ..., \mathcal{X}_n(t)$. Entonces la prueba de hipótesis para el contraste en la fase I es:

$$H_0 : \mathcal{X}_i(t) \overset{d}{=} \mathcal{X}_j(t), \forall i, j \in \{1, ..., n\}$$
$$H_a : \mathcal{X}_i(t) \overset{d}{=} \mathcal{X}_j(t), \textit{para algún } i, j \in \{1, ..., n\}$$

a este proceso de estabilización se debe implementar a un método iterativo que detecte y elimine a los valores que se aleje o desvíe de la forma que generen las demás funciones. (Febrero)[3], propuso un método que usa las profundidades funcionales en los cuales se asume independencia teniendo un contraste de hipótesis para identificar si las funciones son atípicas o no. Conocemos que los datos atípicos son las funciones con menor profundidad, de esta manera el se debe calcular un cuantil o $trim$ (corte), que identifica los datos por debajo este valor como atípicos en este proyecto se usara la profundidad modal para la detección de atípicos.
Entonces consideramos la v.a.f $\mathcal{X}$, de la donde tomamos una muestra

aleatoria $\mathcal{X}_1(t), \mathcal{X}_2(t), ..., \mathcal{X}_n(t)$. Así, procedemos con los siguientes pasos para realizar el grafico de control.

1. Calculamos la profundidad de cada valor respecto a los datos, $D(\mathcal{X}_2)_{i=1}^n$ realizando sus respectivos gráficos de la profundidad.
2. Escojemos el limite de control inferior en función del α del grafico de control, en el presente proyecto se hará uso de bootstrap basado en recorte para estimar el LCI.
3. Si existen funciones tal que $D(\mathcal{X}_i) \leqslant LCI$, a esta curva o función se la considera atípica y por ende el proceso no esta en control.
4. Además, se grafica un las funciones iniciales y su envolvente funcional replicas bootstrap con un alto índice de profundidad.

con las funciones detectadas como atípicas procedemos a eliminarlas repitiendo los pasos hasta que el proceso se encuentre en control.

Medida de profundidad modal

Como el nombre lo indica esta profundidad utiliza la definición de moda la cual indica que ocupa el kernel para ponderar una función en contra de las otras funciones este procedimiento se realizara mediante el uso del paquete $fda.usc$ y la función $depth.MD$. Además, tomaremos un $\alpha = 0{,}15$.

Ahora que definimos la profundidad procedemos a ocupar la función qcr del paquete qcr para realizar los 4 pasos directamente la cual esta compuesta por las funciones de $outliers.depth.trim$ que realiza el remuestreo bootstrap suavizado con 200 muestras y un parámetro de bootstrap suavizado de valor $\gamma = 0{,}10$ con la función de profundidad modal, incluye también la función $quantile$ que da como resultado el cuantil de profundidad LCI para detectar los datos atípicos, esta función qcr nos retorna dos gráficos las curvas iniciales con sus respectivas bandas obtenidas por bootstrap suavizado y las curvas atípicas, también el gráfico de control con el quantil calculado, sin embargo realizamos una modificación al paquete para que se ajuste a nuestros requerimientos. Ahora proseguimos

a calcular las cartas de control por estación:

En el conjunto de Figuras notamos claramente la existencia de datos atípicos en cada uno de los gráficos de control y el grafico que hace uso de la envolvente funcional la cual indica las funciones atípicas que están representadas gráficamente por las líneas entrecortadas de color negro, debido a la difícil detección del día que la función se encontró fuera de control procedemos a realizar una tabla con las funciones atípicas y su profundidad para cada estación con un resumen de lo obtenido.

Estación ALAO

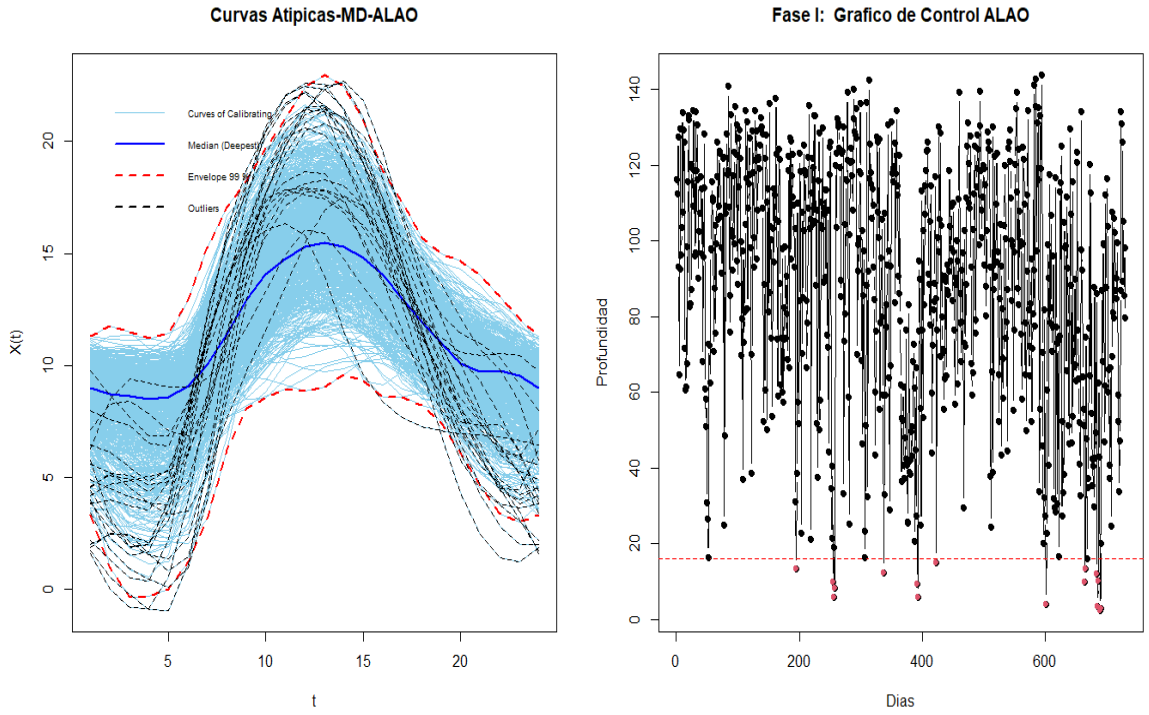


Figura 2.6: Gráfico con base a la envolvente funcional y gráfico de control funcional para la estacion de ALAO

n	FECHA	Profundidad	n	FECHA	Profundidad
1	2015-07-14	13,25	11	2016-10-27	13,3
2	2015-09-12	9,676	12	2016-11-15	12,05
3	2015-09-14	5,671	13	2016-11-16	3,288
4	2015-09-15	8,133	14	2016-11-17	10,07
5	2015-12-03	12,23	15	2016-11-20	2,413
6	2016-01-27	9,378	16	2016-11-21	2,932
7	2016-01-28	5,734	17	2015-02-21	14,87
8	2016-02-27	15	18	2015-11-03	13,53
9	2016-08-24	3,971	19	2016-09-14	15,44
10	2016-10-26	9.835	20	2016-10-30	13,41

Cuadro 2.4: Tabla de funciones atípicas de ALAO

Mediante la funcion $outliers.depth.trim$ se obtuvo el cuantil con un valor de $15{,}688$ de la estación de ALAO con el cual se identifica las funciones atípicas de la estación de ALAO en el gráfico de control funcional, obteniendo un total de 20 días atípicos los cuales se indican en la tabla $(2{,}4)$.

