ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

"ANÁLISIS GEOESTADÍSTICO DE DATOS FUNCIONALES DE TEMPERATURA DEL AIRE EN LA PROVINCIA DE CHIMBORAZO"

Trabajo de Titulación

Tipo: Proyecto de Investigación

Presentado para optar el grado académico de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

AUTORA: MARISOL CAROLINA CHECA GAMARRA

DIRECTORA: ING. AMALIA ISABEL ESCUDERO VILLA

Riobamba - Ecuador

2020

Yo, MARISOL CAROLINA CHECA GAMARRA, declaro que el presente trabajo de titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autora asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación. El patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 10 de febrero de 2020

Marisol Carolina Checa Gamarra
060451172-5

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

El Tribunal del trabajo de titulación certifica que: El trabajo de investigación: Tipo: Proyecto de Investigación, **"ANÁLISIS GEOESTADÍSTICO DE DATOS FUNCIONALES DE TEMPERATURA DEL AIRE EN LA PROVINCIA DE CHIMBORAZO",** realizado por la señorita MARISOL CAROLINA CHECA GAMARRA, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Titulación, el mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal Autoriza su presentación.

	FIRMA	**FECHA**
Dr. Celso Guillermo Recalde Moreno **PRESIDENTE DEL TRIBUNAL**		2020-02-10
Ing. Amalia Isabel Escudero Villa **DIRECTORA DEL TRABAJO DE TITULACIÓN**		2020-02-10
Ing. Carlos Rolando Rosero Erazo **MIEMBRO DEL TRIBUNAL**		2020-02-10

DEDICATORIA

A mis padres, Fany y Carlos por su cariño, confianza y apoyo incondicional, en especial a mi madre quien día a día me motiva para salir adelante y cumplir una a una las metas que me he propuesto.

A mi hermanita Kristel, que con su inocencia y ocurrencias siempre logra sacarme una sonrisa, por estar junto a mí y compartir los más bellos momentos de mi vida.

Marisol

AGRADECIMIENTO

Mi más sincero agradecimiento primeramente a Dios, por la vida, la familia, por haberme guiado y darme la fortaleza para seguir adelante y nunca darme por vencida.

A la Escuela Superior Politécnica de Chimborazo, por abrirme sus puertas y darme la oportunidad de prepararme profesionalmente.

A mi tutora, Ing. Isabel Escudero V. por la confianza puesta en mí durante la realización de este proyecto de investigación, por compartir sus conocimientos y por todo el tiempo y paciencia invertidos en el mismo.

Al Ing. Carlos Rosero E. por todas las recomendaciones y sugerencias dadas para mejorar la investigación realizada.

A cada uno de mis maestros y a todas aquellas personas que de alguna forma han contribuido con sus conocimientos a mi formación como Ingeniera en Estadística Informática.

Al Grupo de Investigación de Energías Alternativas y Ambiente, por darme la oportunidad de realizar mis prácticas preprofesionales y facilitar toda la información necesaria para llevar a cabo este trabajo de investigación.

Finalmente, de manera especial y de corazón deseo agradecer a mi familia por toda la paciencia, consejos y apoyo en los momentos de dificultad.

Marisol

TABLA DE CONTENIDO

ÍNDICE DE TABLAS ..... **xi**

ÍNDICE DE FIGURAS ..... **xii**

ÍNDICE DE GRÁFICOS ..... **xiii**

ÍNDICE DE ANEXOS ..... **xv**

RESUMEN ..... **xvi**

SUMMARY ..... **xvii**

INTRODUCCIÓN **1**

CAPITULO I

1 MARCO TEÓRICO REFERENCIAL **10**

1.1 Meteorología **10**

1.1.1 Una tradición empírica. Climatología ***10***

1.1.2 Una tradición teórica. Física de la atmosfera ***11***

1.1.3 Una tradición práctica. Predicción del tiempo ***11***

1.2 Variables Meteorológicas **12**

1.2.1.1 Temperatura del aire *12*

1.3 Estación Meteorológica **12**

1.3.1 Dispositivos para medir la temperatura del aire ***13***

1.3.2 Especificaciones funcionales de las estaciones meteorológicas ***14***

1.4 Relleno de valores faltantes **15**

1.4.1 Dato faltante ***15***

1.4.2 Imputación Múltiple ***15***

1.4.3 MICE (Multivariate Imputation by Chained Equations) ***16***

1.4.3.1 Paquete MICE en R *17*

1.4.3.2 Validación del modelo de imputación *17*

1.5 Estadística espacial **18**

1.5.1 Áreas de la estadística espacial ***18***

1.5.2 Tipos de datos espaciales ***18***

1.6 Geoestadística **19**

1.6.1 Variable regionalizada ***20***

1.6.2 Isotropía ***20***

1.6.3 Estacionariedad ***20***

1.6.3.1 Estacionariedad de Segundo Orden *21*

1.6.3.2 Estacionariedad Débil o Intrínseca *21*

1.6.4 Análisis estructural ***22***

1.6.4.1 Semivariograma *23*

1.6.5 Kriging: predicción e interpolación ***26***

1.6.5.1 Kriging Ordinario (KO) *26*

1.6.5.2 Kriging Universal (KU) *27*

1.7 Análisis de Datos Funcionales **28**

1.7.1 Definiciones ***29***

1.7.2 Representación en funciones de una base ***30***

1.7.2.1 Bases de Fourier *31*

1.7.2.2 Bases B-Splines *32*

1.7.3 Suavización de datos funcionales ***33***

1.7.4 Elección del número de funciones base ***35***

1.7.5 Análisis descriptivo funcional ***36***

1.7.6 Datos funcionales atípicos ***37***

1.7.6.1 BAGPLOT para datos funcionales *37*

1.7.7 Análisis de la Varianza Funcional ***39***

1.8 Kriging Ordinario para datos funcionales **40**

1.8.1 Predicción y estimación de parámetros ***41***

1.8.2 Estimación de la traza del semivariograma ***43***

1.8.3 Criterio de evaluación ***44***

CAPITULO II

2 MARCO METODOLÓGICO **45**

2.1 Tipo y diseño de investigación **45**

2.2 Localización del área de estudio 45
2.3 Población de estudio 46
2.4 Recolección de información 46
2.5 Identificación de variables 46
2.6 Operalización de variables 47
2.7 Alcances de investigación 47
2.8 Análisis de datos 47
CAPITULO III
3 RESULTADOS Y DISCUSIÓN 48
3.1 Estructura de la matriz de datos original 48
3.2 Análisis exploratorio de datos 49
3.3 Imputación múltiple en R 50
3.4 Análisis de Datos Funcionales 53
3.4.1 Selección de la Base y número de funciones 53
3.4.2 Suavización de las curvas 54
3.4.3 Análisis Descriptivo Funcional 56
3.4.4 Detección de datos funcionales atípicos 62
3.4.5 Análisis de la Varianza Funcional (FANOVA) 65
3.4.5.1 FANOVA para curvas diarias de temperatura por hora 65
3.4.5.2 FANOVA para curvas anuales de temperatura por día 66
3.5 Análisis descriptivo espacial 68
3.6 Kriging Ordinario para datos funcionales de temperatura 69
3.6.1 OKFD para curvas medias de temperatura diaria por hora 70
3.6.1.1 Análisis estructural 71
3.6.1.2 Validación Cruzada Funcional 72
3.6.1.3 Estimación 74
3.6.2 OKFD para curvas medias de temperatura anual por día 76
3.6.2.1 Análisis estructural 77
3.6.2.2 Validación Cruzada Funcional 78

3.6.2.3 Estimación 81

CONCLUSIONES 85

RECOMENDACIONES 87

GLOSARIO

BIBLIOGRAFÍA

ANEXOS

ÍNDICE DE TABLAS

Tabla 1-1: Codificación de los tipos de estación meteorológica..... 13
Tabla 2-1: Alcance máximo de la capacidad de medición de la Temperatura..... 14
Tabla 3-1: Instrucciones en R para el proceso de imputación múltiple. 17
Tabla 1-2: Operalización de la variable 47
Tabla 1-3: Porcentaje de datos faltantes por estación y año. 48
Tabla 2-3: Resumen del análisis de regresión múltiple para la variable temperatura. 52
Tabla 3-3: Resultados de la prueba FANOVA por estación (Caso 1). 66
Tabla 4-3: Resultados de la prueba de FANOVA por estación (Caso 2)..... 67
Tabla 5-3: Principales tipos de kriging y sus propiedades (Caso 1). 71
Tabla 6-3: Resumen de la SSE de validación cruzada para el OKFD, 2014. 73
Tabla 7-3: Coeficientes del OKFD y distancias más representativas, Alao 2014..... 74
Tabla 8-3: Principales tipos de kriging y propiedades (Caso 2). 78
Tabla 9-3: Resumen de la SSE de la validación cruzada para el OKFD. 80
Tabla 10-3: Coeficientes del OKFD y distancias más representativas, Amulá Casaloma..... 82

ÍNDICE DE FIGURAS

Figura 1-1: Procedimiento de Imputación Múltiple..... 16
Figura 2-1: Representación de una superficie interpolada para una variable regionalizada estacionaria..... 21
Figura 3-1: Representación de una superficie interpolada para una variable regionalizada no estacionaria..... 22

ÍNDICE DE GRÁFICOS

Gráfico 1-1: Comportamiento típico de un semivariograma y sus parámetros básicos. 24
Gráfico 2-1: Comparación de los modelos esférico, exponencial y Gaussiano. 25
Gráfico 3-1: Aproximación mediante base Fourier para 5 y 7 funciones base. 31
Gráfico 4-1: Aproximación mediante base B-Splines de orden 2, 3 y 4..... 32
Gráfico 5-1: BAGPLOT Funcional y Bivariado. 38
Gráfico 1-2: Ubicación de la provincia de Chimborazo y estaciones meteorológicas. 46
Gráfico 1-3: Diagramas de caja de Temperatura para Atillo 2014 y Matus 2015. 49
Gráfico 2-3: Patrón de datos faltantes de la estación Quimiag 2017. 50
Gráfico 3-3: Modelo de imputación múltiple..... 51
Gráfico 4-3: Densidad de la variable temperatura original vs imputada..... 51
Gráfico 5-3: Curvas No suavizadas de temperatura diaria, Alao 2014-2017..... 53
Gráfico 6-3: Ajuste de las curvas y la varianza residual de sus variaciones. 54
Gráfico 7-3: Curvas suavizadas de temperatura del aire de las 11 estaciones meteorológicas de Chimborazo, 2014-2017..... 55
Gráfico 8-3: Media y Desviación Funcional de la temperatura, Estación Alao. 56
Gráfico 9-3: Media y Desviación Funcional de la temperatura, Estación Atillo. 57
Gráfico 10-3: Media y Desviación Funcional de la temperatura, Estación Cumandá. 57
Gráfico 11-3: Media y Desviación Funcional de la temperatura, Estación ESPOCH. 58
Gráfico 12-3: Media y Desviación Funcional de la temperatura, Estación Matus..... 58
Gráfico 13-3: Media y Desviación Funcional de la temperatura, Estación Multitud..... 59
Gráfico 14-3: Media y Desviación Funcional de la temperatura, Estación Quimiag..... 60
Gráfico 15-3: Media y Desviación Funcional de la temperatura, Estación San Juan. 60
Gráfico 16-3: Media y Desviación Funcional de la temperatura, Estación Tixán. 61
Gráfico 17-3: Media y Desviación Funcional de la temperatura, Estación Tunshi. 61
Gráfico 18-3: Media y Desviación Funcional de la temperatura, Estación Urbina. 62
Gráfico 19-3: Bagplots funcional y bivariado de temperatura diaria. 64
Gráfico 20-3: FANOVA de temperatura para la estación de Alao (Caso 1)..... 65
Gráfico 21-3: FANOVA de temperatura para la estación de Alao (Caso 2)..... 67
Gráfico 22-3: Dispersograma de temperatura para 11 (a) y 29 (b) sitios en Chimborazo. 68
Gráfico 23-3: Localización de las estaciones meteorológicas, puntos sistemáticos y a estimar (rojo) en Chimborazo. 69
Gráfico 24-3: Curvas medias de temperatura diaria de las 11 estaciones de Chimborazo. 70
Gráfico 25-3: Semivariograma experimental, 2014. 71
Gráfico 26-3: Curvas medias de temperatura diaria estimadas por VCF, 2014. 72

Gráfico 27-3: Residuos de VCF de las estaciones, 2014. 73
Gráfico 28-3: Estimación de temperatura diaria, Alao 2014. 74
Gráfico 29-3: Estimación de temperatura diaria por horas. 75
Gráfico 30-3: Curvas medias de temperatura anual en 11 (a) y 15 sitios (b) de Chimborazo. . 76
Gráfico 31-3: Semivariograma experimental y modelos teóricos ajustados. 77
Gráfico 32-3: Curvas medias de temperatura anual estimadas por VCF para 11 (a) y 15 (b) sitios en Chimborazo. 79
Gráfico 33-3: Residuos de VCF para 11 (a) y 15 (b) sitios en Chimborazo. 80
Gráfico 34-3: Estimación de temperatura anual, Amulá Casaloma. 81
Gráfico 35-3: Estimación de temperatura anual en sitios no muestreados. 82
Gráfico 36-3: Mapa de temperatura promedio anual en horas de la mañana en la provincia de Chimborazo, 2014-2017. 83
Gráfico 37-3: Mapa de temperatura promedio anual en horas de la noche en la provincia de Chimborazo, 2014-2017. 84

ÍNDICE DE ANEXOS

ANEXO A: COORDENADAS DE LAS 11 ESTACIONES METEOROLÓGICAS EN LA PROVINCIA DE CHIMBORAZO.

ANEXO B: CÓDIGO EN R PARA EL ANÁLISIS EXPLORATORIO E IMPUTACIÓN DE DATOS FALTANTES.

ANEXO C: CÓDIGO EN R PARA ANÁLISIS DE DATOS FUNCIONALES.

ANEXO D: CÓDIGO EN R PARA EL ANÁLISIS GEOESTADÍSTICO DE DATOS FUNCIONALES.

ANEXO E: DIAGRAMAS DE CAJA DE TEMPERATURA DEL AIRE POR ESTACIÓN Y AÑO (ORIGINAL).

ANEXO F: MAPAS DE TEMPERATURA DEL AIRE PROMEDIO POR MES.

RESUMEN

El presente trabajo de investigación tuvo como objetivo estimar la temperatura del aire en sitios no muestreadas de la provincia de Chimborazo, período 2014-2017 a través del análisis geoestadístico de datos funcionales considerando las 11 estaciones meteorológicas que monitorea el GEAA. Se consideró como dato funcional la temperatura diaria por horas y anual por días, suavizadas mediante B-Splines Cúbico y Fourier, con 15 y 365 bases según el comando $min.basis()$ de R y una varianza residual de 0.238 y 0.047 respectivamente. Para determinar el comportamiento de la temperatura se identificó las funciones: media, desviación estándar y atípicas (fueron separadas del análisis) de cada una de las estaciones. A fin de definir el dato funcional para la modelación geoestadística se realizó un FANOVA, tanto para las curvas medias por hora y por día; no se rechazó la hipótesis nula por día, por tal motivo para la modelación espacial se tomó la temperatura promedio de los años en estudio. Mediante validación cruzada se obtuvo menor suma de residuos con el modelo esférico para las estimaciones con kriging ordinario funcional (OKFD), mismo que mientras más datos muestrales se disponga el ajuste es mejor, motivo por el cual se generaron sistemáticamente 29 puntos, de los cuales 4 permitieron mejorar el modelo, por lo que se definió con 15 puntos georreferenciados, con una desviación estándar de 3102.62 grados centígrados. Se estimó la temperatura en cuatro zonas de cultivo de quinua: Amulá Casaloma, Majipamba, San Pedro de Yacupamba y Columbe Grande, cuyos resultados fueron comparados con las temperaturas descargadas de la NASA, obteniendo sumas de cuadrados del error de 355.13, 1878.12, 1465.88 y 765.05 respectivamente. Se recomienda aplicar la misma metodología para el análisis de otras variables meteorológicas.

Palabras clave: <ANÁLISIS DE DATOS FUNCIONALES>, <GEOESTADÍSTICA>, <KRIGING ORDINARIO PARA DATOS FUNCIONALES>, <METEREOLOGÍA>, <TEMPERATURA DEL AIRE>.

SUMMARY

The present titling work had as aim to estimate the air temperature in unsampled sites of the Chimborazo province, period 2014-2017 through the geostatistics analysis of functional data considering the 11 weather stations that the GEAA monitors. The daily, hourly and annual temperatures per day, softened by B-Splines Cubic and Fourier, with 15 and 365 bases according to the $min.basis()$ command of R and a residual variance of 0.238 and 0.047 respectively, were considered as functional data. To determine the behavior of the temperature, functions were identified: mean, standard deviation and atypical (were separated from the analysis) of each of the stations. In order to define the functional data for geostatistical modeling, a FANOVA was carried out, both for the average curves per hour and per day; the null hypothesis was not rejected for the day, for this reason for the spatial modeling the average temperature of the years under study was taken. By means of cross-validation, a smaller sum of residues was obtained with the spherical model for estimating with ordinary kriging functional (OKFD), even if the more sample data the adjustment is available, the better, which is why 29 points were systematically generated, of which 4 allowed to improve the model, so it was defined with 15 georeferenced points, with a standard deviation of 3102.62 degrees Celsius. The temperature in four quinoa cultivation zones was estimated: Amulá Casaloma, Majipamba, San Pedro de Yacupamba and Columbe Grande, whose results were compared with the temperatures downloaded from NASA, obtaining sums of error squares of 355.13, 1878.12, 1465.88 and 765.05 respectively. It is recommended to apply the same methodology for the analysis of other methodological variables.

Keywords: <FUNCTIONAL DATA ANALYSIS>, <GEOSTATISTICS>, <ORDINARY KRIGING FOR FUNCTIONAL DATA>, <METEROLOGY>, <AIR TEMPERATURE>.

INTRODUCCIÓN

Desde el ámbito ambiental hasta el social, el componente meteorológico es un factor indispensable para el diseño y desarrollo de proyectos relacionados con diferentes campos de las ciencias. La temperatura es una de las variables meteorológicas más importantes en este tipo de estudios, su papel es primordial en los ecosistemas naturales, al influir en el desarrollo de una gran variedad de especies vegetales y animales. En el área agrícola, determina el manejo adecuado del agua, la longitud de los ciclos de cultivo y aprovechamiento de los insumos. En el campo humano es fundamental en el confort térmico, pero sobre todo es más trascendental en estudios sobre cambio climático. Sin embargo, no siempre se puede disponer de datos meteorológicos en sitios específicos para desarrollar estudios en beneficio de la sociedad, ya que la obtención de los mismos conlleva significativos recursos económicos.

El Análisis de Datos Funcionales (ADF) ha tenido un gran interés durante los últimos años debido a la versatilidad en diversos campos de estudio, muy usado a nivel mundial a la hora de trabajar con grandes cantidades de datos, que se presentan en forma continua y en la mayoría de casos suelen reflejarse por medio de curvas (datos funcionales). Actualmente existen diversos trabajos que tienen como propósito la investigación de curvas vistas como realizaciones de funciones aleatorias. Los libros más sobresalientes de entre toda la literatura disponible en este tipo de estudio son los de Ramsay y Silverman (2005) y Ferraty y Vieu (2006), que son referencias bibliográficas fundamentales para resolver varios problemas desde un contexto funcional. El libro de Ramsay, Hooker y Graves (2009) trata estos problemas desde un aspecto computacional usando softwares estadísticos como: R y MATLAB.

El análisis geoestadístico ha tenido un gran uso en varias ciencias como: agronomía, meteorología, hidrología, y otras, donde la variable de interés depende de la posición geográfica y del parámetro tiempo. En el contexto funcional se han adaptado algunas técnicas geoestadísticas para realizar la predicción espacial de curvas cuando se dispone de una muestra de curvas o funciones en una región con continuidad espacial. Goulard y Voltz (1993), trataron el problema de predicción espacial en sitios no muestreados bajo el supuesto de estacionariedad, en este trabajo las funciones son conocidas a partir de un conjunto finito de puntos y un modelo paramétrico los ajusta para reconstruir la curva completa. Giraldo, Delicado y Mateu (2010) retoman este trabajo para proponer la metodología de kriging ordinario para datos funcionales (OKFD) donde la función a estimar es una combinación lineal de las curvas observadas, las cuales son suavizadas mediante un ajuste no-paramétrico y un parámetro que es elegido a través de validación cruzada funcional (Ginzo, 2011; Cardona, 2015).

El Grupo de Investigación de Energías Alternativas y Ambiente (GEAA) monitorea 11 estaciones meteorológicas en: Alao, Atillo, Cumandá, ESPOCH, Matus, Multitud, Quimiag, San Juan, Tixán, Tunshi y Urbina; sin embargo, no son suficientes para la obtención de resultados con menor sesgo posible debido a que la provincia de Chimborazo se encuentra en la zona Andina caracterizada por la existencia de microclimas. Aplicando métodos geo-funcionales el objetivo de este proyecto de investigación es modelar y estimar la temperatura del aire (ºC) en zonas no muestreadas de la provincia, a partir de curvas observadas, debido a la ausencia de estaciones meteorológicas.

En el capítulo I se realiza una investigación teórica, con los conceptos necesarios para el desarrollo del proyecto de investigación, donde se presenta una breve descripción de la geoestadística tradicional y aplicada a datos funcionales, resaltando la obtención de la forma funcional mediante bases de funciones, además del análisis descriptivo, detección de atípicos, análisis de varianza funcional y la adaptación del kriging ordinario clásico a los datos funcionales.

En el capítulo II se describe el marco metodológico, en el se define la zona de estudio que cuenta con 11 estaciones meteorológicas en la provincia Chimborazo, el diseño de investigación, población de estudio, operacionalización de la variable y el análisis de datos.

En el capítulo III se muestran los resultados obtenidos en la depuración de los datos mediante las técnicas estadísticas tradicionales, especificaciones de la Organización Mundial de Meteorología (OMM), y el relleno de faltantes con el método de imputación múltiple MICE. Se realiza el suavizado de los datos mediante bases B-Splines y Fourier, donde los parámetros de suavización son elegidos a través de la validación cruzada generalizada (VCG). El análisis descriptivo y detección de atípicos funcional permite determinar el comportamiento de la temperatura a lo largo de las horas y días, mientras que el Análisis de Varianza Funcional (FANOVA) identifica las diferencias significativas entre las medias funcionales de los años de cada una de las estaciones. La modelación de las curvas de temperatura media se realiza mediante el análisis estructural por validación cruzada funcional (VCF) aplicando 4 modelos de semivariograma para 11 y 15 referencias geográficas, con el fin de seleccionar el modelo más adecuado para la estimación de temperatura en zonas no muestreadas de la provincia de Chimborazo con menor SSE (Suma de cuadrados del error).

Antecedentes

Hasta el siglo XX, el término meteorología de manera científica se utilizó para tres actividades: una empírica encargada de recolectar datos e inferir a partir de ellos, una teórica por la aplicación de la física a los fenómenos atmosféricos y una práctica de predicción del tiempo que surgió de la mano del astrónomo francés Urban Le Verrier (1811-1877). Estas actividades con el tiempo se unificaron, debido a la relación que presentaban y a los avances producidos en cuanto al uso de potentes ordenadores y sofisticados modelos numéricos, obtención de medidas de diversas fuentes incluyendo satélites, gran almacenamiento de datos, y demás herramientas estándar en meteorología (Iturralde, 2003, pp. 145-155).

El Instituto Ecuatoriano de Meteorología e Hidrología (INAMHI) creado el 4 de agosto de 1961 es la entidad con competencias de implementación y operación de estaciones meteorológicas en Ecuador, las cuales registran grandes volúmenes de información tales como la temperatura del aire, humedad, presión atmosférica, radiación, etc. (Narváez, 2012, p. 1). Sin embargo, algunas actividades de gestión e investigación, requieren información en sitios no muestreados, por ello para aprovechar la gran cantidad de datos que generan las estaciones por segundo, minuto u hora, una alternativa para modelar y estimar las variables meteorológicas, es utilizar una de las técnicas del Análisis de Datos Funcionales aplicado a la geoestadística.

El análisis de datos funcionales (ADF) tiene sus inicios desde el trabajo pionero de Deville (1974) de Ramsay y Dalzell (1991), fue entonces que a partir de la década de los 90, debido al incremento y mejora de las aplicaciones informáticas, se han producido grandes avances en técnicas estadísticas que permiten trabajar con gran cantidad de datos. El ADF presenta muchos de los problemas de la estadística clásica como: descripción de variables, modelización, clasificación, inferencia, entre otros (Tarrío y Naya, 2011, pp. 212-214).

El interés de la comunidad estadística en la mejora constante de técnicas de análisis de datos funcionales, hoy en día ha permitido que nuevos métodos estén disponibles, como la geoestadística funcional, una de las categorías de métodos funcionales espaciales que ha tenido una gran variedad de estudios en interpolación de curvas cuando se cuenta con una muestra de curvas en una región con continuidad espacial. Este método de predicción funcional es una adaptación de las técnicas de predicción más clásicas como Kriging, adecuado al caso de las curvas espacialmente correlacionadas, existen diferentes tipos como: kriging ordinario (OK), kriging universal (UK), cokriging y otros (Mateu y Romano, 2017, pp. 1-3).

El trabajo de Goulard y Voltz (1993, pp. 805-806) titulado "Geostatistical Interpolation of Curves: A Case Study in Soil Science", fue crucial para futuros estudios en geoestadística funcional, ya que resolvió el problema de predicción espacial de funciones en sitios no muestreados bajo el supuesto de estacionariedad. La predicción se realizó mediante funciones conocidas a partir de un conjunto finito de puntos, y un modelo paramétrico que se suponía conocido para reconstruir la curva entera. El documento describe teóricamente tres enfoques geoestadísticos, dos de ellos multivariantes con un salto al cokriging y el tercero similar al kriging clásico, aplicados a datos de las características de retención del agua del suelo.

En la tesis doctoral "Geostatistical analysis of functional data", se aplicó algunas técnicas de predicción espacial asumiendo estacionariedad, entre ellos el predictor kriging ordinario para datos funcionales (OKFD), kriging continuo en el tiempo para datos funcionales (CTKFD), cokriging basado en datos funcionales (CBFD) y kriging funcional: modelo total (FKTM). Se utilizó datos agronómicos pertenecieron a perfiles de resistencia a la penetración del suelo (MPa) obtenidos en el año 2004 de una granja agrícola en la Estación Experimental Marengo (Universidad Nacional de Colombia) en 32 puntos de muestreo, donde de cada uno se recolectó 334 observaciones en profundidades entre 0 a 45 cm. El objetivo de este análisis fue establecer estrategias de manejo del suelo a diferentes profundidades, y así aumentar la rentabilidad y sostenibilidad en la producción de cultivos. También utilizó datos meteorológicos de temperatura diaria promediada entre 1960 – 1994 registrada en 35 estaciones meteorológicas y marítimas en Canadá que a diferencia de los anteriores cubren un área más pequeña y homogénea. En todos los casos, se utilizó un enfoque no paramétrico para suavizar los datos mediante B-Splines y Fourier con un número de funciones básicas elegidas por validación cruzada funcional (VCF). Para los tres conjuntos de datos se estimó los parámetros funcionales para la aplicación de un adecuado modelo geoestadístico y realizar predicciones en sitios no muestreados. Por último se realizó una comparación entre los modelos a través de la validación cruzada funcional de los errores al cuadrado, donde el FKTM y OKFD fueron los mejores, sin embargo OKFD resultó el más adecuado desde el punto de vista práctico y computacional (Giraldo, 2009, pp. 1-102).

En el artículo científico "Ordinary kriging for function-valued spatial data" publicado en la revista Enviromental and Ecological Statistics, se trabajó con datos reales de resistencia a la penetración del suelo cuyo objetivo era realizar predicciones en ubicaciones sin muestrear, siempre y cuando los valores de datos sean funciones. Las curvas suavizadas se obtuvieron mediante funciones básicas B-Splines, donde se aplicó un ajuste no-paramétrico para pre-procesar las funciones observadas y obtener el parámetro de suavización, el cual fue elegido mediante un proceso llamado validación cruzada funcional (VCF). Gráficamente se detectó un valor atípico y se trabajó con 31 curvas. Para la aplicación de la metodología del kriging ordinario para datos funcionales,

donde la función a predecir es una combinación lineal de las curvas observadas propuesto inicialmente por Goulard y Voltz (1993), se calculó para varios retrasos espaciales el traza-semivariograma, donde se ajustó un modelo esférico estimado mediante mínimos cuadrados ordinarios. Por ello para evaluar la bondad del ajuste del modelo elegido, comparó gráficamente las curvas observadas con las predichas, obtenidas mediante validación cruzada funcional por kriging ordinario funcional. La predicción se realizó en un punto no muestreado de coordenadas 11179 (longitud) y 9750 (latitud). (Giraldo et al., 2011, p. 411-425).

Ramón Giraldo, Jorge Mateu y Pedro Delicado en el artículo denominado "$geofd$: An R Package for Function-Valued Geostatistical Prediction" propusieron el uso de la librería $geofd$ dentro del software estadístico R. Este trabajo implementó funciones para modelar la función traza-variograma y realizar predicciones espaciales utilizando el método kriging ordinario para datos funcionales. Para ilustrar su uso se analizó datos reales y simulados. Primero se aplicó la metodología a los datos promedio de temperatura diaria de 35 estaciones meteorológicas ubicadas en las provincias marítimas canadienses como: Nueva Escocia, Nuevo Brunswick y la Isla del Príncipe Eduardo, utilizados en varios trabajos de estos autores, datos disponibles en R cargando el comando data(maritimes). Las curvas de temperatura se ajustaron mediante bases Fourier y B-Splines con un número de base igual a 65. Los parámetros para realizar la predicción de kriging ordinario para la curva de temperatura en la estación de Moncton, se obtuvo a través de la resolución estimada de un traza-variograma basado en un sistema lineal, donde el mejor modelo fue el exponencial con una suma mínima de cuadrados del error. Los datos simulados fueron suavizados mediante el uso de bases Splines con 15 funciones, obteniendo 365 curvas para realizar la predicción con el uso de la función $okfd$. Los avances en este paquete fueron posibles debido a varias contribuciones a CRAN como el paquete fda que proporciona métodos para suavizar datos mediante funciones básicas y $geoR$ para modelar la función trace-variograma, entre otros (Giraldo et al., 2012, pp. 385-405).

En el trabajo de investigación "Generación de superficies climáticas usando datos funcionales de temperatura y precipitación por medio de métodos geoestadístico para el Valle del Río Cauca, Colombia", con el propósito de sistematizar, analizar y generar superficies climáticas interpoladas usando conjuntamente el software R y ArcGIS, mediante técnicas funcionales y geoestadísticas, se utilizó un enfoque teórico-práctico, empezando por la identificación y exploración de los datos de temperatura máxima y mínima (ºC) y precipitación (ml) de 28 estaciones climáticas puntuales a los largo del Valle del Río Cauca, Colombia registrados entre los años 1997-2012. Luego a través de análisis estadísticos se aplicó un proceso de control de calidad, donde se detectó los datos anómalos y se rellenó faltantes mediante el paquete RMAWGEN (Generador meteorológico regresivo automático de sitios múltiples). El proceso de suavización de los datos lo realizó usando

bases Fourier, resumiendo las 28 series multianuales (5846 registros) a promedios diarios de un año (365 días). Aplicando los métodos geo-funcionales propuestos por Giraldo y colaboradores halló los parámetros del variograma usando la técnica de validación cruzada, siendo el mejor modelo de variograma el exponencial. Finalmente, para la predicción en sitios no muestreados bajo la ausencia de estacionariedad, aplicó la técnica de kriging ordinario funcional y generó las superficies climáticas para su lugar de estudio (Cardona, 2015, pp. 13-63).

En el trabajo de postgrado de Técnicas Estadísticas realizado por María José Ginzo Villamayor titulado "Análisis geoestadístico de datos funcionales", se hizo una revisión crítica de los métodos geoestadísticos funcionales que se han considerado previamente en algunos estudios utilizando en su mayoría datos de temperatura, debido a su componente espacial y funcional. Es por ello que para este proyecto utilizó datos de temperatura promedio del ambiente de 66 estaciones meteorológicas para Galicia durante el año 2009 en provincias como Coruña, Lugo, Orense y Pontevedra. La base de datos fue suavizada con 35 bases de Fourier, mediante el método funcional min.np que usa una estimación no paramétrica kernel. Los modelos de semivariograma isotrópicos como: Circular, Exponencial, Esférico y Matérn, aplicados a los datos fueron similares, sin embargo, seleccionó el esférico para fines prácticos. Una vez seleccionado el modelo, mediante validación cruzada (dejando un punto fuera de la muestra), utilizó kriging ordinario funcional (OKFD) para predecir una curva entera en un sitio no muestreado observando que los valores predichos estén cerca de los observados, en este caso correspondía a la Facultad de Matemáticas con coordenadas 536.07(Este) y 4747.02(Norte). Además, realizó predicciones con el kriging universal funcional (UKFD) con 8 modelos, el kriging continuo variable en el tiempo para datos funcionales (CTKFD) que combina el kriging ordinario y el modelo funcional lineal concurrente (punto-wise), el cokriging funcional (CBFD) y el kriging funcional: modelo total (FKTM); concluyendo que el predictor OKFD sin los efectos de longitud y latitud es la mejor opción, debido a que es más simple y computacionalmente rápido para trabajar con grandes conjuntos de datos que presenten curvas homogéneas (Ginzo, 2011, pp. 1-92).

Planteamiento del problema

Una de las principales preocupaciones de los meteorólogos es contar con información depurada y confiable, para el conocimiento del tiempo atmosférico y la situación climática. La instalación de estaciones, dispositivos (sensores), recolección de datos meteorológicos, procesamiento y validación de los mismos requiere una significativa inversión por parte del Estado, por ello las tareas de vigilancia, control y planificación mediante la medición de variables meteorológicas son demasiado costosos, por lo que limita dichas tareas en todas las zonas no muestreadas de la Provincia de Chimborazo y por ende su caracterización, especialmente de la temperatura del aire, variable objeto de este estudio. Este trabajo de investigación pretende estimar la temperatura en sitios no muestreados y aprovechar la gran cantidad de datos que genera la red de estaciones de la provincia, mediante el análisis geoestadístico de datos funcionales.

¿El análisis geoestadístico de datos funcionales permitirá modelar y estimar el comportamiento de la temperatura del aire en las zonas donde no se cuenta con alguna estación meteorológica dentro de la provincia de Chimborazo?

Justificación

En meteorología la predicción del tiempo ha sido una tarea permanente del ser humano desde tiempos remotos, factor importante para varios tipos de modelos hidrológicos, agrometeorológicos, medio ambientales, de crecimiento poblacional, mortalidad, económicos, entre otros. En Ecuador las condiciones meteorológicas afectan a numerosos sectores, frecuentemente al sector agrícola que anualmente pierde entre el 5% y 10% de la producción agrícola a consecuencia de condiciones meteorológicas desfavorables como heladas y sequías. El Banco Interamericano de Desarrollo con el aporte de la Consultoría del Instituto Hidráulico de Dinamarca (DHI), mediante estudios realizados sobre el cambio climático hasta el año 2039, mencionó que el incremento de la temperatura en 1 ºC, significaría una pérdida del 12% en caudal de varias fuentes, 1% por retroceso de glaciares, 5% por pérdida de superficie de páramos y 6% por cambio de precipitación (INAMHI).