Estación ATILLO

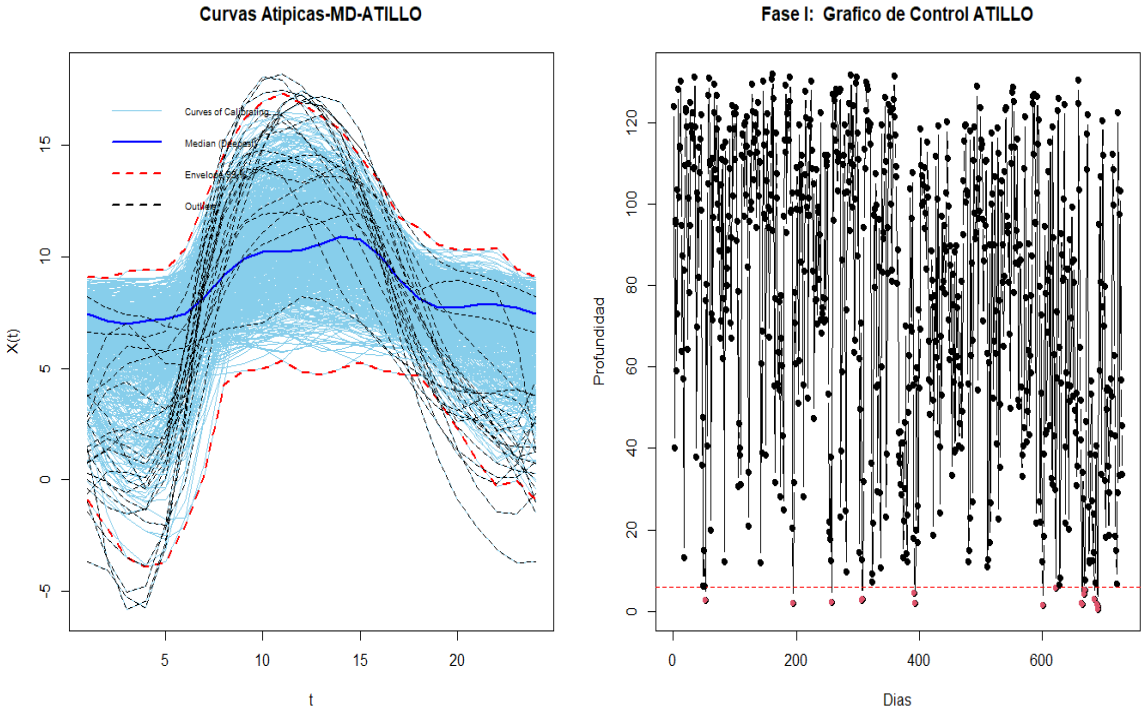


Figura 2.7: Gráfico con base a la envolvente funcional y gráfico de control funcional para la estación de ATILLO

n	FECHA	Profundidad	n	FECHA	Profundidad
1	21/2/2015	25,342	12	30/10/2016	40,494
2	14/7/2015	19,059	13	31/10/2016	52,098
3	15/9/2015	22,375	14	16/11/2016	28,491
4	3/11/2015	27,492	15	20/11/2016	17,255
5	4/11/2015	28,146	16	21/11/2016	0,520
6	27/1/2016	43,119	17	22/11/2016	12,655
7	28/1/2016	19,412	18	18/2/2015	56,186
8	24/8/2016	14,013	19	20/2/2015	55,535
9	14/9/2016	57,099	20	20/9/2016	57,631
10	26/10/2016	19,596	21	22/12/2016	51,034
11	27/10/2016	17,257			

Cuadro 2.5: Tabla de funciones atipicas de ATILLO

En este caso obtuvimos el cuantil con el valor de $5,98$ de la estación de ATILLO con el cual se identifica las funciones atípicas del grafico de control funcional ya que es LCI, obteniendo un total de 21 días atípicos los cuales se indican en la tabla $(2{,}5)$.

Estación ESPOCH

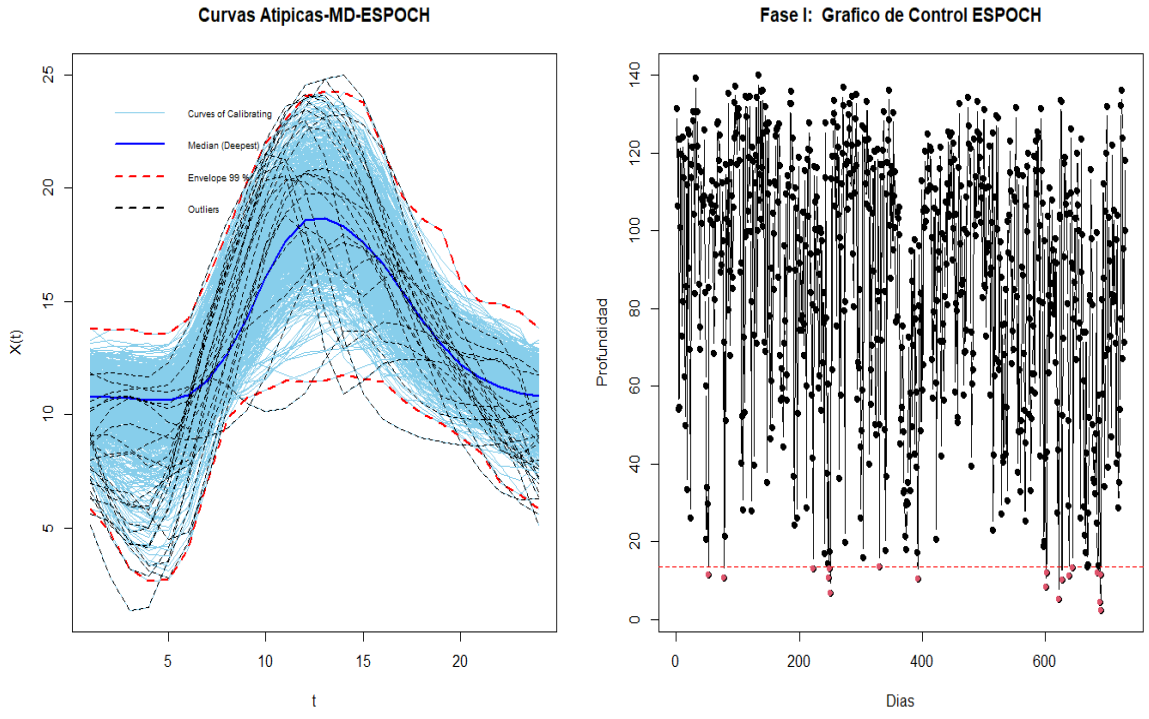


Figura 2.8: Gráfico con base a la envolvente funcional y gráfico de control funcional para la estación de ESPOCH

n	FECHA	Profundidad	n	FECHA	Profundidad
1	2015-02-21	11,288	12	2016-09-19	10,178
2	2015-03-18	10,726	13	2016-10-01	11,161
3	2015-08-11	13,006	14	2016-10-07	13,257
4	2015-09-05	10,518	15	2016-11-16	11,993
5	2015-09-06	13,018	16	2016-11-20	4,270
6	2015-09-08	6,803	17	2016-11-21	2,375
7	2015-11-26	13,382	18	2016-11-22	11,402
8	2016-01-28	10,404	19	2015-09-03	11,649
9	2016-08-24	8,212	20	2016-10-30	11,177
10	2016-08-25	11,830	21	2016-10-31	12,617
11	2016-09-14	5,126	22	2016-11-17	12,154

Cuadro 2.6: Tabla de funciones atípicas de ESPOCH

Para la estación de ESPOCH el cuantil calculado tiene el valor de $13,4323$, con el cual se identifica las funciones atípicas del grafico de control funcional ya que es el LCI, obteniendo un total de 22 días atípicos los cuales se indican en la tabla $(2{,}6)$.

Estación QUIMIAG

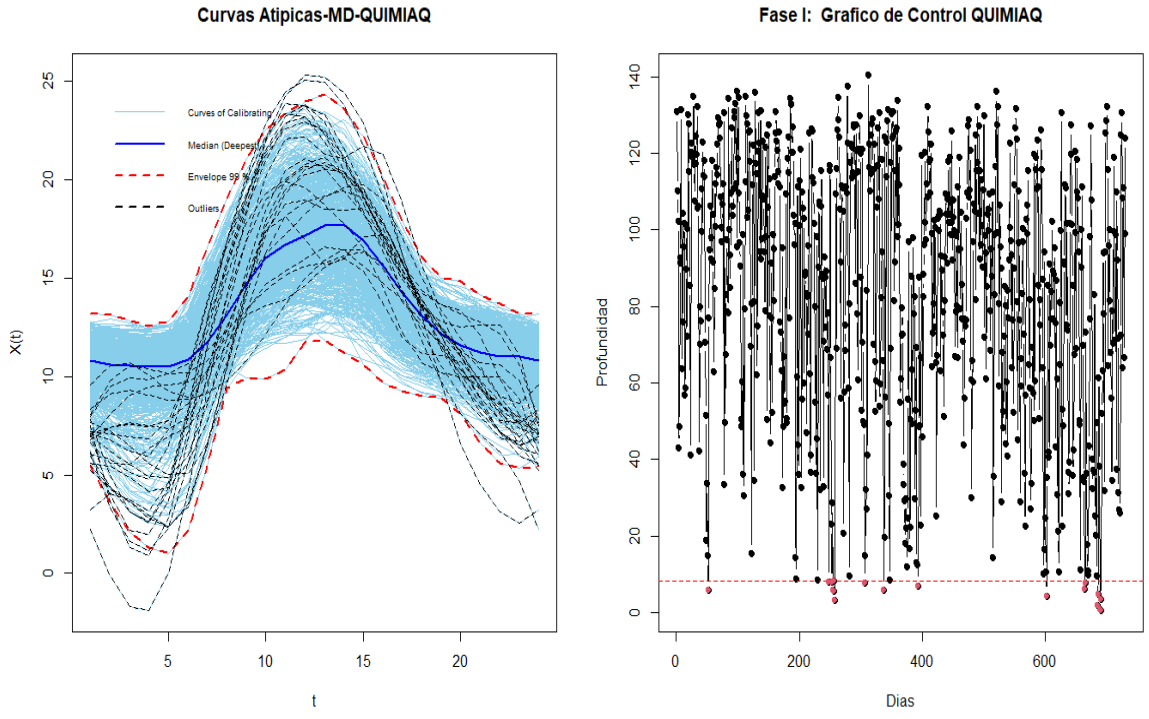


Figura 2.9: Gráfico con base a la envolvente funcional y gráfico de control funcional para la estación de QUIMIAG

n	FECHA	Profundidad	n	FECHA	Profundidad
1	2015-02-21	56,882	13	2016-10-27	76,676
2	2015-09-05	79,171	14	2016-11-16	18,084
3	2015-09-10	78,174	15	2016-11-17	47,770
4	2015-09-12	57,338	16	2016-11-20	0,702
5	2015-09-13	54,256	17	2016-11-21	0,622
6	2015-09-14	81,411	18	2016-11-22	33,158
7	2015-09-15	31,967	19	2015-07-14	69,349
8	2015-11-03	77,592	20	2015-08-18	62,144
9	2015-12-03	57,493	21	2015-10-08	71,856
10	2016-01-28	69,289	22	2015-12-13	61,128
11	2016-08-25	42,488	23	2016-11-15	79,274
12	2016-10-26	60,286			

Cuadro 2.7: Tabla de funciones atípicas de QUIMIAG

El cuantil con el valor de $8,1733$ de la estación de QUIMIAQ identifica las funciones atípicas del grafico de control funcional, con un total de 23 días atípicos se indican en la tabla $(2{,}7)$. Como observación los días $2016-11-20$, $2016-11-21$ tienen una profundidad aproximadamente de 0, por lo cual se realiza un análisis en estos días.

Estación SAN JUAN

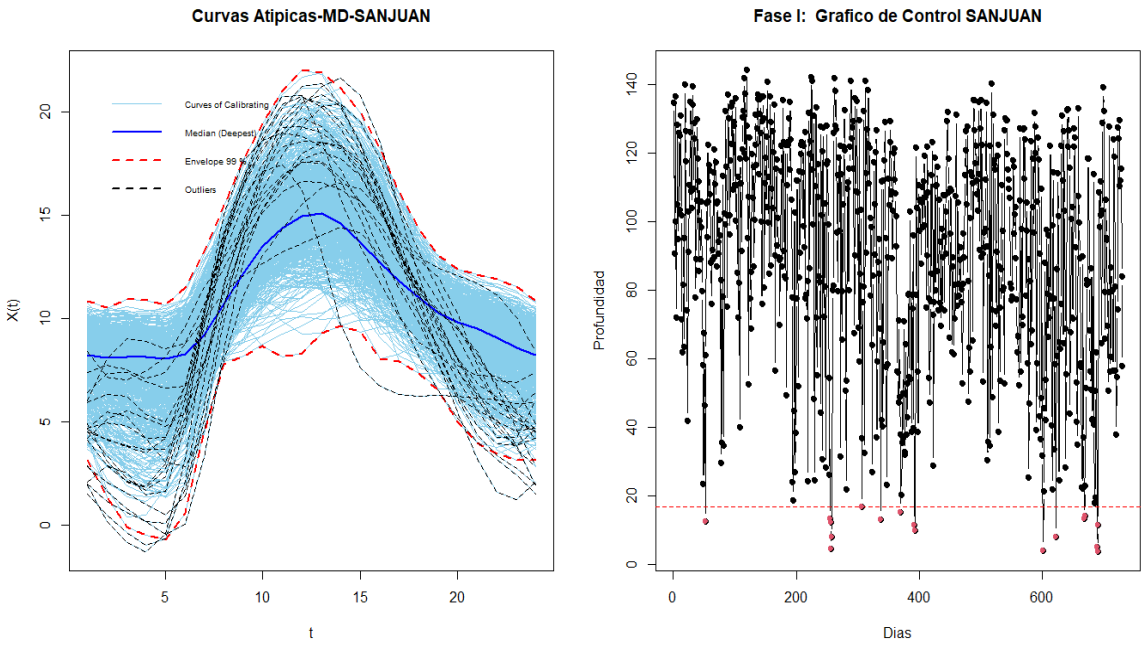


Figura 2.10: Gráfico con base a la envolvente funcional y gráfico de control funcional para la estación de SAN JUAN

n	FECHA	Profundidad	n	FECHA	Profundidad
1	2015-02-21	12,384	11	2016-08-24	3,929
2	2015-09-12	13,135	12	2016-09-14	7,888
3	2015-09-13	12,186	13	2016-10-30	13,282
4	2015-09-14	4,365	14	2016-10-31	14,134
5	2015-09-15	7,859	15	2016-11-20	4,920
6	2015-11-03	16,639	16	2016-11-21	3,755
7	2015-12-03	13,068	17	2016-11-22	11,361
8	2016-01-05	15,129	18	2015-07-14	16,200
9	2016-01-27	11,392	19	2016-11-16	15,054
10	2016-01-28	9,742			

Cuadro 2.8: Tabla de funciones atípicas de SAN JUAN

Obtuvimos el cuantil con el valor de $16,6756$ de las estación de SAN JUAN con el cual se identifica las funciones atípicas del grafico de control funcional ya que es el LCI, obteniendo un total de 19 días atípicos los cuales se indican en la tabla $(2{,}8)$. Como observación los días $2016-08-24$, $2016-11-21$ tienen una profundidad aproximadamente de $3{,}8$, por lo cual se realiza un análisis en estos días.

Estación TIXAN

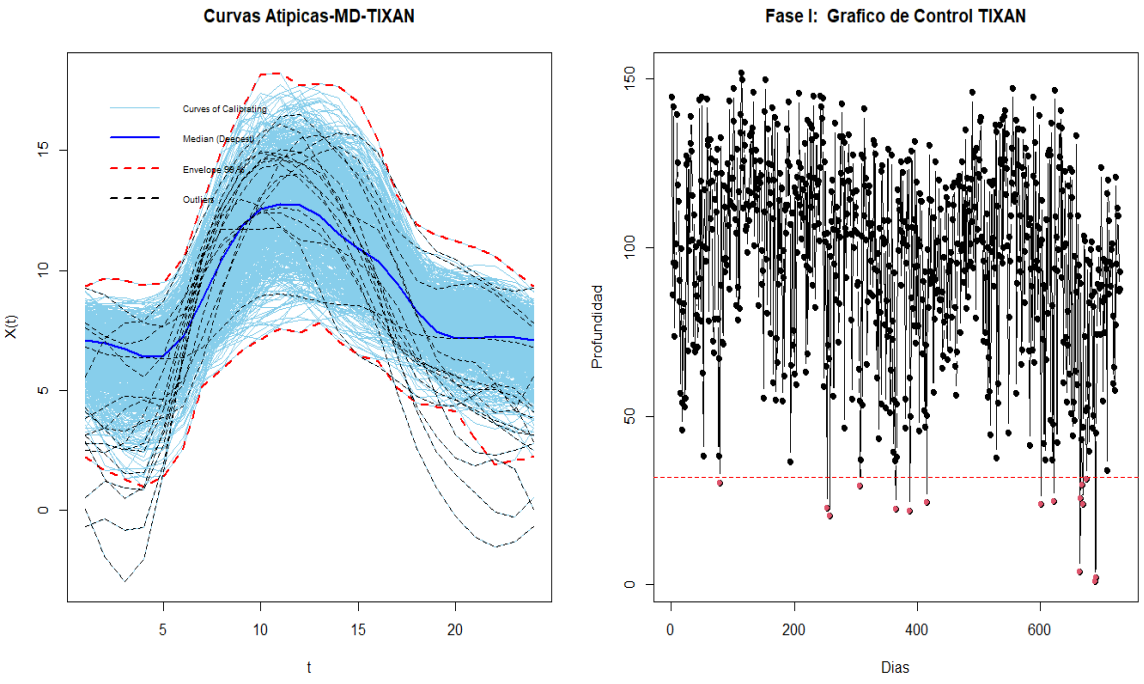


Figura 2.11: Gráfico con base a la envolvente funcional y gráfico de control funcional para la estación de TIXAN

n	FECHA	Profundidad	n	FECHA	Profundidad
1	2015-03-20	30,201	11	2016-10-27	25,464
2	2015-09-10	22,728	12	2016-10-30	29,499
3	2015-09-15	20,416	13	2016-10-31	23,906
4	2015-11-03	29,385	14	2016-11-06	31,240
5	2015-12-31	22,487	15	2016-11-20	1,099
6	2016-01-23	21,958	16	2016-11-21	2,074
7	2016-02-19	24,301			
8	2016-08-24	23,711			
9	2016-09-14	24,637			
10	2016-10-26	3,708			