En la actualidad el avance de las nuevas tecnologías han permitido el diseño de dispositivos para la medición de variables meteorológicas y con ello la inclinación por almacenar grandes masas de datos, incrementado el número de situaciones donde se dispone de datos funcionales que

requieren ser analizados estadísticamente para dar soporte a la toma de decisiones (Salmerón, 2008, pp. 28-31). La presente investigación tiene relevancia en el análisis estadístico de datos funcionales (FDA), debido a su capacidad para trabajar con grandes cantidades de datos en diversas áreas como: la salud, meteorología, medioambiental, entre otras. Los métodos tradicionales como: regresión, ANOVA y componentes principales, entre otros, han sido consideradas desde el punto de vista funcional, en general, se centran en variables funcionales independientes e idénticamente distribuidas y son aplicables en muchas disciplinas de las ciencias, es así además que existe gran interés en la modelización de datos funcionales espacialmente correlados (Ginzo, 2011, p. 2). Un enfoque apropiado para el análisis estadístico de diferentes fenómenos naturales que se suscitan en el diario vivir es la teoría de análisis geoestadístico de datos funcionales, mismo que permitirá estimar la temperatura del aire (curvas) en sitios no muestreados de la provincia de Chimborazo con el mayor grado de confiabilidad, sin dejar de lado la importancia del uso del software estadístico R, por su potencia y flexibilidad. R compila sobre varias plataformas como: UNIX, Windows y MacOS, además es el más usado por investigadores a nivel mundial, porque permite el desarrollo de código de acuerdo a las necesidades y tiene un fácil acceso a las librerías y manuales requeridos para el análisis estadístico de datos funcionales (R Core Team, 2018).

La obtención de datos meteorológicos conlleva una fuerte inversión, sin embargo, el Centro de Investigación de Energías Alternativas y Ambiente (CEAA) conjuntamente con otras instituciones está a cargo del monitoreo de las variables meteorológicas de 11 estaciones instaladas en la provincia de Chimborazo y pone a disposición de la ciudadanía dichos datos, por ello este trabajo de investigación está avalado por el GEAA (Grupo de Investigación de Energías Alternativas y Ambiente).

Objetivos

Objetivo General

Modelar y estimar la temperatura del aire en zonas no muestreadas de la provincia de Chimborazo mediante técnicas geoestadísticas con datos funcionales.

Objetivos Específicos

- Realizar un análisis exploratorio de la base de datos de temperatura del aire de las estaciones meteorológicas.
- Realizar un análisis estadístico descriptivo funcional de la variable temperatura del aire.
- Modelar la temperatura del aire mediante herramientas geoestadísticas con datos funcionales.
- Estimar la temperatura del aire en zonas no muestreadas de la provincia de Chimborazo.

CAPITULO I

1 MARCO TEÓRICO REFERENCIAL

1.1 Meteorología

La meteorológica es una ciencia que permite estudiar y predecir numerosos fenómenos que se producen en la atmósfera en espacio y tiempo, siendo clave para diversos ámbitos de la investigación tales como: ambiental, agrícola, aeronáutica, médica, energética, ecológica entre otros.

En la actualidad, la meteorología es una ciencia muy avanzada, basada en la Física y en el uso de la tecnología, donde los meteorólogos predicen el tiempo hasta con una semana de anticipación con un porcentaje mínimo de error. Este estudio se basa en el conocimiento de una serie de magnitudes o variables meteorológicas como: la temperatura, la radiación solar, la presión atmosférica o la humedad (Rodríguez et al., 2014, p. 6).

A continuación, se presenta una breve descripción del desarrollo de cada una de las tradiciones (empírica, teórica y práctica) relacionadas con la meteorología previo al escenario de la predicción numérica del tiempo.

1.1.1 Una tradición empírica. Climatología

Hasta el siglo XVII con la invención del termómetro y del barómetro las observaciones meteorológicas se realizaron de forma sistemática, dando lugar a un cambio profundo en las descripciones del tiempo, que pasaron de carácter cualitativo a cuantitativo. Esto fue debido a que en el siglo XVII la temperatura, la humedad, la presión atmosférica, la cantidad de precipitación, la fuerza de viento, entre otras se pudieron medir. Cabe recalcar que se atribuye a Galileo Galilei (1564-1642) la construcción del primer termómetro en los últimos años del siglo XVII, Evangelista Torricelli (1608-1647) construyó el primer barómetro en 1943. En esta centuria se inventaron algunos aparatos para medir las precipitaciones, la dirección y fuerza del viento, lo que produjo un aumento considerable de datos meteorológicos, y a su vez propició el planteamiento de nuevas cuestiones teóricas para tratar dichos datos (Iturralde, 2003, pp. 146-147).

1.1.2 Una tradición teórica. Física de la atmosfera

La obra de Aristóteles (384-322 a.C.) "Meteorología", escrita alrededor de los años 340 a.C., estableció una tradición teórica acerca de la meteorología y se aseguró que esta ciencia fuera estudiada como una parte de la filosofía natural. Este tratado fue la base de todos los estudios teóricos de meteorología hasta comienzos del siglo XVII, donde experimentó grandes cambios. A mediados del siglo XIX, a pesar de los datos proporcionados por el termómetro y el barómetro, y a la relevancia de las matemáticas en la mecánica a partir de la obra Isaac Newton (1642-1727), la mayor parte de las teorías meteorológicas seguían siendo totalmente cualitativas, sólo pocas cuestiones eran tratadas matemáticamente, por ejemplo, la relación entre la altitud y la presión atmosférica, fue así que en la segunda mitad del siglo XIX se explicaron muchos fenómenos atmosféricos, lo cual a su vez fomentó el gran desarrollo de los estudios teóricos.

Los meteorólogos teóricos defendían que la meteorología debería ser física aplicada y los datos observados debían explicarse de forma deductiva, considerándose como pioneros de una nueva meteorología, en contraposición con los empíricos que afirmaban que la meteorología era una ciencia independiente cuyas leyes se tenían que inducir directamente de los datos (Iturralde, 2003, pp. 147-148).

1.1.3 Una tradición práctica. Predicción del tiempo

Durante siglos los pronósticos se apoyaban en signos naturales, por ejemplo, un pequeño halo alrededor del sol como presagio de lluvia, que en algunos casos se expresaban en forma de refranes, interpretaciones ambiguas que poseían cierta validez local para previsiones a muy corto plazo. A finales del siglo XIX la predicción del tiempo presentaba el siguiente proceso: cada día los pronosticadores construían mapas sinópticos con los valores de las distintas variables atmosféricas recogidos a una misma hora en cien o más localidades y que les eran enviados por telégrafo. Uno de los más importantes era el mapa de líneas isobaras y que mostraba la distribución de las presiones barométricas recogidas. Una vez realizado el retrato del tiempo actual, la principal tarea de los pronosticadores era hacer un mapa pronóstico del día siguiente en los lugares considerados. Este trabajo desembocaba en una predicción que era una simple descripción verbal del tiempo "lluvioso", "ventoso", "despejado" y "frío".

Actualmente existen varios modelos de predicción del tiempo, cada uno con sus ventajas e inconvenientes, como norma general, cuando más preciso sea el método más cálculos habrá que hacer y por lo tanto más tiempo se tardará en ejecutarlo. Sin embargo para que una predicción tenga sentido se debe ejecutarla en un plazo de tiempo relativamente corto, por lo que se elegirán métodos precisos más rápidos (Iturralde, 2003, pp. 148-153).

1.2 Variables Meteorológicas

Las variables meteorológicas son parámetros o elementos que caracterizan el estado del tiempo (a corto plazo) o del clima (a largo plazo) mediante magnitudes que son medibles, y que a través de su comportamiento permiten conocer la condición de la atmósfera. Las variables meteorológicas más importantes que nos ayudan a conocer esta condición son:

- Temperatura del aire.
- Presión atmosférica.
- Humedad relativa
- Velocidad y dirección del viento.
- Radiación solar global y difusa.
- Evaporación, entre otras.

En este proyecto se hará énfasis en la variable temperatura del aire.

1.2.1.1 Temperatura del aire

La temperatura del aire es una de las variables meteorológicas más utilizadas para describir el estado de la atmósfera, la cual varía entre el día y la noche, entre una estación y otra, entre una ubicación geográfica y otra, además de estar influenciada por factores como el viento o la humedad. Por ello debido a que está sometida a numerosas oscilaciones, está condicionada por la longitud, latitud y altura s.n.m. Normalmente, en invierno puede estar bajo los 0 ºC mientras que en verano puede superar los 40 ºC (Rodríguez et al., 2014, pp. 12-15).

1.3 Estación Meteorológica

Una estación meteorológica es una instalación destinada a medir y registrar habitualmente numerosas variables que alteran el estado de la atmósfera, las cuales son recolectadas para realizar predicciones meteorológicas y estudios climáticos. La situación climática en una zona de estudio será más detallada y exacta si se cuenta con numerosas estaciones meteorológicas. Las estaciones meteorológicas pueden ser convencionales o automáticas (Tipos de estaciones meteorológicas):

- Estaciones convencionales. - Necesitan de una persona calificada para recopilar y transmitir información meteorológica, como lo es el observador meteorológico.
- Estaciones automáticas. - Son configuradas para que el registro de datos en lugares remotos o inaccesibles lo transmitan de manera automática.

Dependiendo de las variables a medir, las estaciones se clasifican en diferentes tipos, a continuación, se detalla la descripción de la codificación utilizada para cada estación (Tabla 1-1):

Tabla 1-1: Codificación de los tipos de estación meteorológica.

Código	Tipo de estación
AN	Anemográfica
AP	Agrometeorológica
AR	Aeronáutica
CE	Climatológica Especial
CO	Climatológica Ordinaria
CP	Climatológica Principal
PC	Plataforma Colectora de Datos
PG	Pluviográfica
PV	Pluviométrica
RS	Radio Sonda

Fuente: (INAMHI).

Realizado por: Checa G., Marisol C.,2020.

La instalación de la red de estaciones, así como la ubicación de los instrumentos para el registro de los datos son realizadas según las normas internacionales que establece la Organización Meteorológica Mundial (OMM).

1.3.1 Dispositivos para medir la temperatura del aire

La temperatura del aire es señalada mediante un termómetro expuesto al aire y protegido de la radiación solar directa, inventado por Galileo en 1953. Se mide en grados Celsius y décimas de grado. Según el Anuario que presentó el INAMHI en el año 2017 los tipos de termómetros que miden la temperatura del aire son:

- Termómetro seco. - Utiliza la diferencia de dilatación del líquido (mercurio o alcohol), y vidrio que lo contiene para la medición de la temperatura.
- Termómetro de mínima. - Indicador de la temperatura mínima alcanzada durante un intervalo de tiempo dado. Para su lectura a las 07:00 siempre queda en posición mínima.
- Termómetro de máxima. - Indicador de la temperatura máxima alcanzada durante un intervalo de tiempo dado. Se lee diariamente a las 19:00.

- Termógrafo. - Instrumento proporcionado de un mecanismo que registra continuamente la temperatura de manera gráfica durante un intervalo de tiempo. Es útil ante la incapacidad del ser humano de observar de forma continua la variación de la temperatura en el tiempo.

Los primeros 3 dispositivos son colocados de acuerdo a las normas técnicas de la OMM, donde sugiere ventilación, protección de la precipitación y radiación solar directa, y una altura del suelo determinada a 2 metros (INAMHI).

1.3.2 Especificaciones funcionales de las estaciones meteorológicas

Las estaciones meteorológicas automáticas presentan los siguientes parámetros de especificación para la variable temperatura:

Tabla 2-1: Alcance máximo de la capacidad de medición de la Temperatura.

Variable	Alcance máximo	Resolución mínima transmitida	Modo de observación
Temperatura del aire	-80 º C - +60 º C	0,1 K	I, V
Temperatura del punto de rocío	-80 º C - +60 º C	0,1 K	I, V
Temperatura del suelo	-80 º C - +80 º C -50 º C - +50 º C	0,1 K	I, V
Temperatura de la nieve	-80 º C - 0 º C	0,1 K	I, V
Temperatura del agua-pozo, río, lago, mar	-2 º C - +100 º C	0,1 K	I, V

Fuente: (OMM)

Realizado por: Checa G., Marisol C.,2020.

Según la Guía de Instrumentos y Métodos de la OMM, el modo de observación es el tipo de datos transmitidos, es decir:

- I: Instantáneos – valor de 1 minuto
- V: Variabilidad – promedio, desviación típica, máxima, mínima, intervalo, mediana, etc., de las muestras donde la información transmitida depende de la variable.

1.4 Relleno de valores faltantes

1.4.1 Dato faltante

Son aquellos valores que no constan debido a diferentes problemas como por ejemplo errores en la transcripción de los datos, falla en los dispositivos de medición y otros. No tomar en cuenta los faltantes puede acarrear repercusiones graves, ya que el trabajo de investigación que se lleva a cabo puede presentar resultados sesgados. La eliminación de NA (datos ausentes) es factible si se presentan aleatoriamente y su número es inferior al 5%. Cuando no se pueda ignorar estos faltantes, la forma más adecuada de tratarlos es llenar estos espacios faltantes con valores plausibles, a este procedimiento se le conoce como imputación (Useche y Mesa, 2006, pp. 130-140).

1.4.2 Imputación Múltiple

Los métodos de imputación múltiple reemplazan cada valor perdido por un conjunto de *m* valores, obteniéndose así *m* conjuntos completos de datos, lo que da lugar a *m* estimaciones con sus respectivas varianzas o errores estándar. Estos métodos minimizan el sesgo y la pérdida de potencia estadística causada por datos faltantes en forma completamente aleatoria (Missing Completely at Random MCAR) o datos faltantes en forma aleatoria (Missing at Random MAR).

Aquellos que promueven la imputación múltiple como el método más adecuado para rellenar o reponer información omitida, afirman que este procedimiento genera buenos resultados, aún con omisiones de 30%, 40% o 50%, pero en la medida de que la falta de datos supera el umbral establecido se pone en riesgo la confiabilidad estadística de las variables. Por ello no se recomienda imputar datos si la omisión de valores supera el 20%.

El procedimiento de esta metodología consta de varias etapas, y son:

1) **Etapa de imputación. -** Crea varias copias de conjuntos de datos (m), y cada uno contiene diferentes estimaciones de valores perdidos.
2) **Etapa de análisis. -** Analiza los conjuntos de datos rellenados, llevándonos a la obtención de *m* parámetros estimados y error estándar.
3) **Etapa de puesta en común. -** Combina todo en un conjunto simple de resultados.

Es decir, para cada simulación se analizan la matriz de datos resultante a partir de métodos estadísticos convencionales, y posteriormente se combinan los resultados para generar estimadores robustos, error estándar e intervalos de confianza.

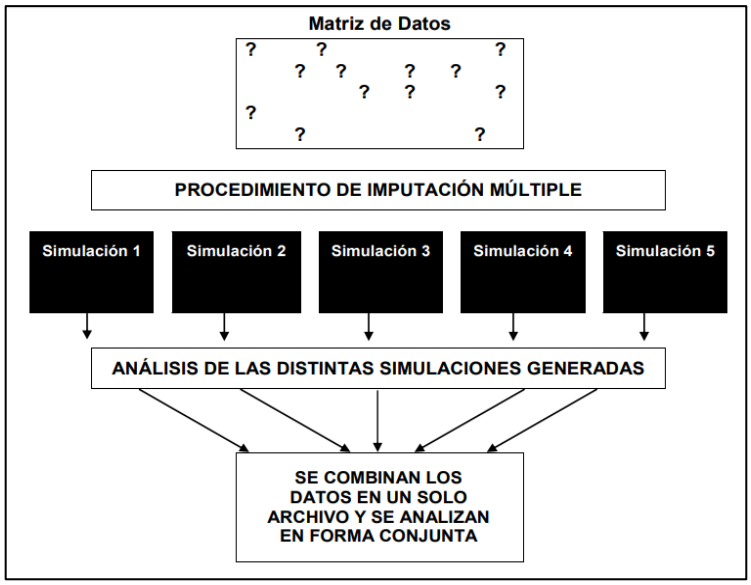


Figura 1-1: Procedimiento de Imputación Múltiple.

Fuente: (Galván y Medina, 2007).

Los supuestos que debe cumplir este tipo de metodología son:

a) El patrón de datos faltantes es aleatorio MAR, lo cual indica que la probabilidad de que existan datos omitidos en la variable X depende de otras variables.
b) El modelo estadístico y/o econométrico utilizado para generar los datos imputados debe ser apropiado, es decir entre la variable a imputar y las covariables debe existir correlación alta.
c) Es preciso que el modelo de análisis guarde relación con el que se utilizó para efectuar el procedimiento de imputación.

Esta metodología ha permanecido durante muchos años en un segundo plano, debido a la inexistencia de herramientas informáticas adecuadas para poder crear las imputaciones. Sin embargo en la actualidad ya existen diferentes métodos de imputación múltiple como el algoritmo MICE (Imputación Multivariada por Ecuaciones Encadenadas), disponible en distintas aplicaciones comerciales y de acceso libre, que además de ser eficiente permite combinar archivos de datos que se generen (Galván y Medina, 2007; Castro, 2014).

1.4.3 MICE (Multivariate Imputation by Chained Equations)

Este algoritmo es un método de imputación múltiple basado en ecuaciones encadenadas que permite imputar valores perdidos con valores de datos plausibles, extraídos de una distribución específicamente diseñada para cada punto de dato faltante, es decir cada variable tiene su propio modelo de imputación. MICE puede imputar mezcla de datos continuo, binarios, categóricos no

ordenados y datos categóricos ordenados, así como datos continuos de 2 niveles y mantener la coherencia entre imputaciones mediante imputación pasiva (Castro, 2014, p. 16).

1.4.3.1 Paquete MICE en R

Este paquete es probablemente el mejor de los disponibles en R, debido a que usa el método FCS (Fully Conditional Specification) de imputación múltiple, además de ser el más completo con mayor número de instrucciones y flexible al implementar un modelo de imputación de acuerdo al tipo de variable que se desea completar. En R el proceso de imputación múltiple se estructura en 3 instrucciones:

Tabla 3-1: Instrucciones en R para el proceso de imputación múltiple.

Etapa	Instrucción	Descripción
Imputación	mice ()	Imputa los datos faltantes y crea un número *m* de datos completos.
Análisis	with ()	Analiza los *m* modelos calculados
Pooling	pool ()	Combina estimaciones de parámetros

Fuente: (Van Buuren y Groothuis-Oudshoorn, 2011).
Realizado por: Checa G., Marisol C.,2020.