Cuadro 2.9: Tabla de funciones atípicas de TIXAN

El cuantil tiene un valor de $31,83069$ de la estación de TIXAN identificando las funciones atípicas del grafico de control funcional, obteniendo un total de 16 días, los cuales se indican en la tabla $(2{,}9)$. Como observación los días $2016-11-20$, $2016-11-21$ y $2016-10-26$ tienen una profundidad aproximadamente de 2, por lo cual se realiza un análisis mas afondo en estos días.

Estación URBINA

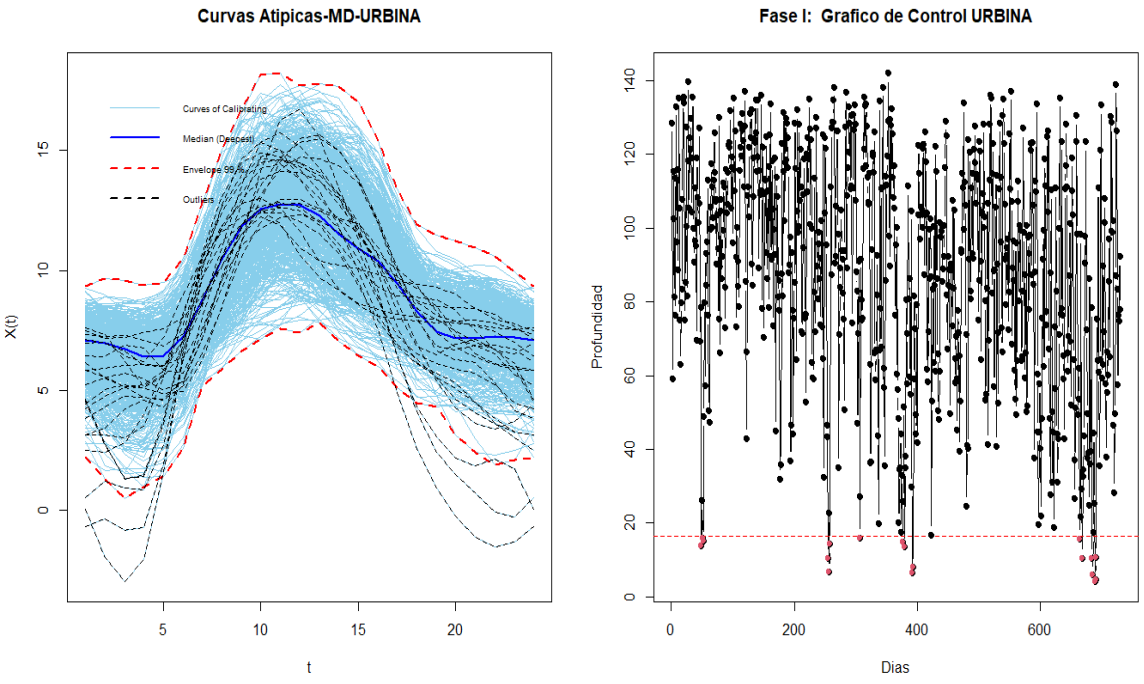


Figura 2.12: Gráfico con base a la envolvente funcional y gráfico de control funcional para la estación de URBINA

n	FECHA	Profundidad	n	FECHA	Profundidad
1	2015-02-18	13,837	12	2016-10-26	15,646
2	2015-02-20	15,818	13	2016-10-30	10,537
3	2015-02-21	15,147	14	2016-11-15	10,438
4	2015-09-12	10,486	15	2016-11-16	5,864
5	2015-09-14	6,689	16	2016-11-20	4,256
6	2015-09-15	14,215	17	2016-11-21	4,743
7	2015-11-03	15,927	18	2016-11-22	10,734
8	2016-01-11	14,942	19	2016-01-09	15,747
9	2016-01-14	13,519	20	2016-02-27	15,473
10	2016-01-27	6,494	21	2016-11-17	15,035
11	2016-01-28	8,019			

Cuadro 2.10: Tabla de funciones atípicas de URBINA

En este caso obtuvimos el cuantil con el valor de $16,4179$ de la estación de URBINA con el cual se identifica las funciones atípicas del grafico de control funcional ya que es el LCI, obteniendo un total de 21 días atípicos los cuales se indican en la tabla $(2{,}10)$. Como observación los días $2016-11-20$, $2016-11-21$ tienen una profundidad aproximadamente de 2, por lo cual se realiza un análisis más minucioso en estos días.

2.1.5. Gráfico de control Multivariante

En la búsqueda de una relación entre las estaciones se propone a realizar un análisis multivariante mediante las profundidades calculadas de cada estación, este análisis se realizara mediante la *FASE I* de un gráfico de control multivariante de rangos con la profundidad de la distancia de Mahalanobis.

Gráfico de control de rangos

Regina Y. Liu[8], propuso 3 modelos de gráficos de control: r, Q y S. Los gráficos de control que son no paramétricas recurre al estadístico $r_n(.)$, su definición está dada de la siguiente manera.

$$r_n(.) = \frac{2}{k} min(m(x_i > x), m(x_i > x)) + \frac{m(x_i = x)}{k}$$

Donde m está dado por el número de observaciones. El estadístico $r_n(.)$, explica el rango con medidas centrales o la representación de las observaciones dentro de un grupo de ellas, tomado a la mediana ya que representa el valor más central.

Profundidad de datos multivariantes

El concepto de una la profundidad de los datos recurre a que la densidad de probabilidades pueda diferenciar los puntos mas $centrales$ de los mas lejanos $perifricos$, asignando cada profundidad calculada y en $\mathbb{R}^k$ valores no negativos y los valores mas altos indicarían el centro de la distribución, las mas bajas vendrían a ser las partes externas, de esta manera se puede describir datos multivariantes, donde el dato mas profundo seria la tendencia central multivariante[10].
la función que cumple las propiedades y se usara para este análisis es la *profundidad de Mahalanobis* [9], y se define por:

$$MD = \frac{1}{[1 + (y - \mu)' \sum^{-1} (y - \mu)]}$$

Con μ vector de medias, $\sum^{-1}$ inversa de la matriz de var-cov de la distribución.
En el caso que los parámetros son no conocidos se calcula del la siguiente manera:

$$MD = \frac{1}{[1 + (y - \overline{Y})' S^{-1} (y - \overline{Y})]}$$

Las medias muéstrales viene a se $\overline{Y}$ y S la matriz de var-cov de la distribución de referencia.

Rango multivariante Para el caso de los rangos multivariantes la idea es la misma se calcula la profundidad de cada variable seguido de una ponderación de todas las observaciones de las veces que dicho dato es más bajo o igual a los datos de profundidad.
Además, el R-estadístico de clasificación de rango multivariante $r(.)$ es:

$$r(x_i) = \frac{\{m(y_j | D(y_j) \leq D(x_i), j = 1, ...n\}}{n}$$

Construcción del grafico de control multivariante

Debemos seguir los siguientes pasos para realizar el GC de rangos.

1. calculamos el vector de medias, su matriz de var-cov y las profundidades de cada dato $D(y_i)$, i=1,2,..,m.
2. obtenemos el estadístico con orden de $D(y_i)$, i=1,2,..,m. presentándose como $y_[1], y_[2], ..., y_[m]$
3. De las nuevas observaciones se calcula su profundidad, donde suponemos que $x_1, x_1, , ..$ siguen una distribución continua.
4. Procedemos a obtener el rango $r(.)$ para todos los $x_1, x_1, , ...$
5. Graficamos los estadísticos por rangos respecto a t $(tiempo)$, tomando en cuenta a limite inferior de control que viene dado por $LCI = \alpha$, con α conocida como la proporción de alarma y el limite central $LC = 0{,}5$.

Lui [8] , llego a la demostración de que el estadístico proviene de una distribución uniforme $[0, 1]$.

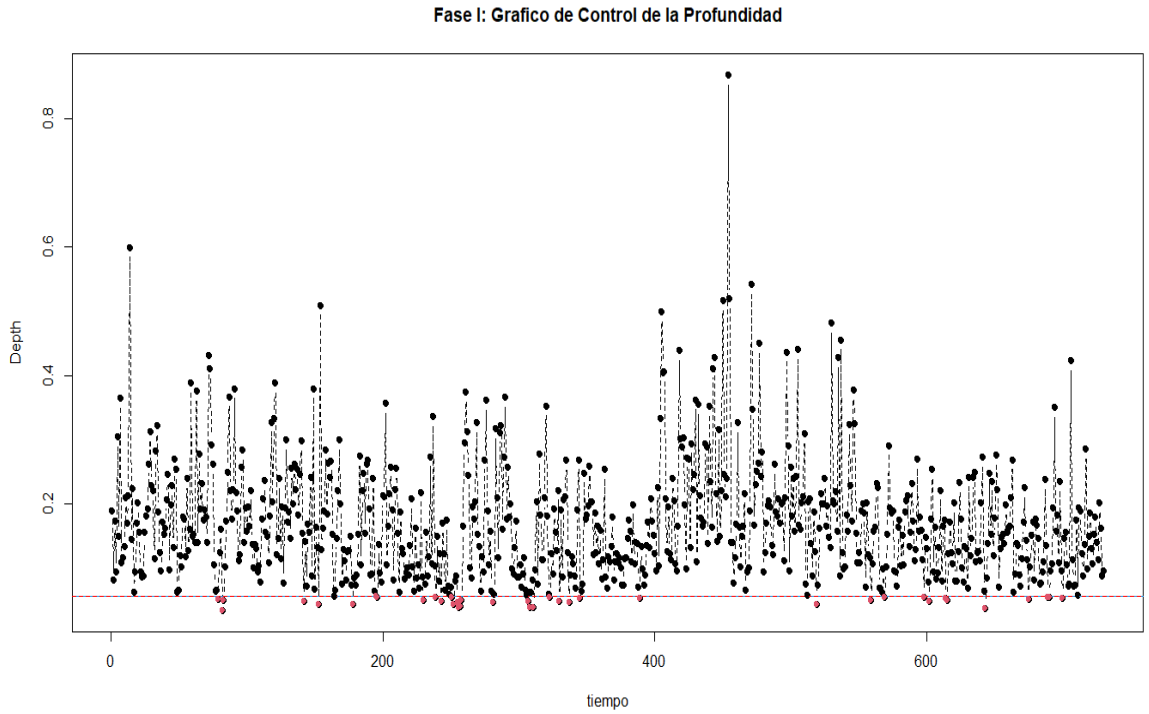


Figura 2.13: Gráfico de control para las 7 estaciones climatológicas

En la fase I de la figura (2.13), investigaremos la existencia de atípicos en los datos históricos, obtenidos de las profundidades de cada estación con 730 dias por los años de 2015-1016 y las 7 estaciones, este análisis se relizará mediante la definición de la profundidad de Mahalanobis para encontrar el cuantil o LCI y detectar los atípicos en el gráfico de control, como se muestra en la siguiente tabla.

n	FECHA	Prof.	n	FECHA	Prof.	n	FECHA	Prof.
1	2015-03-20	0,050	14	2016-12-01	0,053	27	2015-06-02	0,042
2	2015-07-14	0,055	15	2015-03-24	0,049	28	2015-08-31	0,047
3	2015-09-09	0,042	16	2015-08-18	0,049	29	2015-09-15	0,049
4	2015-11-03	0,047	17	2015-09-13	0,037	30	2015-11-26	0,048
5	2015-12-11	0,052	18	2015-11-07	0,038	31	2016-07-23	0,054
6	2016-08-25	0,048	19	2016-06-03	0,043	32	2016-11-06	0,051
7	2016-11-21	0,054	20	2016-09-07	0,049	33	2015-06-27	0,042
8	2015-03-23	0,034	21	2015-05-22	0,048	34	2015-09-08	0,054
9	2015-07-15	0,054	22	2015-08-27	0,053	35	2015-10-08	0,046
10	2015-09-12	0,045	23	2015-09-14	0,039	36	2015-12-03	0,046
11	2015-11-04	0,038	24	2015-11-19	0,054	37	2016-08-21	0,054
12	2016-01-24	0,052	25	2016-07-13	0,049	38	2016-11-20	0,053
13	2016-09-06	0,051	26	2016-10-05	0,036			

Cuadro 2.11: Tabla de funciones atípicas para en análisis compuesto de las 7 estaciones.

En este análisis tenemos 38 datos atípicos un número mayor a comparación del análisis individual de cada estación, esto se debe a la diferencia que existe entre las estaciones de Espoch, Quimiag con las estaciones de Urbina, Atillo, y Tixan con una diferencia aproximadamente de 4 a 5 grados. Sin embargo, puede ser de mucha utilidad ya que pueden indicar una curva típica de toda la región de Chimborazo para eso debemos consolidar con la información de las estaciones individuales.

Capítulo 3

Resultados, conclusiones y recomendaciones

3.1. Resultados

Para la primera parte se presenta el resultado del uso de la validación cruzada, empleando del paquete fda desarrollada por (Ramsay y sylverman)[14], y el criterio de validación cruzada GCV(validación cruzada generalizada) buscando para las bases de $Fourier$ el numero k óptimo y el valor de la rugosidad (λ) de las funciones base para así tener una buena representación de los datos en cada una de las 7 estaciones. Se propuso 8 posibles valores para el numero de funciones base (k) y λ que se presentan por pares en la tabla $(2{,}2)$ y como resultados se presenta la siguiente tabla(3.2):

Estaciones							
	Alao	Atillo	Espoch	Quimiaq	San Juan	Tixan	Urbina
K optimo	15	11	15	15	15	15	15
λ optimo	0.25	0.25	0.25	0.25	0.25	0.25	0.125

Cuadro 3.1: k y λ óptimo de las 7 estaciones

Notamos que el valor del número de bases de k y el valor de la rugosidad λ es casi igual en todas las estaciones esto se debe a que las estaciones tienen un comportamiento muy similar a datos periódicos o cíclicos en la temperatura ambiente.
Como segunda parte, se procedió a transformar los datos discretos a da-

tos funcionales con el k y λ óptimos, obteniendo un total de 730 curvas en un intervalo de 24 horas por estación notando claramente la presencia de curvas atípicas en las estaciones para tratar este problema se propuso realizar la *Fase I* de un gráficos de control funcionales para monitorizar las funciones por estación, la cual se le realizo mediante el uso de la profundidad modal debido a que (Miguel Flores)[15], propone a este análisis como el mejor debido a su efectividad, una vez calculada la profundidad de cada estación, para que se determine las curvas con la menor profundidad funcional se utiliza el paquete $(fda.usc)$ y la función $outliers.deph.trim$, indicando el valor de la profundidad $Modal$, seguido de la detección de las ´funciones atípicas, que realiza mediante Bootstrap seguido del cuantil que se obtuvo del paquete (qcr) obteniendo lo deseado. una ves obtenida los datos atípico funcionales de las 7 estaciones se empieza a realizar un estudio minucioso entre ellas y su relación en los datos atípicos. Realizamos una comparación entre ellas:

Frecuencia Estaciones					
n	FECHA	Frecuencia	n	FECHA	Frecuencia
1	2015-02-21	6	11	2016-10-26	5
2	2015-07-14	4	12	2016-10-27	4
3	2015-09-12	4	13	2016-10-30	6
4	2015-09-14	4	14	2016-10-31	4
5	2015-09-15	6	15	2016-11-16	6
6	2015-11-03	6	16	2016-11-17	4
7	2016-01-27	4	17	2016-11-20	7
8	2016-01-28	6	18	2016-11-21	7
9	2016-08-24	5			
10	2016-09-14	5			

Cuadro 3.2: Tabla de funciones atípicas de las 7 estaciones con frecuencia mayor 3

La tabla (3.2), es de suma importancia ya que representa la frecuencia de un dia atípico en las 7 estaciones y además se encuentran estas 18 curvas en la tabla (2.11) las cual representan las curvas generales compuestas por las 7 estaciones. Esta información nos indicaría que hubo algún tipo de fenómeno meteorológico como vientos muy fuertes, existencia de fríos helados que entro de algún lado, calores extremos o algún suceso anómalo que se debería investigar la razón, esta información no puede ser una falla del sistema o del equipo ya que esta presente en mas

de 3 estaciones. Ahora tomemos como ejemplo los días "$2015-02-20$" y "$2015-02-21$" estas curvas son atípicas y tiene frecuencia de 7 es decir, que esta presente en las 7 estación esto nos indica que toda la región tuvo un fenómeno meteorológico, para consolidar esta información procesemos a consultar este dia en "https://es.weatherspark.com/" , esta página detalla informes meteorológico del clima típico de cualquier lugar específico. Obteniendo la información de la temperatura ambiente de Chimborazo de los días "$2015-02-20$" y "$2015-02-21$" representado en el siguiente gráfico.