El algoritmo MICE requiere una especificación de un método de imputación univariante separadamente para cada variable incompleta, por ello R en su librería MICE considera varios modelos. Normalmente este algoritmo utiliza el método del ajuste de la media predictiva pmm (predictive mean matching) un enfoque semi-paramétrico, que asegura que los valores imputados sean plausibles si se viola el supuesto de normalidad, a diferencia de usar un método de regresión que supone una distribución normal. Además, para tener un grado de precisión satisfactorio en la imputación el número de imputaciones es 5, valores por defecto que se presentan en R **(**Van Buuren y Groothuis-Oudshoorn, 2011; Castro, 2014).

1.4.3.2 Validación del modelo de imputación

La validez del modelo implica un proceso de comparación entre los valores observados con los valores estimados usando estadísticos como: coeficiente de determinación (R^2) para medir la bondad del ajuste, la raíz del error cuadrático medio (RMSE) y el error medio absoluto (MAE) para indicar la magnitud y distribución de los errores del modelo. Además mediante el paquete VIM de R se puede realizar gráficas de visualización y comprobación para valores faltantes e imputados y otros (Calafati, 2017, pp. 1-75).

1.5 Estadística espacial

Estadística espacial es la reunión de un conjunto de metodologías propias para el análisis de datos que corresponden a la medición de variables en diversos sitios (puntos del espacio o agregaciones espaciales) de una región. Es decir la estadística espacial trata con el análisis de realizaciones de un proceso estocástico (campo aleatorio) $\{Z(s): s \in D\}$, en el que $s \in R^d$ representa una ubicación en el espacio euclidiano *d* - dimensional y *Z(s)* es una variable aleatoria en la ubicación *s*, tal que *s* varía sobre un conjunto de índices $D \subset R^d$. Estos índices *D*, pueden ser continuo, discreto o aleatorio (Cressie, 1993; citado en Giraldo, 2007).

1.5.1 Áreas de la estadística espacial

La Geoestadística espacial se subdivide en tres grandes áreas, cada una de ellas asociadas a las características del conjunto *D* de índices del proceso estocástico de interés (Giraldo, 2002, pp. 8-9).

Geoestadística: Las ubicaciones *s* provienen de un conjunto *D* continuo y son seleccionadas a juicio del investigador (*D* fijo*),* es decir a conveniencia o bajo algún esquema de muestreo probabilístico. En esta área el propósito fundamental es la interpolación y si no hay continuidad espacial pueden realizarse predicciones carentes de sentido. En esta área el investigador puede seleccionar los puntos en el espacio a conveniencia o bajo algún esquema de muestreo probabilístico.

Lattice (enmallados): Las ubicaciones *s* pertenecen a un conjunto *D* discreto y son seleccionados por el investigador (*D* fijo*).* Estas pueden estar regularmente o irregularmente espaciadas, lo que se denomina estructura de vecindad. Algunos ejemplos de datos en esta área son: tasa de accidentabilidad en sitios de una ciudad, producción de caña de azúcar por municipios, colores de los pixeles en la interpretación de imágenes de satélite, etc.

Patrones puntuales: Las ubicaciones *s* pertenecen a un conjunto *D* discreto o continuo y su selección no depende del investigador (*D* aleatorio*).* El propósito de análisis en esta área de estudio es determinar si la distribución de los individuos dentro de la región es aleatoria, agregada o uniforme.

1.5.2 Tipos de datos espaciales

Los datos espaciales pueden ser considerados como observaciones del proceso estocástico $\{Z(s): s \in D\}$. Estos datos siempre están correlacionados, a menos que estén suficientemente alejados. De acuerdo con Cressie (1990), se tienen los siguientes tipos de datos espaciales:

Datos geoestadísticos o georreferenciados (geostatistical data). - Son mediciones tomadas en puntos fijos con localizaciones continuas en el espacio, la variable medida puede ser continua como discreta. El análisis de este tipo de datos puede contemplar modelización o predicción de la variable en puntos donde no se ha muestreado.

Datos de rejilla o datos en un área (lattice data). - Son observaciones procedentes de un proceso aleatorio sobre una colección contable de regiones espaciales, que pueden estar regular o irregularmente distribuidas. Matemáticamente una rejilla es un conjunto de lados y vértices, es decir de un conjunto de índices de localizaciones con un conjunto asociado de vecinos.

Datos de procesos puntuales (point processes data). – Consiste en un número finito de localizaciones observadas en una región determinada, cuyo objetivo es el de conocer la variación de la intensidad de los eventos sobre la región de estudio y buscar modelos que permitan explicar o comprender el fenómeno (Ginzo, 2011; Cardona, 2015).

1.6 Geoestadística

La Geoestadística es una rama de la estadística que trata fenómenos espaciales, además comprende un conjunto de herramientas y técnicas que sirven para analizar y predecir los valores de una variable distribuida en el espacio o tiempo de una forma continua. Debido a su aplicación orientada a los SIG (Sistemas de Información Geográfica), también se la conoce como la estadística relacionada con los datos geográficos. Fue desarrollada originalmente para predecir la probabilidad de distribución en operaciones mineras, pero hoy en día es aplicada a diferentes disciplinas como: meteorología, control ambiental, geología de petróleos, epidemiología, hidrología, geoquímica, ecología, entre otros (Moral, 2004, p. 79).

La modelación espacial es la adición más reciente a la literatura estadística; cualquier disciplina que trabaja con datos recolectados en diferentes localizaciones espaciales necesita desarrollar modelos que indiquen cuando hay dependencia espacial entre las medidas de los diferentes sitios. Usualmente dicha modelación concierne con la predicción espacial, aunque existen otras áreas importantes. Por ello el proceso de estimación y modelación de la función que describe la correlación entre puntos del espacio es conocido como análisis estructural. Una vez realizado el análisis estructural, se realiza la predicción en sitios de la región no muestreados por medio de *kriging* por ejemplo, siendo este proceso el que calcula un promedio ponderado de las observaciones muestrales (Petitgas, 1996, pp. 113-114).

En resumen, todo trabajo geoestadístico tiene que llevar a cabo tres etapas (Moral, 2004, p. 79):

1) **Análisis exploratorio de los datos:** Se estudian los datos muestrales, sin tener en cuenta su distribución geográfica. En esta etapa se comprueba la consistencia de los datos mediante técnicas estadísticas tradicionales.
2) **Análisis estructural:** Se encarga del estudio de la continuidad espacial de la variable, se calcula el semivariograma u otra función que permita explicar la variabilidad espacial.
3) **Predicciones:** Estimaciones de la variable en los puntos no muestreados, considerando la estructura de correlación espacial seleccionada.

1.6.1 Variable regionalizada

Una variable regionalizada es aquella que está medida en el espacio de manera que presente una estructura de correlación. Además, se la define como un proceso estocástico con dominio en un espacio euclidiano d-dimensional $R^d, \{Z(s): s \in D \subset R^d\}$, donde este proceso es una colección de variables aleatorias indexadas; es decir, para cada *s* en el conjunto de índices *D*, *Z(s)* es una variable aleatoria, que en la práctica puede percibirse como una medición en un punto *s* de una región de estudio.

En el caso de que las mediciones *Z(s)* se realicen en una superficie, normalmente la variable aleatoria está asociada a ese punto del plano, donde *s* representa las coordenadas: planas o geográficas, y *Z* la variable en cada una de ellas. Por ejemplo estas variables aleatorias pueden representar la magnitud de una variable ambiental medida en un conjunto de coordenadas de la región de estudio (Giraldo, 2002; Ginzo, 2011).

1.6.2 Isotropía

Un proceso isotrópico es aquel cuya correlación entre los datos no depende de la dirección en la que está se calcule. La isotropía es estudiada mediante el cálculo de funciones de auto-covarianza o de semivarianza muestrales en varias direcciones (Giraldo, 2002; Cardona, 2014).

1.6.3 Estacionariedad

La variable regionalizada es estacionaria si su función de distribución conjunta o acumulada es invariante respecto a cualquier traslación del vector h, o lo que es igual, la función de distribución del vector aleatorio $\mathbf{Z}(s) = [Z(s_1), \dots, Z(s_n)]^t$ es idéntica a la del vector $\mathbf{Z}(s+h) = [Z(s_1+h), \dots, Z(s_n+h)]^t$, para cualquier h.

La hipótesis de estacionariedad establece el grado de homogeneidad espacial del fenómeno, además el tipo de estacionariedad que se asuma indica qué tipo de inferencia estadística puede realizarse con el modelo probabilístico.

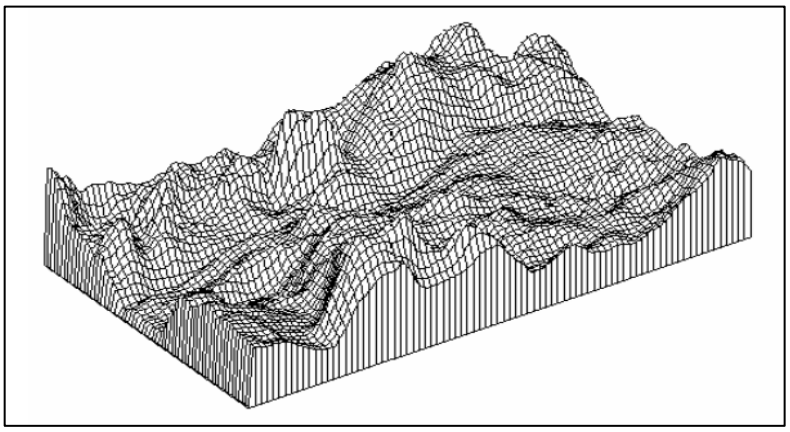


Figura 2-1: Representación de una superficie interpolada para una variable regionalizada estacionaria.

Fuente: (Giraldo, 2002).

1.6.3.1 Estacionariedad de Segundo Orden

Sea $\{Z(s): s \in D \subset R^d\}$ una variable regionalizada definida en un dominio D contenido en R^d, se dice que *Z(s)* es estacionario de segundo orden si cumple:

a) $E\big(Z(s)\big) = m, \quad \forall\, s \in D \subset R^d,$ con $m \in R.$

b) $Cov\big[Z(s_i), Z\big(s_j\big)\big] = C(h) < \infty.$

Estas dos condiciones implican que la media y la varianza son constantes en la región y que la covarianza depende solo de las distancias entre los sitios y no de su posición dentro del área de estudio (Giraldo, 2002; Cardona 2014).

1.6.3.2 Estacionariedad Débil o Intrínseca

Existen fenómenos físicos reales en los que la varianza no es finita. En estos casos se trabaja sólo con la hipótesis que pide que los incrementos $[Z(s) - Z(s+h)]$ sean estacionarios. Este supuesto es muy usado en la práctica del análisis geoestadístico, esto es:

a) $E\left(Z(s_i) - Z\big(s_j\big)\right) = 0, \quad \forall\, (s_i, s_j) \in D \subset R^d.$

b) $V\big[Z(s_i) - Z\big(s_j\big)\big] = E\big[Z(s_i) - Z\big(s_j\big)\big]^2 = 2\gamma(h), \;\; \forall\, \big(s_i, s_j\big) \in D \subset R^d$ y $h = \big\|s_i - s_j\big\|.$

A $2\gamma(h)$ se le denomina variograma y es la función comúnmente empleada para hacer estimación de la autocorrelación espacial.

Finalmente, una variable regionalizada será *no estacionaria* si $E[Z(s)] = m(s)$, es decir si su esperanza matemática no es constante (Giraldo, 2002; Cardona 2014).

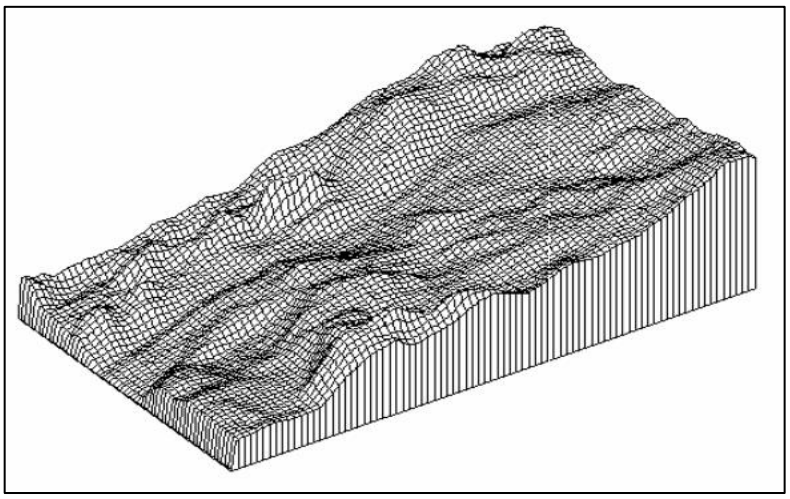


Figura 3-1: Representación de una superficie interpolada para una variable regionalizada no estacionaria.

Fuente: (Giraldo, 2002).

En casos prácticos resulta compleja la identificación de la estacionariedad, por lo que este requerimiento es estudiado de manera empírica empleando dispersogramas de los valores de la variable de interés respecto a las coordenadas de medición. (Giraldo, 2007, pp. 117-118).

1.6.4 Análisis estructural

El análisis estructural es uno de los tópicos más importantes de la geoestadística, consiste en estimar y modelar una función que refleje la correlación espacial de la variable regionalizada a partir de la adopción de una hipótesis adecuada, es decir que en dependencia de las características de estacionariedad del fenómeno se modelará la función de covarianza, semivarianza o correlograma.

Por su importancia, generalidad y dado que el semivariograma es la única que no requiere estimaciones de parámetros, en la práctica se empleó esta función para el proceso de estimación y modelación (Díaz, 2002, p. 19).

1.6.4.1 Semivariograma

El semivariograma es la herramienta central de la geoestadística, que permite analizar el comportamiento espacial de una variable sobre un área definida. Sus características son las siguientes:

- Es un estimador no paramétrico.
- No es necesario conocer la media de la variable para el cálculo de su función.
- Es empleada para tratar datos que presentan continuidad espacial (datos geoestadísticos).

Cuando se definió la estacionariedad débil se asumía que la varianza de los incrementos de la variable regionalizada era finita. La función denotada por $2\gamma(h)$ se le denomina variograma, y la mitad del variograma $\gamma(h)$, se conoce como la función de semivarianza que caracteriza las propiedades de dependencia espacial del proceso.

El estimador de momentos del semivariograma está definido como:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)}\left[(Z(s_i) - Z(s_j)\right]^2 \qquad (1.1)$$

donde $Z(s_i)$ es el valor de la variable en un sitio *s*, $Z(s_j)$ es otro valor muestral separado del anterior por una distancia h y $N(h)$ el número de parejas que se encuentran separadas por dicha distancia.

La función de semivarianza se calcula para varias distancias *h;* debido a la irregularidad en el muestreo y por ende en las distancias entre los sitios muestreados, se toman intervalos de distancia $\{(0, h].(h, 2h], (2h, 3h], \dots\}$, y el semivariograma empírico o experimental corresponde a una distancia promedio entre parejas de sitios dentro de cada intervalo y no a una distancia *h* específica, por ello el número de parejas de puntos *s* dentro de los intervalos no es constante. Para la interpretación del semivariograma experimental se parte del criterio de que a menor distancia entre los sitios mayor similitud o correlación espacial entre las observaciones (Giraldo, 2002; Ginzo, 2011, Cardona, 2015).

Dado que el semivariograma experimental es calculado para algunas distancias promedios particulares, es necesario el ajuste de modelos que generalicen lo observado en este semivariograma. Todos estos modelos tienen tres parámetros en común (Gráfico 1-1), los cuales se describen a continuación (Cardona, 2015, pp. 22-23):

- **Pepita (Nugget) (τ):** Representa una discontinuidad puntual del semivariograma en el origen. Puede ser debido a errores de medición en la variable o a la escala de la misma. En algunas ocasiones es indicativo de que parte de la estructura espacial se concentra a distancias inferiores a las observadas.
- **Meseta (Sill) (σ^2):** Es la varianza a priori, es el máximo valor que alcanza un semivariograma, ya que es un estimador de la varianza de las variables del proceso. También puede definirse como el límite del semivariograma cuando la distancia (*h)* tiende al infinito, indicando la escala bajo la cual los datos definen un proceso estacionario de segundo orden.
- **Rango (ϕ):** Corresponde a la distancia hasta la cual hay correlación espacial El rango se interpreta como la zona de influencia.

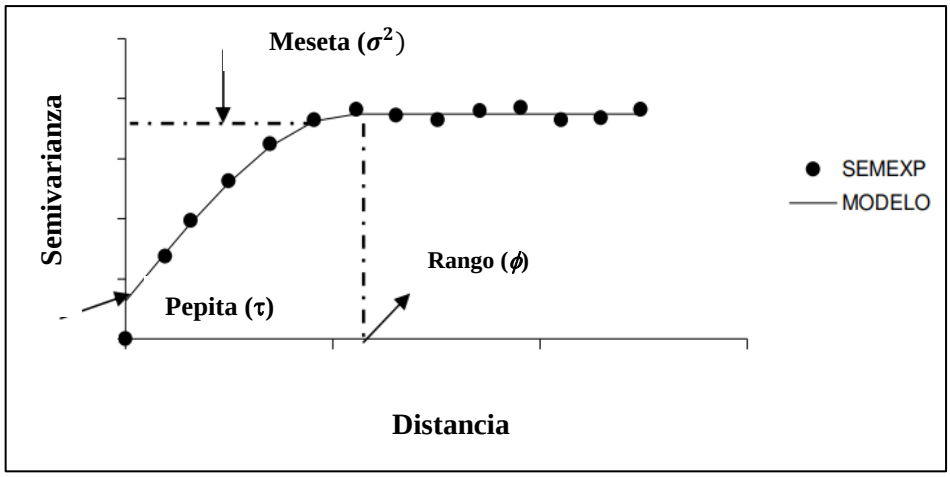


Gráfico 1-1: Comportamiento típico de un semivariograma y sus parámetros básicos.
Fuente: (Giraldo, 2002).

A continuación, se presenta algunos de los modelos teóricos de semivarianza usados en la práctica que pueden ajustarse al semivariograma experimental (Giraldo, 2012; Dueñas, 2017):

Modelo Esférico. - Presenta un crecimiento rápido cerca al origen (Gráfico 2-1), pero los incrementos marginales van decreciendo para distancias grandes, hasta que para distancias superiores al rango los incrementos son nulos. Su expresión matemática es la siguiente:

$$\gamma(h) = \begin{cases} 0 & si\ \ |h| = 0 \\ \sigma^2 \left[\dfrac{3|h|}{2\phi} - \dfrac{1}{2}\left(\dfrac{|h|}{2\phi}\right)^3 \right] & si\ 0 < |h| < \phi \\ \sigma^2 & si\ \ |h| > 0. \end{cases} \qquad (2.1)$$

Modelo Exponencial: Es el más usado y se aplica cuando la dependencia espacial tiene un crecimiento exponencial respecto a las distancias entre las observaciones. El valor del rango es

igual para la cual el semivariograma toma un valor igual al 95% de la meseta (Gráfico 2-1). Su expresión matemática es la siguiente:

$$\gamma(h) = \begin{cases} 0 & si \ \ |h| = 0 \\ \sigma^2 \left[1 - exp\left\{-\frac{3|h|}{\phi}\right\}\right] & si \ \ |h| > 0. \end{cases} \quad (3.1)$$

Modelo Gaussiano: Es similar al exponencial, es decir la dependencia espacial solo se desvanece en una distancia que tiende a infinito. Su distintivo es su forma parabólica cerca del origen (Gráfico 2-1). Su expresión matemática es la siguiente:

$$\gamma(h) = \begin{cases} 0 & si \ \ |h| = 0 \\ \sigma^2 \left[1 - exp\left\{-3\left(\frac{|h|}{\phi}\right)^2\right\}\right] & si \ \ |h| > 0. \end{cases} \quad (4.1)$$

Familia Matérn: Es una función positiva definida en términos de la función de la semivarianza, dada como:

$$\gamma(h) = \begin{cases} 0 & si \ \ |h| = 0 \\ \sigma^2 \left[1 - \left(2^{k-1}\Gamma(k)\right)^{-1} \left(\frac{|h|}{\phi}\right)^k K_k\left(\frac{|h|}{\phi}\right)\right] & si \ |h| > 0. \end{cases} \quad (5.1)$$

La familia Matérn corresponde al modelo exponencial cuando el parámetro de órdenes k=0.5 y Gaussiano cuando k tienda $\rightarrow \infty$ (Uribe, 2015, pp. 15-16).

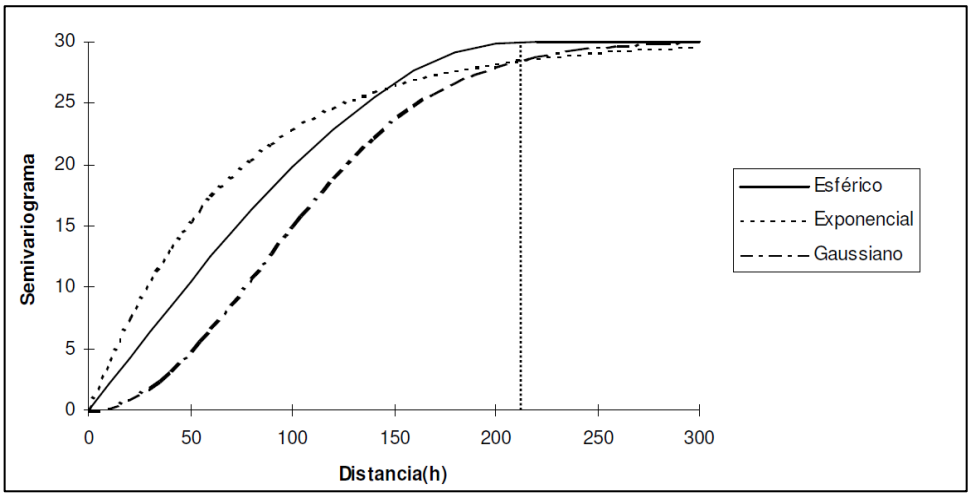


Gráfico 2-1: Comparación de los modelos esférico, exponencial y Gaussiano.

Fuente: (Giraldo, 2002).

1.6.5 Kriging: predicción e interpolación

Kriging procede del nombre del geólogo sudafricano D. G. Krige, cuyo trabajo en la predicción de reservas de oro realizada en la época de los cincuenta (1950), se considera pionero en los métodos de interpolación espacial. Es un nombre genérico adoptado en geoestadística para dar nombre a una metodología de predicción e interpolación espacial basada en una familia de algoritmos de regresión generalizada por mínimos cuadrados. Es un método muy similar a una regresión lineal múltiple aplicada en el contexto espacial, donde las variables aleatorias $Z(s)$ se comportan como variables de regresión y la variable aleatoria en el punto donde interesa la predicción (no muestreado) $Z(s_0)$es la variable dependiente (Cardona, 2015, p. 24-25).

El predictor kriging depende del modelo que se seleccione para la función aleatoria $Z(s)$. Por lo general, $Z(s)$ se descompone en dos componentes: tendencia y residual, tal como se presenta en la siguiente expresión:

$$Z(s) = m(s) + \varepsilon(s)$$

donde para $\varepsilon(s)$ el variograma o el covariograma se supone conocido.

Las variantes de kriging dependen del modelo de variograma que se adopte para la tendencia $m(s)$, entre ellos se considera (Ginzo, 2011; Cardona, 2015):

- Kriging simple (KS), supone $m(s) = m$, es decir que la media *m(s)* es conocida en todo el dominio.
- Kriging ordinario (KO), supone que la tendencia $m(s) = m$ es contante pero desconocida. Además, se ciñe a fluctuaciones locales de la media dentro de una vecindad W(s), dentro de la cual se puede considerar la media estacionaria.
- Kriging universal (KU), se considera que la media *m(s)* es una función que varía suavemente en todo el dominio D, donde la tendencia se modela generalmente mediante modelos de superficie que resultan ser combinaciones lineales de las coordenadas espaciales (latitud y longitud).

1.6.5.1 Kriging Ordinario (KO)

Uno de los intereses de la geoestadística es predecir $Z(s_0)$ en algún sitio $s_0 \notin \{s_1, \dots, s_n\}$ no muestreado a partir de un conjunto de observaciones de un atributo espacial $Z(s_1), \dots, Z(s_n)$, bajo las consideraciones anteriores, la metodología KO propone que el valor de la variable en el

sitio no muestreado puede predecirse como una combinación lineal de la n variables, como se muestra a continuación:

$$Z^*(s_0) = \sum_{i=1}^{n} \lambda_i \, Z(s_i) = \lambda_1 \, Z(s_1) + \cdots + \lambda_n \, Z(s_n) \qquad (6.1)$$

donde λ_i representa los pesos o ponderaciones de los valores de las variables en los sitios muestreados, dichos pesos se calculan en función de la distancia entre los puntos muestreados y el punto donde se lleva a cabo la predicción, cuya suma debe ser igual a uno para que la esperanza del predictor sea igual a la esperanza de la variable, esto se conoce como el requisito de insesgamiento, y en este caso $Z^*(s_0)$ sería el mejor predictor lineal que minimicen la varianza del error de predicción, es decir que minimice (Caballero, 2011; Ginzo, 2011):

$$V\big(Z^*(s_0) - Z(s_0)\big) \quad sujeto\ a \ \sum_{i=1}^{n} \lambda_i \ = 1.$$

Los pesos λ_i óptimos se determinan por la aplicación del método de LaGrange como técnica de optimización en conjunto con la determinación de la matriz de covarianzas.

1.6.5.2 Kriging Universal (KU)

En muchos casos para el kriging ordinario planteado en (6.1), la variable regionalizada no cumple con el supuesto de estacionariedad y presenta un tipo de tendencia, siendo $m(s)$ la función determinística que describe esta componente, más una estocástica de media cero. La tendencia se expresa como (Caballero, 2011; Ginzo, 2011):

$$m(s) = \sum_{l=1}^{p} a_l f_l(s)$$

donde las funciones $f_i(s)$ son conocidas y p es el número de términos empleados para ajustar $m(s)$. El predictor universal viene dado por:

$$Z^*(s_0) = \sum_{i=1}^{n} \lambda_i \, Z(s_i)$$

y será insesgado si:

$$\sum_{i=1}^{n} \lambda_i \, f_l(s_i) = f_l(s_0) \ \text{para todo } l = 1, \dots, p.$$

1.7 Análisis de Datos Funcionales

El análisis de datos funcionales (ADF) es una rama de las matemáticas, en concreto de la estadística, que estudia y analiza la información contenida en curvas, superficies, o cualquier otro elemento que generalmente varía en el tiempo, y surgen de manera natural en varias áreas donde se trabaja con grandes conjuntos de datos. En el ADF la unidad básica de información es la función completa más que un conjunto de valores. En general, cualquier observación que varíe en un continuo se puede considerar como dato funcional (Ramsay y Dalzell, 1991).

Los avances tecnológicos han hecho que el ADF se convierta en una disciplina emergente de la estadística con multitud de estudios y publicaciones en diferentes revistas de gran impacto. Las técnicas de ADF más utilizadas son el Análisis de Componentes Principales (ACPF), ANOVA y los modelos de regresión funcional, aunque en las últimas décadas se ha incursionado en técnicas de detección de atípicos, clúster, kriging, entre otros.

Según Ginzo (2011), Caballero (2011) y Cardona (2015) los problemas que enfrenta el ADF corresponden a los mismos que la estadística clásica, los cuales se categorizan de la siguiente manera:

a) Explorar y describir el conjunto de datos funcionales resaltando sus características más importantes.
b) Explorar y modelar la relación entre una variable dependiente y una independiente (modelos de regresión).
c) Métodos de clasificación Supervisado o no Supervisado de un conjunto de datos respecto a alguna característica.
d) Contraste, validación y predicción.

La aplicación de ADF tiene características deseables que no se podrían tener al realizar un análisis estadístico de datos individuales (clásico). Algunas de las principales características son:

- No se requieren supuestos paramétricos para la modelación.
- Toda técnica estadística descriptiva, inferencial, de clasificación, modelación o multivariada puede ser aplicada para datos funcionales.
- Por lo general, los procesos tienen alta afijación y poco ruido, además se puede suponer que el proceso que genera los datos es suave y continuo.

1.7.1 Definiciones

Según Ferraty y Vieu (2006, pp.5-7) definen una variable aleatoria funcional $\boldsymbol{\mathcal{X}}$ como una variable aleatoria que toma valores en un espacio de funciones, es decir, un espacio infinito dimensional (espacio funcional).

A continuación, se presentan entre otras, la definición del espacio funcional, en el cual se representan los datos funcionales.

Definición 1: Una variable aleatoria $\boldsymbol{\mathcal{X}}$ es una variable funcional si toma valores en $L^2(T)$. Una observación $\mathcal{X}$ de $\boldsymbol{\mathcal{X}}$ se denomina dato funcional.

Definición 2: Un conjunto de datos funcionales $S_n = \{\mathcal{X}_i\}_{i=1}^n$ (también denotado como $\mathcal{X}_i, \dots, \mathcal{X}_n$) es la observación de *n* variables funcionales $\boldsymbol{\mathcal{X}}_1, \dots, \boldsymbol{\mathcal{X}}_n$ con igual distribución que $\boldsymbol{\mathcal{X}}$.

Definición 3: Un dato funcional $\mathcal{X}_i(t), t \in T = [a, b] \subset \mathbb{R}$, es representado usualmente con un conjunto finito de pares $(t_i, \mathcal{X}_{ij})$, $t_j \in T$ y $j = 1, \dots, M$, donde M representa la cantidad de puntos en los cuales es observada la variable de interés y $y_{ij} = \mathcal{X}_i(t_j)$ (si no existe error observacional) o $y_{ij} = \mathcal{X}_i(t_j) + \varepsilon_j$ (en caso contrario).

Definición 4: Sea $L^2(T)$, el espacio de Hilbert separable definido por las funciones $f(t)$ de cuadrado integrable en el intervalo $T = [a, b] \subset \mathbb{R}$:

$$L^2(T) = \{f: T \to \mathbb{R}, tal\ que \int_T f(t)^2 dt < \infty\}.$$

Con un producto interno definido por:

$$\langle f, g \rangle = \int_T f(t)\, g(t) dt.$$

En general, un dato funcional es la observación de la variable aleatoria a lo largo de un intervalo continuo fijo (usualmente de tiempo o frecuencia).

1.7.2 Representación en funciones de una base

El primer paso para el tratamiento funcional consiste en reconstruir la forma funcional a partir de observaciones discretas, que permita su evaluación en cualquier intervalo de tiempo *t*, por ello el modo más usual de resolver este problema consiste en asumir una expansión de cada curva muestral en términos de una base de funciones y aproximar los coeficientes básicos utilizando un suavizado o interpolación.

Las bases de funciones son muy útiles en el proceso de obtener las trayectorias funcionales pues conservan muy bien la información de los datos, son flexibles y además optimizan el tiempo computacional necesario para realizar los procesos, por ello para obtener un buen resultado es necesario que las funciones básicas compartan características con las funciones a estimar.

Formalmente una base es un conjunto de funciones conocidas e independientes $\{\phi_k\}_{k\in N}$, tal que cualquier función puede ser aproximada, tan bien como se quiera, mediante una combinación lineal de *K* de ellas, de esta manera, la observación funcional se expresa de la siguiente forma:

$$\boldsymbol{\mathcal{X}}_i(t) = \sum_{k=1}^{K} a_{ik}\,\phi_k(t), \quad i = 1, \dots, N. \tag{7.1}$$

Donde:

a_{ik}: Son los coeficientes básicos en la nueva base, que se podrán obtener por ejemplo a partir de mínimos cuadrados.

K: Es el número de funciones de la base, parámetro que mide el grado de interpolación o suavizado de la función. Si *K* es bajo el modelo será manejable pero posiblemente se pierda información relevante, mientras que si *K* es alto habrá un problema de dimensión, aunque los datos estén bien representados.

Existen diferentes técnicas de aproximación, tales como interpolación o la proyección en un espacio finito-dimensional generado por una base de funciones, y entre ellas técnicas de estimación no paramétrica para obtener la verdadera forma funcional de las curvas y un gran número de bases a utilizar dependiendo de las características de las curvas y observaciones que se disponga. Las bases más comunes son base Fourier y B-Splines (Santofimia, 2011; Aguilera, 2009; Escudero, 2016).

1.7.2.1 Bases de Fourier

Uno de los sistemas de bases de funciones más antiguos y conocidos viene dado por las series de Fourier, formada por senos y cosenos, las funciones base son:

$$1, \sin(wt), \cos(wt), \sin(2wt), \cos(2wt), \ldots \quad (8.1)$$

definidas por:

$$\phi_0(t) = 1$$
$$\phi_{2r-1}(t) = \sin(rwt)$$
$$\phi_{2r}(t) = \cos(rwt).$$

Luego, una función $\mathcal{X}(t)$**,** se aproxima por el valor de $\widehat{\mathcal{X}}(t)$**,** dado por la siguiente expresión:

$$\widehat{\mathcal{X}}(t) = c_0 + c_1 \sin(wt) + c_2 \cos(wt) + c_3 \sin(2wt) + c_4 \cos(2wt) + \cdots \quad (9.1)$$

que forma una base periódica, donde el parámetro *w* se conoce como la frecuencia de la señal. Determina el periodo $\frac{1}{r\omega}$ si se mide en hertzios (Hz) o $\frac{2\pi}{r\omega}$ en radianes por segundo.

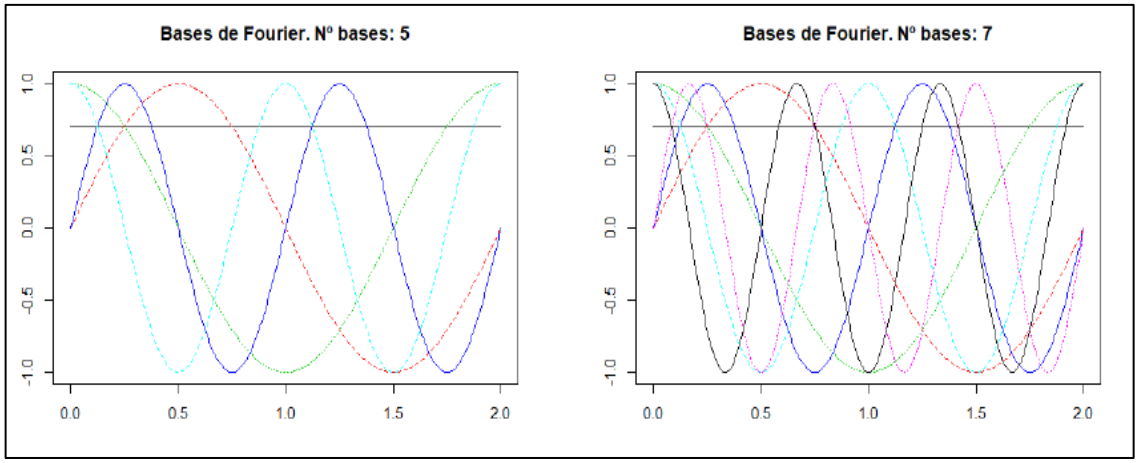


Gráfico 3-1: Aproximación mediante base Fourier para 5 y 7 funciones base.

Fuente: (Millán, 2017).

Tradicionalmente ha sido utilizada para series temporales largas debido a que la transformación rápida de Fourier (FFT) permite calcular estos coeficientes de manera eficiente (en O (*p log p)* operaciones) cuando el número de puntos *p* es potencia de 2 y los instantes en los que se obtiene las observaciones están igualmente espaciados.