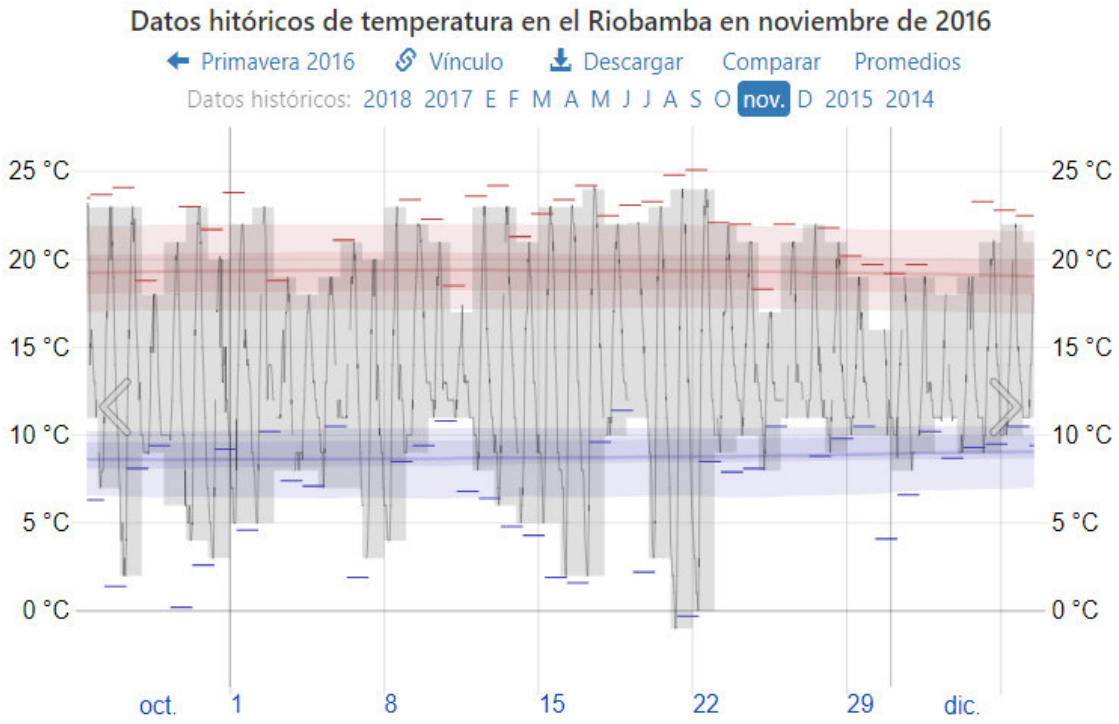


Figura 3.1: Gráfico del comportamiento de la temperatura Amb. de Chimborazo del mes de noviembrel del 2016

Notamos claramente en la grafica (3.1) que los días 20 y 21 hubo la presencia de un frio extremo que registro alrededor de -1 grado centígrado en las horas de la madrugada en los 2 días lo cual no es común en esta región esto fue a causa de la presencia de un frio extremo que ingreso por la parte norte de Chimborazo.
Lo mismo sucede con los otros datos atípicos de la tabla (3.2) su causa

se debe a algún fenómeno natural.
De esta manera consolidamos la efectividad del análisis y monitoreo de la Fase I del grafico de control funcional y multivariante, el cual fue aplicado a los datos de series meteorológicas proporcionados por el (GEAA) de la provincia de Chimborazo.

3.2. Conclusiones y recomendaciones

En este proyecto se busca la aplicación del análisis de datos funcionales y la *Fase I* de los gráficos de control, la combinación de estos métodos da como resultado el grafico de control funcional, monitorizando la presencia de curvas anómalas de un conjunto de datos funcionales y de este modo definir si existe una anomalía es a causa de un fallo del sistema o el entorno. Esté método se lo puede considerar relativamente nuevo y va tomando una gran acogída para el análisis de grandes volúmenes de datos con una dimensión alta.
Mediante el uso de diferentes métodos para el análisis estadísticos por horas de la temperatura ambiente de 7 estaciones de la provincia de Chimborazo durante los años de 2015 y 2016. Una propuesta fue la aplicación de transformar los datos discretos a funcionales esto se realizó mediante el uso del análisis de datos funcionales brindando ventajas como como disminuir la dimensión, tener más información y herramientas para el tratamiento de variables climatológicas medidas en el tiempo, este tipo de datos tienen periocidad, por lo cual se utilizó las bases de $Fourier$ asociando los datos escalares a una curva o función en el intervalo de 24 horas, se realizó con el criterio de suavisamiento con el k optimo, incluyendo la penalización por rugosidad obteniendo los datos estimados ajustados que describen los datos de mejor manera mediante (GCV) validación cruzada generalizada.
Ahora para detectar los datos atípico se empleo el concepto de medida de profundidad funcional el cual ordena el conjunto de función según que tan alejada del función central.

Para realizar la Fase I del gráfico de control funcional nos basamos en el método (Febrero)[2], propone un método de remuestreo para la detección de datos atípicos, tomando el concepto de profundidad funcional el

cual ordena el conjunto de función según la distancia de la función central, el método aplicado en este proyecto es la profundidad modal. Así, se obtiene el gráfico de control tipo R, el cual calcula de manera directa el LCI con la suposición del que el estadístico proviene de una distribución uniforme [0,1] [8]
Este método se aplicó a los datos meteorológico funcionales obteniendo un resultado eficiente para la clasificación de los datos como atípicos monitoreados en el tiempo t.
De esta manera se concluye que el gráfico de control funcional es una de las mejores alternativas en el caso de que exista o no normalidad, siempre y cuando las observaciones que se desee monitorear se pueda transformar a datos funcionales. Mejorando la eficacia de diferentes herramienta clásicas para el análisis de grandes volúmenes de datos.

3.2.1. Recomendaciones

El estudio del gráfico de control funcional, se lo puedes extender al análisis funcional mediante otros tipos de profundidades y estimadores los cuales fueron mencionados en el proyecto, ya que el estudio de datos atípicos está presente en todo tipo de información y pueda que para diferentes casos como la profundidad modal no sea la más indicada, se debería llevar a una comparación de resultados. Además, el análisis de datos meteorológicos tiene un campo muy grande e importante en la sociedad y economía de una región, entonces se debería analizar las diferentes variables como humedad, precipitación, entre otros.