Al ser una base de funciones periódicas se ajustará mejor a problemas en los que los datos muestren una cierta periodicidad. Esta representación es especialmente útil para funciones estables, sin grandes variaciones y cuya curvatura es la misma aproximadamente en todo el intervalo (más o menos constante). En cambio, son inadecuadas si se sospecha que existe algún grado de discontinuidad en las trayectorias a aproximar (Torecilla, 2010; Aguilera, 2013; Millán, 2017).

1.7.2.2 Bases B-Splines

La base B-Splines son trozos de polinomios de grado p conectados entre sí, más utilizados en el caso de datos no periódicos reemplazando de alguna manera a los polinomios que quedan contenidos en ellos. A diferencia de otros métodos no padecen de los efectos frontera comunes, como en algunos suavizados tipo núcleo, en donde al extender la curva ajustada fuera del dominio de los datos ésta tiende hacia cero. Posee las siguientes características:

- Consiste en $p + 1$ trozos de polinomios de orden p que se unen en p nodos internos.
- Las derivadas hasta el orden $p - 1$ son continuos en los puntos de unión.
- B-Splines es positivo en el dominio expandido por $p + 2$ nodos y cero en el resto.
- Salvo los extremos, se solapa con $2p$ trozos nodos y cero en el resto.
- Para cada x, son no nulos $p + 1$ B-Splines.

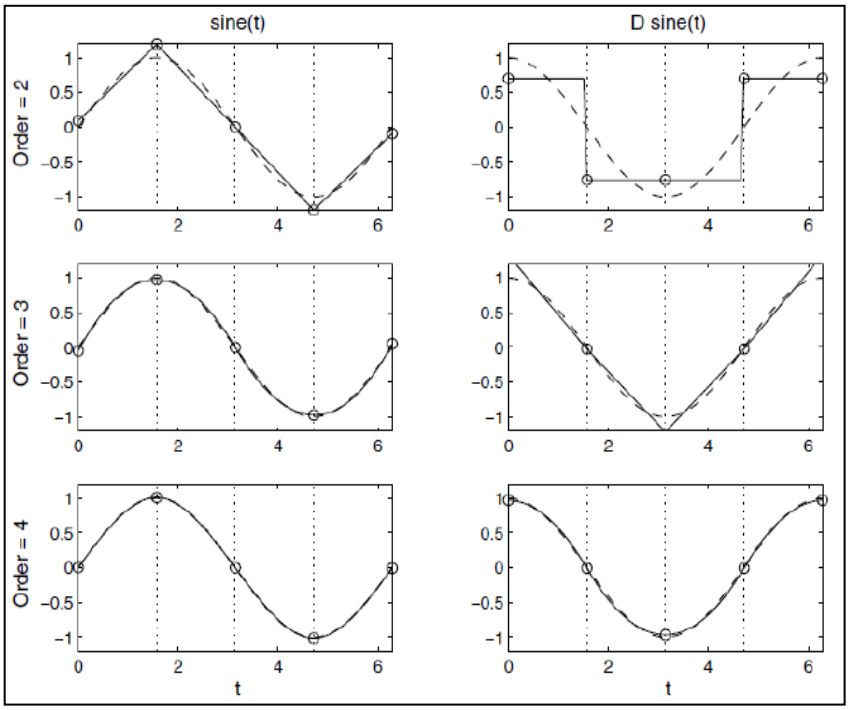


Gráfico 4-1: Aproximación mediante base B-Splines de orden 2, 3 y 4.

Fuente: (Torrecilla, 2010).

La función que denota el valor del j-ésimo B-Spline de grado *p* en el punto *t* está dado por la forma $x(t) = \sum_{j=1}^{n} c_j \, B_j(t, p)$. Estas bases se pueden ser calcular fácilmente y de forma numéricamente estable con el algoritmo de Boor, cuya fórmula recursiva se define de la siguiente manera:

$$B_{j,1} = \begin{cases} 1 & t_{j-2} \leq t \leq t_{j-1} \\ 0 & en\ caso\ contrario \end{cases}$$

$$B_{j,p+1}(t) = \frac{t - t_{j-2}}{t_{j+p-2} - t_{j-2}} B_{j,p}(t) + \frac{t_{j+p-1} - t}{t_{j+p-1} - t_{j-1}} B_{j+1,p}(t) \qquad (10.1)$$

con $q = 1, 2, \ldots \quad ; \ j = -1, 0, \ldots, s - p + 4.$

Las funciones Splines tienen un mejor comportamiento local que la trigonométricas y polinómicas, de ahí su popularidad, siendo las más utilizadas las de grado *p=3*. Estas funciones base se denominan B-Splines cúbicos, utilizados para la aproximación de curvas muestrales regulares y que tienen soporte compacto. De manera que los B-Splines cúbicos se denotarán como:

$$B_{j,4}(t) = B_j(t) \, , \quad j = -1, 0, \ldots, p + 1.$$

Aunque existen otros métodos de suavización, este es el más utilizado y está implementado en varios lenguajes como R, que combinan la eficiencia computacional de los polinomios con una mayor flexibilidad, que varias veces hace que la *K* necesaria sea pequeña para obtener los mejores resultados (Aguilera, 2009; Torrecilla, 2010; Escudero, 2016).

1.7.3 Suavización de datos funcionales

Para convertir los datos recogidos en forma discreta a una función, se utiliza una técnica conocida como suavizado y que trata de ajustar los datos a una base de funciones seleccionada, permitiendo eliminar el ruido registrado al obtener las observaciones.

Existen algunos métodos de suavizado entre los que destacan el suavizado por mínimos cuadrados, mediante kernels o por regularización.

- **Suavizado por mínimos cuadrados**

El objetivo es ajustar una curva $x(t)$ a las observaciones discretas $y_j, j = 1, \dots, p$; usando una combinación lineal de funciones base para $x(t)$ de la forma:

$$x(t) = \sum_{k=1}^{K} a_k \, \phi_k(t) = \boldsymbol{c}^t \boldsymbol{\phi}.$$

Los vectores **c** y $\boldsymbol{\phi}$ son de longitud *K* y contienen los coeficientes a_k y las funciones base ϕ_k respectivamente.

Ajuste por mínimos cuadrados no ponderados

Los coeficientes a_k se determinan por el criterio de mínimos cuadrados, dado por:

$$MSSE(y|a) = \sum_{j=1}^{p} \left[y_j - \sum_{k=1}^{K} a_k \phi_k(t_j) \right]^2 . \qquad (11.1)$$

Esta aproximación es adecuada cuando los residuos ε_j son independientes e idénticamente distribuidos con media cero y varianza constante.

Ajuste por mínimos cuadrados ponderados

Cuando las varianzas de los ε_j no son constantes o no están idénticamente distribuidos se debe aportar un peso diferente a los distintos residuos. Por ello, una extensión del criterio de mínimos cuadrados viene de la forma:

$$SMSSE(y|a) = (\boldsymbol{y} - \boldsymbol{\Phi c})^{\boldsymbol{t}} \mathbf{W} (\boldsymbol{y} - \boldsymbol{\Phi c}). \qquad (12.1)$$

- **Suavizado mediante Kernels**

El caso más simple y clásico de un estimador que hace uso de pesos locales es el estimador kernel. La estimación en un punto dado es una combinación lineal de las observaciones locales, para algunas funciones de peso S_j definidas adecuadamente.

$$\hat{x}(t) = \sum_{j=1}^{p} S_j(t) \, y_j.$$

El estimador kernel más popular es el estimador Nadaraya-Watson, el cual se construye usando los pesos:

$$S_j(t) = \frac{Kern\left[t_j - t/h\right]}{\sum_r Kern[t_r - t/h]}. \tag{13.1}$$

La principal ventaja de los kernels es la computación rápida, sin embargo, su principal problema es que no aproxima bien ceca de los extremos, sobre todo *h* si es grande en relación a la tasa de muestreo (Aguilera, 2009; Millán, 2017).

1.7.4 Elección del número de funciones base

Para elegir el número de funciones base *K*, se debe ten en cuenta que mayor sea *K* mejor será el ajuste, sin embargo, se corre el riesgo de ajustar el ruido que puede distorsionar los resultados fácilmente. Mientras que en el caso de *K* demasiado bajo se puede estimar una función muy suave y que pierda características importantes.

Para elegir el número de funciones base, existen varios métodos como la Validación Cruzada (VC) y Validación Cruzada Generalizada (VCG).

Validación Cruzada:

$$VC(v) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{-i}^{v})^2\, w_i \tag{14.1}$$

donde $\hat{y}_{-i}^{v}$, indica el estimador basado en omitir el par (t_i, y_i) y w_i el peso en el punto t_i.

Validación Cruzada Generalizada:

$$VCG(v) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{-i}^{v})^2\, w_i\, \Xi(v) \tag{15.1}$$

donde $\Xi(v)$ indica el tipo de función penalizadora (Febrero-Bande y Oviedo de la Fuente, 2012; Aguilera, 2013).

1.7.5 Análisis descriptivo funcional

En Ramsay y Silverman (2005) se considera la aplicación de los estadísticos descriptivos funcionales de manera similar a los estadísticos clásicos, ya que están definidos con el mismo criterio de medición. A partir de un conjunto de datos funcionales $\{\mathcal{X}_1, \dots, \mathcal{X}_n\}$, definidos en $t \in T \subset \mathbb{R}$, las correspondientes funciones muestrales descriptivas están dadas por las siguientes expresiones:

Media funcional:

$$\overline{\boldsymbol{X}}(t) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i(t). \tag{16.1}$$

Varianza funcional:

$$Var\big(\boldsymbol{X}(t)\big) = \frac{1}{n-1}\sum_{i=1}^{n}\big[\boldsymbol{X}_i(t) - \overline{\boldsymbol{X}}(t)\big]^2. \tag{17.1}$$

Desviación estándar:

$$D.E\big(\boldsymbol{X}(t)\big) = \sqrt{Var(\boldsymbol{X}(t))}. \tag{18.1}$$

Covarianza:

$$Cov(\boldsymbol{X}(t_1), \boldsymbol{X}(t_2) = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\boldsymbol{X}_i(t_1) - \overline{\boldsymbol{X}}(t_1)\Big)\Big(\boldsymbol{X}_i(t_2) - \overline{\boldsymbol{X}}(t_2)\Big). \tag{19.1}$$

Correlación:

$$Corr\big(\boldsymbol{X}(t_1), \boldsymbol{X}(t_2)\big) = \frac{Cov\big(\boldsymbol{X}(t_1), \boldsymbol{X}(t_2)\big)}{\sqrt{Var\big(\boldsymbol{X}(t_1)\big)Var\big(\boldsymbol{X}(t_2)\big)}}. \tag{20.1}$$

1.7.6 Datos funcionales atípicos

En ADF, se considera que una curva es un outlier o atípico si ha sido generado por un proceso estocástico con una distribución diferente que el resto de las curvas, las cuales pueden ser causadas por errores de medición o ser sospechosas de ser erróneas porque no siguen el mismo patrón que la mayoría de las curvas.

Las observaciones funcionales pueden desviarse de la mayoría de las curvas en diferentes formas:

- **Outliers aislados. -** Presentan un comportamiento extraño durante un intervalo breve de tiempo.
- **Outliers persistentes. -** Presentan un comportamiento extraño durante la mayor parte de su dominio. Dentro de este grupo se encuentran 3 tipos de outliers:
  Desplazados: Presentan igual forma que la mayor parte de las curvas, pero de manera desplazada.
  De la Forma: Presentan una forma distinta al resto de las curvas, sin ser perceptibles en ningún instante de tiempo.
  En Amplitud: Presentan la misma forma que el resto de curvas, pero difieren en su escala.

También se considera como datos atípicos funcionales aquellos valores "significativamente" pequeños en profundidad, que difieren en magnitud o forma del resto de las curvas. La mediana funcional de la muestra de curvas, tiene el mayor valor de profundidad. Por esta razón, el análisis de profundidad es uno de los métodos para detectar outliers.

Actualmente existen múltiples herramientas para visualizar datos funcionales como el gráfico de arcoíris, bagplots y boxplots, que sirven como métodos gráficos de análisis. Son muy útiles ya que facilitan el hallazgo de características que podrían no ser evidentes con el uso de estadísticos y modelos matemáticos, como lo es la detección de outliers funcionales con gran velocidad de cálculo y precisión (Millán, 2017, pp. 49-53).

1.7.6.1 BAGPLOT para datos funcionales

Los bagplots o diagramas de mochila fueron propuestos inicialmente por Rousseeuw, Ruts y Tukey en 1999, como una extensión al caso multivariante del boxplot. Como un boxplot univariado el bagplot bivariado tiene un punto central, en este caso la mediana de Tukey, una región interna (*bag*), y una región externa (*fence*), además de los valores atípicos. El *bag* es la

región profunda más pequeña que contiene al menos el 50% de las curvas centrales, y la *fence* que contiene las observaciones no consideradas atípicas.

El bagplot se construye de la siguiente manera:

- Sea D_k número de puntos muestrales. Se calcula el valor de *k* tal que $D_k \leq [n/2] \leq D_{k-1}$
- Luego se interpola linealmente entre D_k y D_{k-1} (relativo al punto T* correspondiente a la mediana muestral). Esta región convexa *B*, es la mochila y contiene el 50% de los puntos con mayor profundidad.
- La valla se obtiene inflando *B* (en relación a T*) por un factor que usualmente es ρ = 3 (valor experimental obtenido a través de simulaciones), aunque también se usa ρ = 2.58. Los puntos fuera de la valla son considerados como outliers.

Una propiedad curiosa que posee el diagrama de mochila (bagplot), es que, debido a la invariabilidad de la profundidad de Tukey por transformaciones afines, si trasladamos o rotamos la muestra, el bagplot no cambia.

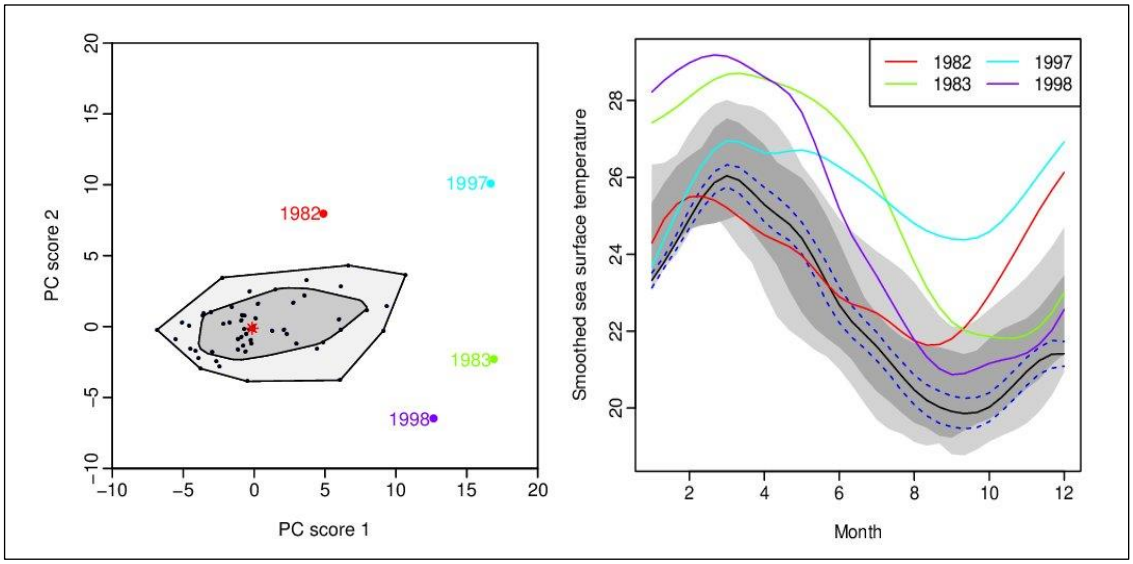


Gráfico 5-1: BAGPLOT Funcional y Bivariado.

Fuente: (Hyndman y Shang, 2010)

En el caso funcional no es más que establecer un mapeo entre los scores de las componentes principales funcionales y las curvas. Sea $\{y_i(t) = 1, 2, \dots, n\}$ un conjunto de curvas, se calcula las dos primeras componentes principales y se proyecta sobre $\mathbb{R}^2$. En el caso de trasladar los resultados a curvas nos encontraremos con dos regiones (interna y externa) y la curva mediana que es la curva con mayor profundidad (Hyndman y Shang, 2010; Pérez, 2018).

Medida de Profundidad de Tukey (semiespacio)

La profundidad del semiespacio de un punto $x \in \mathbb{R}^d$ con respecto a una muestra $x_1, \dots, x_n$ también en $\mathbb{R}^d$ es la menor fracción de puntos de la muestra que hay en un semiespacio cerrado que contenga ha x, es decir:

$$H\,D_n(x) = \frac{min_{u\in\mathbb{R}^d}\#\{i: \langle x_i, u\rangle \geq \langle x, u\rangle\}}{n}.$$

Esta medida de profundidad puede aplicarse a cualquier espacio de dimensión finita, donde en el caso 1 coincide con el orden natural $\mathbb{R}$, mientras que cuando d=2 se tiene una muestra en el plano. Se toma un $x \in \mathbb{R}^2$ y se calcula todos los hiperplanos que pasan por *x (*en este caso serán rectas). Para cada hiperplano, se cuenta los puntos que hay en un semiespacio u otro y se selecciona el mínimo, para posteriormente dividir por el tamaño muestral y obtener $H\,D_n(x)$. La mediana será el valor más profundo, el cual maximice $H\,D_n(x)$ (Pérez, 2018, p. 16).

1.7.7 Análisis de la Varianza Funcional

El método de Análisis de la Varianza Funcional (FANOVA) considera L grupos de funciones aleatorias independientes $\mathcal{X}_{ij}(t)\ con\ i = 1, \dots, L\ ;\ j = 1, \dots, n_i$ definidas en un intervalo compacto $T = [a, b]$. Se denota a $\bar{X}_i$ como la media funcional muestral para el grupo *i*, $t \in T$.

La hipótesis para el modelo FANOVA para *k* muestras independientes de datos funcionales es probar que:

$$\begin{aligned} H_0 &: \quad \mu_1(t) = \mu_2(t) = \cdots = \mu_L(t) \\ H_1 &: \quad \exists\, k, m \quad \mu_k(t) \neq \mu_m(t). \end{aligned}$$

El estadístico de prueba es análogo al F clásico para el ANOVA de una vía, donde el numerador mide la variabilidad externa entre las muestras y el denominador la variabilidad interna. La expresión del estadístico viene dada por:

$$F_n = \frac{\sum_{i=1}^{L} \frac{n_i \|\bar{X}_{i.} - \bar{X}_{i..}\|^2}{L-1}}{\sum_{i=1}^{L} \frac{\left\|\bar{X}_{ij} - \bar{X}_{i.}\right\|^2}{n-L}}. \qquad (21.1)$$

Donde:

$$\bar{X}_{i.}(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}(t)$$

$$\bar{X}_{..}(t) = \frac{1}{n}\sum_{i=1}^{L} n_i X_{i.}(t)$$

$$n = \sum_{j=1}^{n_i} n_i$$

$$\|x\| = \left(\int_a^b x^2(t)dt\right)^{1/2}.$$

La prueba F es sólo una entre varias posibles maneras de formalizar la idea de rechazar Ho. Para contrastar la hipótesis nula del FANOVA de igualdad de funciones medias, se utiliza el siguiente estadístico que mide la variabilidad entre grupos:

$$V_n = \sum_{k<j} n_i \left\|\bar{X}_k - \bar{X}_j\right\|. \tag{22.1}$$

Para calcular el valor crítico del estadístico V_n, se utiliza técnicas de remuestreo dado que no hay una distribución de referencia, lo que evita el requerimiento de la hipótesis de homocedasticidad de un ANOVA tradicional entre los grupos en estudio. De esta manera, se rechaza la hipótesis nula si el valor observado de V_n de la muestra es mayor que el valor crítico $V_\propto$ calculado a partir de una aproximación por Monte Carlo mediante el remuestreo de bootstrap (Cuevas et al., 2004; Escudero, 2016).

1.8 Kriging Ordinario para datos funcionales

Actualmente varias técnicas se han extendido en geoestadística para el análisis de datos funcionales, que permiten realizar la predicción espacial. Una primera aproximación al problema de predicción se lo conoce como kriging funcional, donde la curva a predecir resulta ser una combinación lineal de las curvas observadas, y los coeficientes son números reales. La técnica más utilizada en este contexto es el OKFD que permite predecir $\mathcal{X}_{s0}(t)$, el valor de un proceso aleatorio funcional en s_0 , donde s_0 es una ubicación sin muestrear (localización geográfica), para un conjunto de instantes de tiempo *t.*

En este contexto se ha desarrollado dos enfoques principales: el enfoque paramétrico de Goulard y Voltz (1993) y el enfoque no paramétrico propuesto por Giraldo y colaboradores (2011). Este último está implementado de acuerdo a las tendencias actuales del ADF.

Sea $\boldsymbol{\mathcal{X}}_s(t)$: $s \in D \subseteq \mathbb{R}^d, t = [a,b] \subset \mathbb{R}$, generalmente *d=2* un proceso o campo aleatorio funcional tal que $\boldsymbol{\mathcal{X}}_s(t)$ es una variable funcional para cualquier $s \in D$. Sean $s_1, \dots, s_n$ puntos arbitrarios en *D*, y asumiendo que podemos observar una realización del proceso aleatorio funcional $\boldsymbol{\mathcal{X}}_s(t)$, en *n* sitios $\mathcal{X}_{s1}(t), \dots, \mathcal{X}_{sn}(t)$.

Por lo general, se supone que el proceso aleatorio funcional es estacionario de segundo orden e isotrópico, es decir las funciones de media y varianza son constantes y la covarianza depende solo de la distancia entre los puntos de muestreo (sin embargo, esta metodología también se desarrolla sin asumir estas condiciones):

a) $E\left(\boldsymbol{\mathcal{X}}_s(t)\right) = m(t), \quad \forall\, t \in T,\ s \in D$

b) $V\left(\boldsymbol{\mathcal{X}}_s(t)\right) = \sigma^2(t), \quad \forall\, t \in T, \quad s \in D$

c) $Cov\left(\boldsymbol{\mathcal{X}}_{si}(t), \boldsymbol{\mathcal{X}}_{sj}(t)\right) = C(h,t), \quad \forall\, s_i, s_j \in D \in T \ \ donde\ h = \left\|s_i - s_j\right\|$

d) $\frac{1}{2}V\left(\boldsymbol{\mathcal{X}}_{si}(t), \boldsymbol{\mathcal{X}}_{sj}(t)\right) = \gamma(h,t) = \gamma_{s_i s_j}(t) \ \ \forall\, s_i, s_j \in D \in T \ \ donde\ h = \left\|s_i - s_j\right\|.$

La función $\gamma(h,t)$, es una función *h*, denominada semivariograma de $\boldsymbol{\mathcal{X}}(t)$ (Giraldo, 2009; Ginzo, 2011, Giraldo, 2012).

1.8.1 Predicción y estimación de parámetros

En la familia de predictores lineales, el predictor OKFD se define por:

$$\widehat{\boldsymbol{\mathcal{X}}}_{s0}(t) = \sum_{i=1}^{n} \lambda_i \boldsymbol{\mathcal{X}}_{si}(t) \quad , \lambda_1, \dots, \lambda_n \in \mathbb{R} \tag{23.1}$$

donde λ_i son los coeficientes que muestran la influencia de las curvas que están alrededor de la localización no muestreada donde se llevara a cabo la predicción.

El predictor (23.1) tiene la mima expresión que el predictor del kriging ordinario clásico (6.1), con la diferencia de que considera curvas en lugar de variables. Cada curva es un dato completo, y este enfoque la trata como una entidad singular, donde las curvas de las ubicaciones más cercanas al punto de predicción naturalmente tendrán más influencia que las más lejanas. Este es el primer paso en el modelado de datos espaciales funcionales.

Tanto en la geoestadística univariada como multivariada el mejor predictor lineal insesgado (BLUP) se obtiene minimizando la varianza del error de predicción, en el segundo caso es para *n* variables en una ubicación no muestreada s_0, dada por:

$$\sigma^2{}_{s0} = \sum_{i=1}^{n} V\left(\hat{Z}_i(\boldsymbol{s}_0) - Z_i(s_0)\right) \quad , \lambda_1, \dots, \lambda_n \in \mathbb{R} \tag{24.1}$$

sujeta a restricciones que garanticen la condición de insesgadez, esto es, minimizar la traza de la matriz del error cuadrático medio.

Extendiendo este criterio al contexto funcional, la sumatoria es reemplazada por una integral, y para encontrar el BLUP, los *n* parámetros en el predictor kriging de $\boldsymbol{\mathcal{X}}_{s0}$ están dados por el siguiente problema de optimización:

$$\underset{\lambda_1, \dots, \lambda_n}{min} \int_T (\widehat{\boldsymbol{\mathcal{X}}}_{s0}(t) - \boldsymbol{\mathcal{X}}_{s0}(t))dt \quad \text{donde} \quad \sum_{i=1}^{n} \lambda_i = 1 \tag{25.1}$$

donde $\sum_{i=1}^{n} \lambda_i = 1$ son las restricciones de insesgadez.

Luego de la manipulación algebraica se obtiene la función objetivo, que se escribe como:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \int_T (C_{ij}(h,t)dt + \int_T \sigma^2(t)dt - 2\sum_{j=1}^{n} \int_T (C_{i0}(h,t)dt + 2\mu\left(\sum_{j=1}^{n} \lambda_i - 1\right) \tag{26.1}$$

donde μ es el multiplicador de Lagrange usado para tener en cuenta la restricción de insesgadez.

Minimizando la ecuación anterior con respecto a $\lambda_1, \dots, \lambda_n$ y μ, se obtiene *n+1* ecuaciones, que se expresan de manera matricial, y mediante los supuestos de estacionariedad para obtener estimaciones basadas en la traza del variograma, se tiene:

$$\int_T C_{ij}(t)dt = \int_T \sigma^2(t) - \int_T \gamma_{s_i s_0}(t)dt \tag{27.1}$$

que reemplazando en la forma matricial de las *n+1* ecuaciones, por la ecuación anterior, se obtiene el sistema lineal:

$$\begin{pmatrix} \int_T \gamma_{s_1 s_1}(t)dt & \dots & \int_T \gamma_{s_1 s_n}(t)dt & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \int_T \gamma_{s_n s_1}(t)dt & \cdots & \int_T \gamma_{s_n s_n}(t)dt & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \int_T \gamma_{s_0 s_1}(t)dt \\ \vdots \\ \int_T \gamma_{s_0 s_1}(t)dt \\ 1 \end{pmatrix} \tag{28.1}$$

donde la función $\gamma(h) = \int_T \gamma_{s_i s_j}(t)dt$, $h = \|s_i - s_j\|$, se denomina traza-variograma, y corresponde a la extensión del semivariograma al caso en que las mediciones son curvas.

Al resolver este sistema lineal, se obtienen los pesos λ_i óptimos para la predicción, pero para ello se necesita su estimador. La predicción del OKFD basado en el traza-semivariograma está dado por:

$$\sigma^2{}_{s0} = \int_T V(\widehat{\boldsymbol{X}}_{s0}(t) - \boldsymbol{X}_{s0}(t))dt = \sum_{i=1}^{n} \lambda_i \int_T \gamma_{s_i s_0}(t)dt - \mu. \qquad (29.1)$$

El parámetro definido en esta ecuación es considerado como una medida de incertidumbre global, en el sentido de que es una versión integrada de la clásica varianza de predicción puntual del kriging ordinario. Por esta razón, su estimación no puede usarse para obtener intervalos de confianza para la curva predicha, ya que el caso espacial-funcional el objetivo es predecir una función y no una sola variable en una ubicación no muestreada (Giraldo, 2009; Ginzo, 2011).

1.8.2 Estimación de la traza del semivariograma

Para resolver el sistema de la ecuación 28.1, se necesita un estimador de traza-semivariograma. Dado que se asume que $\boldsymbol{X}_s(t)$ es una función con media constante *m* sobre D,

$$V\left(\boldsymbol{X}_{si}(t) - \boldsymbol{X}_{sj}(t)\right) = E\left[\left(\boldsymbol{X}_{si}(t) - \boldsymbol{X}_{sj}(t)\right)^2\right]$$

y usando el teorema de Fubini,

$$\gamma(h) = \frac{1}{2}E\left[\int_T \left(\boldsymbol{X}_{si}(t) - \boldsymbol{X}_{sj}(t)\right)^2 dt\right], \qquad \text{para } s_i, s_j \in D \text{ con } h = \|s_i - s_j\|$$

se obtiene mediante una adaptación del modelo clásico de los momentos (MoM) para esta cantidad el siguiente estimador:

$$\gamma(h) = \frac{1}{2|N(h)|}\sum_{i,j \in N(h)} \int_T \left(\boldsymbol{X}_{si}(t) - \boldsymbol{X}_{sj}(t)\right)^2 dt. \qquad (30.1)$$

Donde:

$N(h) = \{(s_i, s_j): \|s_i - s_j\| = h\}$, es el número de parejas de sitios separados por *h.*

$|N(h)|$: El número de elementos distintos en $N(h)$.

Debido a la irregularidad de los datos espaciales, generalmente, no hay observaciones suficientes separadas exactamente a una distancia h, *e*ntonces $N(h)$ se modifica por $\{(s_i, s_j): \|s_i - s_j\| \in (h - \varepsilon,\ h + \varepsilon\}$, con $\varepsilon > 0$ un valor pequeño.

Una vez que se haya estimado la traza - semivariograma para una secuencia de *K* valores h_k, se propone un ajuste paramétrico del modelo $\gamma_\propto(h)$ en los puntos $(h_k, \hat{\gamma} h_i), k = 1, \dots, K$, como si se obtuviera bajo los argumentos de la geoestadística clásica con los modelos esférico, Gaussiano, exponencial y Matérn, generalmente, este tipo de ajuste, se hace por mínimos cuadrados ordinarios o mínimos cuadrados ponderados, teniendo en cuenta que este ajuste paramétrico es siempre válido porque sus propiedades son las de una variograma paramétrico ajustado por un conjunto geoestadístico univariante (Giraldo, 2009; Ginzo, 2011).

1.8.3 Criterio de evaluación

El criterio de evaluación en la selección del modelo adecuado consiste en dividir un conjunto de datos en dos partes: una muestra de entrenamiento (estimación) y otra muestra de test (validación), llamada validación cruzada. Sin embargo, este procedimiento no es eficiente, si el tamaño de la muestra no es grande.

En el contexto de datos funcionales espacialmente correlacionados, la técnica de validación cruzada de dejar un dato fuera, se denomina Validación Cruzada Funcional (VCF) y trabaja del siguiente modo: cada curva o función de la región en estudio se saca del conjunto de datos funcionales y mediante una función de suavización, se predice la curva entera de la función $\boldsymbol{\mathcal{X}}_{s0}(t)$ en el punto $\boldsymbol{s}_0$ no muestreado, mediante Kriging funcional como predictor. Por ello se calcula el SSE (Suma de cuadrados del error) de VCF, de la siguiente manera:

$$SSE_{VCF} = \sum_{i=1}^{n} SSE_{VCF}(i) = \sum_{i=1}^{n} \left\| \boldsymbol{\mathcal{X}}_{si}(t_j) - \widehat{\boldsymbol{\mathcal{X}}}_{si}^{(i)}(t_j) \right\|. \qquad (31.1)$$

Donde: $\widehat{\boldsymbol{\mathcal{X}}}_{si}^{(i)}(t_j)$ es la predicción en s_i evaluada en $t_j, j = 1, \dots, M$, sacando el sitio s_i temporalmente fuera de la muestra.

Además, se puede emplear gráficas con los datos de la curva estimada del sitio no muestreado contra sus valores reales (Giraldo, 2009; Ginzo, 2011).

CAPITULO II

2 MARCO METODOLÓGICO

2.1 Tipo y diseño de investigación

La investigación según:

- El método de investigación es cuantitativa, ya que la información se concentra en la variable estadística temperatura del aire (ºC).
- El objetivo es aplicada, ya que a partir de los conocimientos adquiridos se dará solución al problema específico a través técnicas estadísticas y geoestadísticas.
- El nivel de profundización en el objeto de estudio, es exploratorio ya que se aborda en un nuevo campo estadístico como es el análisis de datos funcionales aplicado a la geoestadística.
- El tipo de inferencia es inductivo, debido a que con la información que monitorea las estaciones se estimará un modelo de predicción del comportamiento de la temperatura del aire en sitios no muestreados.
- La manipulación de la variable es no experimental, debido a que la temperatura del aire es un fenómeno natural donde no se tiene control absoluto de la misma, es decir se observa el fenómeno tal y como se da en su contexto natural después ser analizadas.
- El periodo temporal es transversal ya que se trabajará con datos meteorológicos en un período definido de 4 años de acuerdo a la disponibilidad de información por el GEAA.

2.2 Localización del área de estudio

La provincia de Chimborazo se encuentra ubicada en la región sierra o interandina del Ecuador, rodeada por la cordillera de los Andes que recorre de norte a sur, con una extensión jurisdiccional de 6500.66 km^2. Es conocida como la provincia de las altas cumbres, debido a que en ella se encuentran el volcán Chimborazo, el nevado más alto del Ecuador con una altura de 6.310 msnm. La temperatura promedio es de 13ºC. Tiene una población total de 501.584 habitantes según la proyección del Instituto Ecuatoriano de Censos (INEC, 2010), siendo la novena provincia más poblada del Ecuador, donde su principal actividad es la agricultura. Se divide en 10 cantones, 61 parroquias, 45 rurales y 16 urbanas. Además se contó con 11 estaciones para el monitoreo de variables meteorológicas, cuya información recolectada fue proporcionada por el Grupo de Investigación Energías Alternativas y Ambiente (GEAA) para su posterior análisis (Plan de Desarrollo y Ordenamiento Territorial de Chimborazo, 2020).

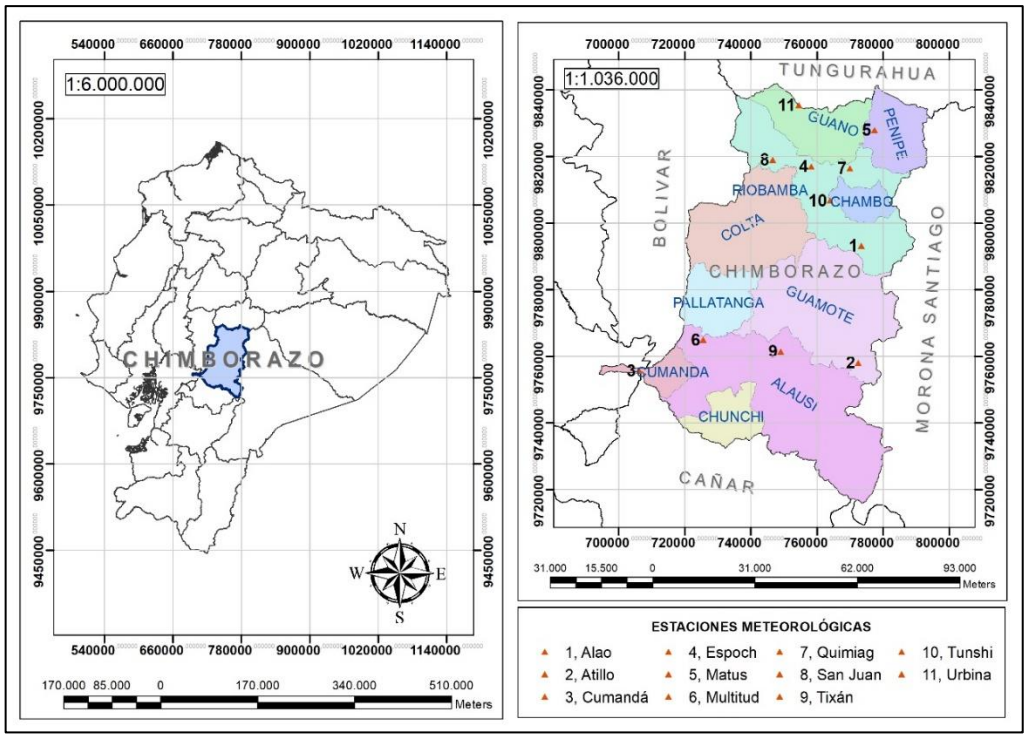


Gráfico 1-2: Ubicación de la provincia de Chimborazo y estaciones meteorológicas.