Capítulo A

Anexos

A.1. Apendice A

```
library(readr)
library(dplyr)
library(tidyverse)
GEAA_s <- read_csv("impu_GEAA.csv", col_types = cols(HORAS = col_time(
    format = "%H:%M:%S")))

####POR HORAS ALAO
dfg <- data.frame( GEAA_s[,1:3])
ALAOH<- reshape(dfg, direction = "wide", idvar = "FECHA", timevar = "
    HORAS")

####POR HORAS ATILLO
dfg <- data.frame(select(GEAA_s,FECHA,HORAS, ATILLO))
ATILLOH<- reshape(dfg, direction = "wide", idvar = "FECHA", timevar = "
    HORAS")

####POR HORAS ESPOCH
dfg <- data.frame(select(GEAA_s, FECHA, HORAS, ESPOCH))
ESPOCH<- reshape(dfg, direction = "wide", idvar = "FECHA", timevar = "
    HORAS")

####POR HORAS QUIMIAQ
dfg <- data.frame(select(GEAA_s, FECHA, HORAS, QUIMIAQ))
```

```
QUIMIAQH<- reshape(dfg, direction = "wide", idvar = "FECHA", timevar =
    "HORAS")


####POR HORAS SANJUAN
dfg <- data.frame(select(GEAA_s, FECHA, HORAS, SANJUAN))
SANJUANH<- reshape(dfg, direction = "wide", idvar = "FECHA", timevar =
    "HORAS")


####POR HORAS TIXAN
dfg <- data.frame(select(GEAA_s, FECHA, HORAS, TIXAN))
TIXANH<- reshape(dfg, direction = "wide", idvar = "FECHA", timevar = "
    HORAS")


####POR HORAS URBINA
dfg <- data.frame(select(GEAA_s, FECHA, HORAS, URBINA))
URBINAH<- reshape(dfg, direction = "wide", idvar = "FECHA", timevar = "
    HORAS")

#########################################################
########    GAFICOS DISCRETOS   ##########################
#########################################################
library(ggplot2)

df <- data.frame(x = seq_along(ALAOH[,1]),
                 ALAOH)
df <- df[,-4]
df <- df[,-2]
df <- df[,-1]
# Formato long
df <- melt(df, id.vars = "x")
ggplot(df, aes(x = x, y = value, color = variable)) +
  geom_point(size=0.7)

#########################################################
########    DATOS FUNCIONALES POR HORAS   ###############
#########################################################

library(reshape2)
library(fda.usc)
library(fda)
```

```
library(ftsa)
library(tictoc)
library(fields)
#####################
rownames(ALAOH)<-ALAOH[,1]
ALAOH<-ALAOH[,-1]
argvals <- 1:24

ALAOHf<- fdata(ALAOH, argvals=argvals, names= list(main='ALAOH',
                                                   xlab='HORAS',ylab='
                                                       TEMPERATURA␣
                                                       AMBIENTE␣ALAO'))

rownames(ATILLOH)<-ATILLOH[,1]
ATILLOH<-ATILLOH[,-1]
ATILLOHf<- fdata(ATILLOH, argvals=argvals, names= list(main='ATILLOH',
                                                       xlab='HORAS',ylab
                                                           ='TEMPERATURA
                                                           ␣AMBIENTE␣
                                                           ATILLO'))

rownames(ESPOCH)<-ESPOCH[,1]
ESPOCH<-ESPOCH[,-1]
ESPOCHf<- fdata(ESPOCH, argvals=argvals, names= list(main='ESPOCH',
                                                     xlab='HORAS',ylab='
                                                         TEMPERATURA␣
                                                         AMBIENTE␣'))

rownames(QUIMIAQH)<-QUIMIAQH[,1]
QUIMIAQH<-QUIMIAQH[,-1]
QUIMIAQHf<- fdata(QUIMIAQH, argvals=argvals, names= list(main='QUIMIAQ',
                                                         xlab='HORAS',
                                                             ylab='
                                                             TEMPERATURA
                                                             ␣AMBIENTE␣'
                                                             ))

rownames(SANJUANH)<-SANJUANH[,1]
SANJUANH<-SANJUANH[,-1]
SANJUANHf<- fdata(SANJUANH, argvals=argvals, names= list(main='SANJUAN',
                                                         xlab='HORAS',
                                                             ylab='
                                                             TEMPERATURA
```

```
                                                          AMBIENTE '
                                                          ))


rownames(TIXANH)<-TIXANH[,1]
TIXANH<-TIXANH[,-1]
TIXANHf<- fdata(TIXANH,argvals=argvals, names= list(main='TIXAN',
                                                    xlab='HORAS',ylab='
                                                      TEMPERATURA
                                                      AMBIENTE '))

rownames(URBINAH)<-URBINAH[,1]
URBINAH<-URBINAH[,-1]
URBINAHf<- fdata(URBINAH,argvals=argvals, names= list(main='URBINA',
                                                      xlab='HORAS',ylab
                                                        ='TEMPERATURA
                                                        AMBIENTE '))

############################################################
############ Validaci n CruZada ############################
############################################################
nb<-floor(seq(5,29,len=8))
l<-2^(-5:10)

opt_al<-optim.basis(ALAOHf, lambda = l, numbasis = nb,
                    type.CV = GCV.S, type.basis = 'fourier')

opt_at<-optim.basis(ATILLOHf, lambda = l, numbasis = nb,
                        type.CV = GCV.S, type.basis = 'fourier')

opt_es<-optim.basis(ESPOCHf, lambda = l, numbasis = nb,
                        type.CV = GCV.S, type.basis = 'fourier')

opt_qm<-optim.basis(QUIMIAQHf, lambda = l, numbasis = nb,
                    type.CV = GCV.S, type.basis = 'fourier')

opt_sj<-optim.basis(SANJUANHf, lambda = l, numbasis = nb,
                    type.CV = GCV.S, type.basis = 'fourier')

opt_tx<-optim.basis(TIXANHf, lambda = l, numbasis = nb,
                    type.CV = GCV.S, type.basis = 'fourier')

opt_ur<-optim.basis(URBINAHf, lambda = l, numbasis = nb,
```

```
                    type.CV = GCV.S, type.basis = 'fourier')

## Data, k, lamda
k <- opt_ur$numbasis.opt
lambda <- opt_ur$lambda.opt

## Grafico de Cada Estacion con el K y lambda optimo
par(mfrow=c(1,2))
dataconf<-ALAOHf
dataconf$data <- ALAOHf$data%*%opt_al$S.opt
plot(datacon, main="Datos no suavizados ALAO", ylab= "TEMPERATURA
    AMBIENTE" )
plot(dataconf, main="Datos suavizados ALAO", ylab= "TEMPERATURA
    AMBIENTE")

dataconf<-ATILLOHf
dataconf$data <- ATILLOHf$data%*%opt_al$S.opt
plot(datacon, main="Datos no suavizados ATILLO" )
plot(dataconf, main="Datos suavizados ATILLO", ylab= "TEMPERATURA
    AMBIENTE")

dataconf<-ESPOCHf
dataconf$data <- ESPOCHf$data%*%opt_al$S.opt
plot(datacon, main="Datos no suavizados ESPOCH" )
plot(dataconf, main="Datos suavizados ESPOCH", ylab= "TEMPERATURA
    AMBIENTE")

dataconf<-QUIMIAQHf
dataconf$data <- QUIMIAQHf$data%*%opt_al$S.opt
plot(datacon, main="Datos no suavizados QUIMIAQ" )
plot(dataconf, main="Datos suavizados QUIMIAQ", ylab= "TEMPERATURA
    AMBIENTE")

dataconf<-SANJUANHf
dataconf$data <- SANJUANHf$data%*%opt_al$S.opt
plot(datacon, main="Datos no suavizados SAN JUAN" )
plot(dataconf, main="Datos suavizados SAN JUAN", ylab= "TEMPERATURA
    AMBIENTE")

dataconf<-TIXANHf
dataconf$data <- TIXANHf$data%*%opt_al$S.opt
plot(datacon, main="Datos no suavizados TIXAN" )
```

```
plot(dataconf, main="Datos␣suavizados␣TIXAN", ylab= "TEMPERATURA␣
    AMBIENTE")

dataconf<-URBINAHf
dataconf$data <- URBINAHf$data%*%opt_al$S.opt
plot(datacon, main="Datos␣no␣suavizados␣URBINA" )
plot(dataconf, main="Datos␣suavizados␣URBINA", ylab= "TEMPERATURA␣
    AMBIENTE")

############################################################
#############MEDIA y DERIVADAS ##########################
############################################################
data <-dataconf$data
tt = dataconf$argvals
rtt = dataconf$rangeval
data.name = dataconf$names
names <- list(main=data.name)
x <- fdata(data,tt,rtt,names)

al<- func.mean(x)
at<- func.mean(x)
es<- func.mean(x)
qu<- func.mean(x)
sj<- func.mean(x)
tx<- func.mean(x)
ur<- func.mean(x)
alm<- func.var(x)
atm<- func.var(x)
esm<- func.var(x)
qum<- func.var(x)
sjm<- func.var(x)
txm<- func.var(x)
urm<- func.var(x)
va<- rbind(al$data,at$data,es$data,qu$data,sjm$data,txm$data,urm$data)
va<-data.frame(va)

URder1<- fdata.deriv(dataconf, nderiv = 1, nbasis = k,
                     lambda=lambda, method = 'fourier')
plot(URder1, main = "URBINA␣1era␣Derivada", xaxt = "n", ylab= "
    TEMPERATURA␣AMBIENTE")
axis(1, at= seq(0,24, by=4), las =2)
#########2dA DERIVADA###########
```

```
URder2<- fdata.deriv(dataconf, nderiv = 2, nbasis = k,
                     lambda=lambda, method = 'fourier')
plot(URder2, main = "Estaciones 2da Derivada", xaxt = "n", ylab= "
    TEMPERATURA AMBIENTE")
axis(1, at= seq(0,24, by=4), las =2)

############################################################
############# Profundidad Modal ############################
############################################################
par(mfrow=c(1,2))

out_al=depth.mode(ALAOHf,trim=0.15,draw=TRUE)
view(out_al$ltrim)
out_at=depth.mode(ATILLOHf,trim=0.15,draw=TRUE)
out_es=depth.mode(ESPOCHf,trim=0.15,draw=TRUE)
out_qm=depth.mode(QUIMIAQHf,trim=0.15,draw=TRUE)
out_sj=depth.mode(SANJUANHf,trim=0.15,draw=TRUE)
out_tx=depth.mode(TIXANHf,trim=0.15,draw=TRUE)
out_ur=depth.mode(URBINAHf,trim=0.15,draw=TRUE)

#########################################################################
############ Bootstrap y Grafico por Estacion mediante QCR ##########
#########################################################################
library(lubridate)
ns =  0.01
smo =0.05
trim = 0.015
nb=200
data <-dataconf$data
tt = dataconf$argvals
rtt = dataconf$rangeval
data.name = dataconf$names
names <- list(main=data.name)
x <- fdata(data,tt,rtt,names)
outdp_ur <- outliers.depth.trim(x,nb = nb, ns =ns,
                                smo = smo, dfunc = depth.mode ,trim = trim
                                    )
Dep <- outdp_ala$Dep
ind_al<-unclass(as.Date(outdp_ala$outliers)- as.Date("2015-1-1"))
if(length(ind_al)>0) fdaenv_al <- x[-ind_al,] else fdaenv_al <- x
LCL <- apply(fdaenv_al$data, 2,min)
fmin <- fdata(LCL, tt, rtt)
UCL <- apply(fdaenv_al$data, 2,max)
```

```
fmax <- fdata(UCL, tt, rtt)
med <- func.med.mode(x,trim=trim)
draw.control = NULL
plot=TRUE
draw.control = list(col = c("skyblue", "blue", "red"),
                    lty = c(1, 1, 2), lwd = c(1, 2, 2))
plot(x, col= "skyblue", main= 'Curvas␣Atipicas-MD-ALAO')
lines(fmin, lwd = draw.control$lwd[3],
      lty = draw.control$lty[3], col = draw.control$col[3])
lines(fmax, lwd = draw.control$lwd[3],
      lty = draw.control$lty[3], col = draw.control$col[3])
lines(med, lwd = draw.control$lwd[2],
      lty = draw.control$lty[2],
      col = draw.control$col[2])
if(length(ind_al)>0)
  lines(x[ind_al,], lwd = 1,
        lty = 2, col = 1)
legend(x = min(tt), y = 0.99 * max(data), bty = "n",
       legend = c("Curves␣of␣Calibrating",
                  "Median␣(Deepest)", paste("Envelope",(1-ns)*100,"%"),
                  "Outliers"),
       lty = c(1,1,2,2),
       lwd = c(draw.control$lwd,draw.control$lwd[2],2),
       col = c(draw.control$col,"black"), cex = 0.6,
       box.col = 0)
######## Grafico de Control##########

plot(Dep, type="b",pch=16, main = "Fase␣I:␣␣Grafico␣de␣Control␣ALAO",
     ylim=c(min(Dep,outdp_ala$quantile),max(Dep)),xlab="Dias",ylab="
        Profundidad")
abline(h = outdp_ala$quantile, lty = 2, col = "red")
out<-(Dep<=outdp_ala$quantile)+1
points(x=1:n, Dep, col=c(out,"red"), lwd=5, cex = 0.2)

#####################################################
##############unimos las profundidades ############
#####################################################
library(dplyr)
deppp<- cbind(dep_al,dep_at,dep_es,dep_qm,dep_sj,dep_tx,dep_ur)

### Grafico de control multivariante
library(xtable)
library(TSA)
```

```
nboot<-500
alpha<- 0.05
###aplicamos bootstrap
btdep<- bootapl(data=prof,nboot=nboot,func=mdepth.MhD)
#limite
lcboot<- Lim(btdep, alpha=alpha)$LCI
mvShapiro.Test(prof)
bootapl
depth.x1<- mdepth.MhD(prof)$dep
plot(depth.x1, lwd=0.5, lty=2, type="b",pch=16,xlab="tiempo",
     ylab="Depth", main = "Fase␣I:␣Grafico␣de␣Control␣de␣la␣Profundidad
         ")
abline(h=lcboot, lty = 2, col = "red")
out<-(depth.x1<=lcboot)+1
points(x=1:n, depth.x1, col=out,lwd=5, cex = 0.2)
#####################################################
##############Funciones ############
#####################################################
bootapl<-function(data, nboot, h=0.001, data.ref=NULL, func=func){
  n<-nrow(data)
  rs<-matrix(NA, ncol = n, nrow = nboot)
  rss<-numeric(n)
  for (k in 1:nboot) {
    if(!is.null(data.ref)){
      dep.ref<-func(data)$dep
      dep<-func(data.ref,data)$dep
      rank<-sapply(dep, function(x) mean(dep.ref <= x))
    }
    else{
      dep.ref<-func(data)$dep
      dep<-func(data)$dep
      rank<-sapply(dep, function(x) mean(dep.ref <= x))
    }
    rss<-sample(dep,n,replace = T)
    rs[k,]<-rss+h*rnorm(n,0,1)
  }
  return(rs)
}

Lim<- function(data, alpha=0.05,LCI=NULL){
  if(is.null(LCI)) LCI <- quantile(data, probs=alpha)
  rp<- mean(data<=LCI)
  result<- list(LCI=LCI, RP=rp)
```

```
}
```

Referencias bibliográficas

[1] M. Colosimo, B. M. y Pacella. *A comparison study of control charts for statistical monitoring of functional data*. International Journal of Production Research, 2010.

[2] Galeano P. y González-Manteiga Febrero, M. *Outlier Detection in Functional Data by Depth Measures, With Application to Identify Abnormal NOx Levesls*. Environmetrics, 2007.

[3] Galeano P. y González-Manteiga Febrero, M. *Outlier Detection in Functional Data by Depth Measures, With Application to Identify Abnormal NOx Levesls*. Environmetrics, 2008.

[4] M. O. Febrero-Bande, M. y de la fuente. *Statistical computing in funtional* . 2012.

[5] P. Ferraty, J. y Vieu. *Noparametric Functional*. Springer Science, 2006.

[6] G. Fraiman, R. y Muniz. *Statistical computing in funtional* . Journal of Statistical Software, 2012.

[7] Teresa Delgado Tejada José F. Vilar Barrio. *Control estadístico de los procesos (SPC)*. Fundación CONFEMETA, 2005.

[8] Regina Y. Liu. *Control charts for multivariate processes*. the American Statistical Association, 1995.

[9] P.C Mahalanobis. *On the generaised distance in statistics*. Volume 2, 1936).

[10] Mehmet Mert. *Nonparametric control charts based on mahalanobis depth. Hacettepe Journal of Mathematics and Statistics*. Hacettepe Journal of Mathematics and Statistics, 2004).

[11] D. C. Montgomery. *Introduction To statistical Quality Control*. Johm Wiley Sons, 1997.

[12] D. C. Montgomery. *Diseño y Análisis de Experimentos*. Johm Wiley Sons, 2007.

[13] B. Ramsay, J. y Silverman. *Funtional Data Analysis*. Springer Science, 1997.

[14] B. Ramsay, J. y Silverman. *Funtional Data Analysis*. Springer Science, 2005.

[15] Miguel Flores Sánchez. *Nuevas aportaciones del análisis de datos funcionales en el control estadístico de procesos*. Universidade da Coruña, 2019.

[16] Humberto Gutierres Pulido. y Román de la Vara Salazar . *Control Estadístico de la calidad 6 Sigma*. McGRAW-HILL, 2009.