Fuente: (INEC, 2010).

Realizado por: Checa G., Marisol C.,2020.

2.3 Población de estudio

El estudio se realizó con los datos de temperatura del aire, registrada cada hora en las 11 estaciones meteorológicas de la provincia de Chimborazo durante el periodo 2014-2017.

2.4 Recolección de información

La temperatura del aire en las diferentes estaciones de la provincia de Chimborazo es tomada mediante el termómetro HMP155, por lo que los datos se obtuvieron de los ordenadores que posee cada estación meteorológica y que fue proporcionada por el GEAA.

2.5 Identificación de variables

En el presente trabajo de investigación, la variable de interés para el respectivo análisis es la temperatura del aire en ºC, debido a que es una de las magnitudes más utilizadas para describir el estado del tiempo, además por la necesidad de proveer información confiable al GEAA para el desarrollo de sus investigaciones en beneficio de la colectividad.

2.6 Operalización de variables

A continuación, se presenta la variable proporcionada por el GEAA, la cual cumple con los requerimientos para el desarrollo de los objetivos planteados.

Tabla 1-2: Operalización de la variable.

Variable	Unidad de medida	Tipo	Escala	Descripción
Temperatura del aire	°C	Cuantitativa Continua	Intervalo	Es una magnitud que indica el calentamiento o enfriamiento del aire que resulta del intercambio de calor entre la tierra y la atmósfera.

Realizado por: Checa G., Marisol C.,2020.

2.7 Alcances de investigación

Teniendo en cuenta el problema planteado el trabajo de investigación, tiene alcance descriptivo y estimativo. El primero se ve reflejado en la descripción del comportamiento de la variable en estudio, mientras que el segundo ayudó a la realización de estimaciones en sitios no muestreados mediante la utilización de Kriging Ordinario para Datos Funcionales, dado que permitió trabajar con una función completa en lugar de un solo valor observado, y así aprovechar toda la información con la que se realizó este estudio.

2.8 Análisis de datos

Para el presente estudio, en primera instancia se realizó un análisis exploratorio, para luego proceder con la imputación de datos faltantes mediante el algoritmo MICE, en los casos posibles. Previo al análisis de datos funcionales se obtuvo la forma funcional de la temperatura del aire para las curvas diarias por hora y anuales por día mediante bases B-Splines y Fourier respectivamente. Se calculó la media y desviación estándar funcional, para conocer el comportamiento y variabilidad de la temperatura en las 11 estaciones meteorológicas de la provincia de Chimborazo. Se realizó la detección de atípicos, y mediante el FANOVA se determinó las diferencias significativas entre las curvas promedio de las estaciones en los 4 años de análisis. Para la modelación y estimación de las curvas de temperatura en sitios no muestreados se trabajó con 11 y 15 puntos georreferenciados en la provincia mediante Kriging Ordinario para Datos Funcionales (OKFD), donde a priori se seleccionó el mejor modelo de semivariograma a través de validación cruzada funcional (VCF).

CAPITULO III

3 RESULTADOS Y DISCUSIÓN

3.1 Estructura de la matriz de datos original

En el presente trabajo de investigación se analizó los registros de las estaciones meteorológicas de la provincia de Chimborazo: Alao, Atillo, Cumandá, ESPOCH, Matus, Multitud, Quimiag, San Juan, Tixán, Tunshi y Urbina en un período de 4 años (2014-2017), proporcionados por el GEAA, de los cuales se seleccionó la variable temperatura del aire (ºC) para el respectivo análisis.

A continuación, se muestra una descripción de la estructura de datos de la variable temperatura contenida en cada una de las estaciones:

Tabla 1-3: Porcentaje de datos faltantes por estación y año.

Estaciones meteorológicas	2014		2015		2016		2017	
	VF	%VF	VF	%VF	VF	%VF	VF	%VF
Alao	67	0.8%	0	0%	0	0%	29	0.3%
Atillo	549	6.3%	0	0%	0	0%	401	4.6%
Cumandá	166	1.9%	74	0.8%	3500	39.8%	2610	29.8%
ESPOCH	2217	25.3%	173	2.0%	3	0%	405	4.6%
Matus	7937	90.6%	107	1.2%	8569	97.6%	5	0.1%
Multitud	473	5.4%	551	6.3%	0	0%	1606	18.3%
Quimiag	13	0.1%	0	0%	0	0%	21	0.2%
San Juan	24	0.3%	0	0%	0	0%	28	0.3%
Tixán	292	3.3%	0	0%	0	0%	22	0.3%
Tunshi	228	2.6%	0	0%	629	7.2%	1788	20.4%
Urbina	31	0.4%	0	0%	0	0%	21	0.2%

Realizado por: Checa G., Marisol C.,2020.

Cabe recalcar que para los años 2014, 2015 y 2017 el total de datos por año era de 8760, mientras que para el 2016 (Bisiesto) era de 8784. Los valores faltantes (VF) fueron imputados mediante el algoritmo MICE.

3.2 Análisis exploratorio de datos

El primer paso en cualquier trabajo de investigación, implica un análisis exploratorio de datos atípicos, faltantes o ausentes (NA's), ya que a priori se necesita conocer el estado de la base de datos con la que se realizará el estudio.

Para la identificación de datos atípicos se realizó diagramas de caja, donde la mayor parte de estaciones mostraban la existencia de datos sospechosos, pero no fueron eliminados ya que cumplían con las especificaciones de la OMM (Tabla 2-1). Sin embargo, la estación de Atillo 2014 presentó temperaturas entre -40 y -70 ºC, y Matus 2015 entre 40 y 60ºC, cuyos valores sobrepasan la temperatura mínima de -1.9 ºC y máxima de 32.4ºC aproximadamente que presentó la región interandina en estos años (INAMHI). El gráfico 1-3 muestra los valores atípicos de dichas estaciones que fueron eliminados con un total de 86 y 85 respectivamente.

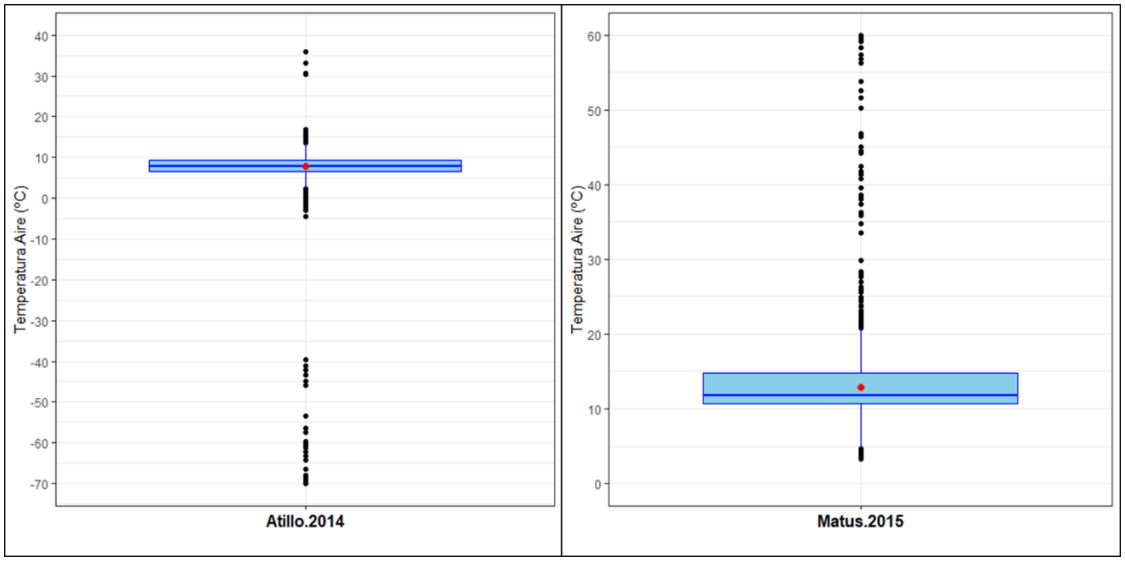


Gráfico 1-3: Diagramas de caja de Temperatura para Atillo 2014 y Matus 2015.

Realizado por: Checa G., Marisol C.,2020.

Posteriormente se identificó los datos faltantes con ayuda del software R y la librería VIM, que permite obtener representaciones gráficas del porcentaje y patrón de faltantes. Para este análisis se tomó como ejemplo la estación meteorológica Quimiag 2017, con 9 variables (temperatura del aire, humedad relativa y temperatura del suelo medida a -5, 0, 5, -10, -20, -50 y -100 cm), debido a que, en un análisis previo en los años 2014, 2015 y 2016, se obtuvo una imputación confiable con estas variables.

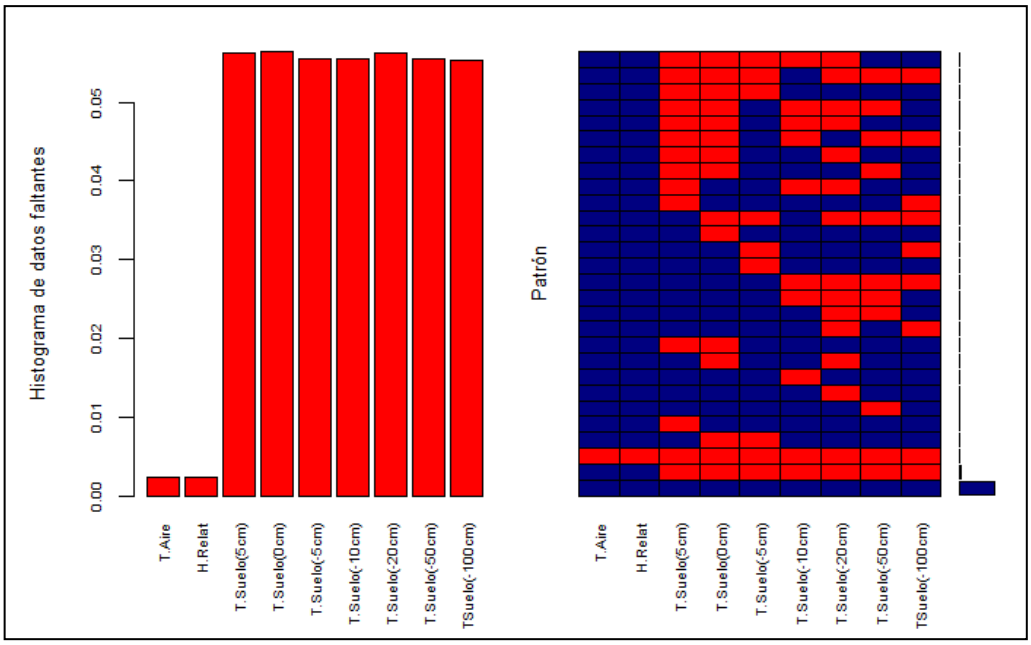


Gráfico 2-3: Patrón de datos faltantes de la estación Quimiag 2017.

Realizado por: Checa G., Marisol C.,2020.

En el gráfico 2-3 se observa el patrón aleatorio de faltantes que no superan el 20%, por lo que se procede a la imputación con el método seleccionado.

3.3 Imputación múltiple en R

Para este análisis se utilizó el algoritmo MICE, que se encuentra disponible en R-Studio. Esta imputación múltiple crea varias versiones completas de los datos, reemplazando los valores faltantes (VF) con valores factibles, obtenidos a través de una distribución modelada específicamente para cada valor faltante.

La imputación MICE de igual manera se realizó con los datos usados anteriormente (Quimiag 2017), debido a que cumplen con los requisitos de este método. El procedimiento a seguir se muestra a continuación:

Con ayuda de la función $mice( )$ se procedió a realizar el relleno. Esta función requiere el total de faltantes, el método y número de imputaciones. En este caso se utilizó el método pmm (predictive mean matching) y 5 imputaciones. En el gráfico 3-3 se muestra las imputaciones (en rojo) que son candidatas a ser imputadas, donde se mantiene muy bien la variabilidad de los datos, en este caso de la nube de puntos formada por la visión bidimensional de las dos variables a prueba (temperatura del aire y humedad relativa).

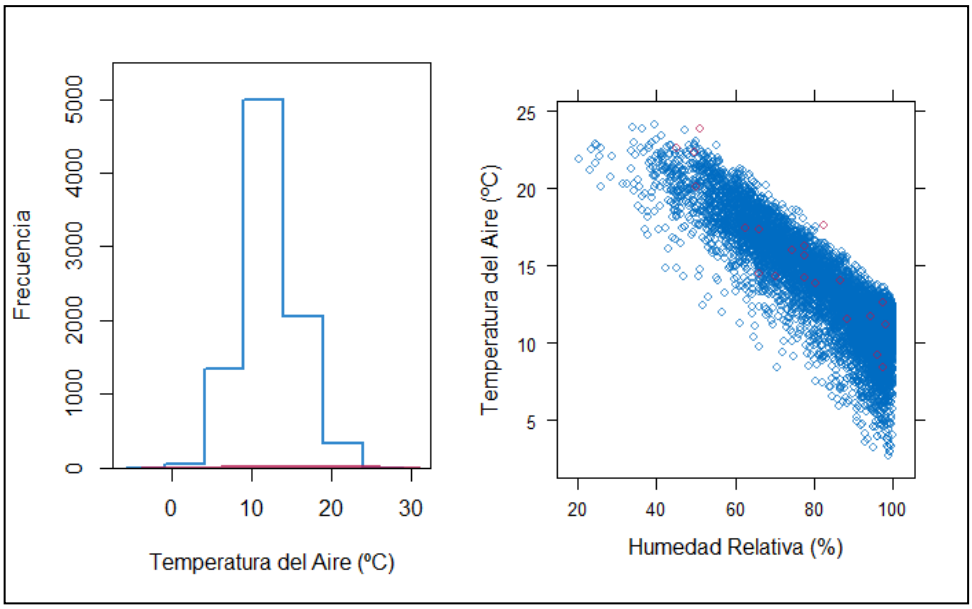


Gráfico 3-3: Modelo de imputación múltiple.

Realizado por: Checa G., Marisol C.,2020.

Para la imputación efectiva de la base de datos, se procedió a completar y obtener dichos datos imputados, para ello se utiliza la función *complete*( ). Una vez realizado este paso, se observó la eficacia del modelo de imputación, comparando la densidad las variables antes de ser imputadas y luego de ser imputadas. En el gráfico 4-3 se observa que la imputación fue muy eficaz, ya que la distribución es prácticamente igual a la variable original.

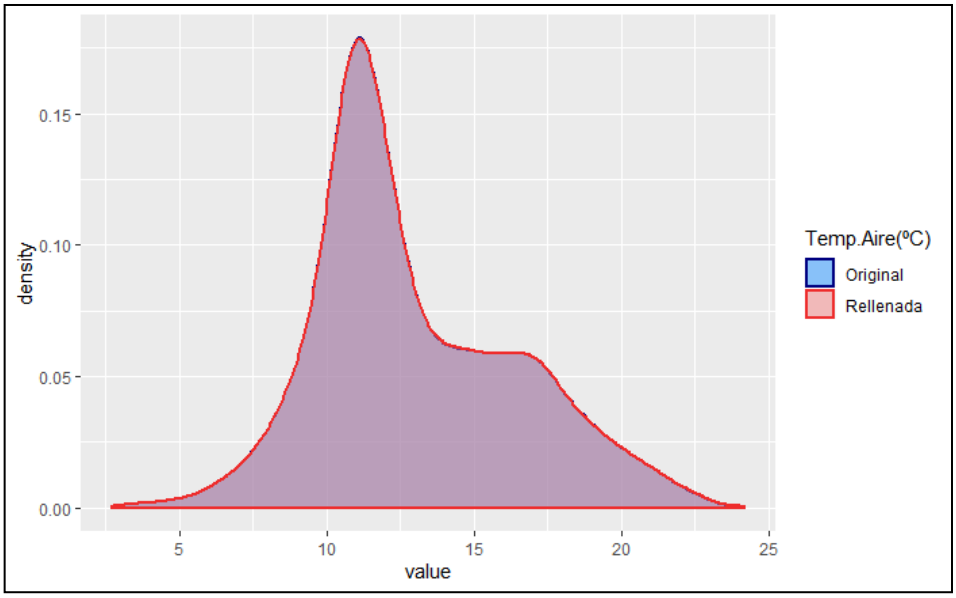


Gráfico 4-3: Densidad de la variable temperatura original vs imputada.

Realizado por: Checa G., Marisol C.,2020.

Además, se verificó que las imputaciones se encuentren dentro de los valores mínimo y máximo de la base de datos original, que cumplen con las normas establecidas por la OMM (Tabla 2-1).

Por último, para una mejor verificación del modelo de imputación se analizó el coeficiente de determinación (R^2) y el error estándar (RSE) obtenidos mediante regresión múltiple. La tabla 2-3 indica los resultados para la variable temperatura sin valores faltantes en las diferentes estaciones meteorológicas, periodo 2014 - 2017:

Tabla 2-3: Resumen del análisis de regresión múltiple para la variable temperatura.

	2014		**2015**		**2016**		**2017**	
Estaciones	**R^2**	**RSE**	**R^2**	**RSE**	**R^2**	**RSE**	**R^2**	**RSE**
Alao	93%	0.875	No se relleno		No se relleno		79%	1.612
Atillo	80%	1.324	No se relleno		No se relleno		83%	1.184
Cumandá	87%	0.838	85%	0.904	No se relleno		No se relleno	
ESPOCH	No se relleno		97%	0.632	No se relleno		74%	1.795
Matus	No se relleno		54%	2.686	No se relleno		96%	0.677
Multitud	70%	0.831	66%	0.817	No se relleno		87%	0.626
Quimiag	95%	0.728	No se relleno		No se relleno		85%	1.348
San Juan	92%	0.930	No se relleno		No se relleno		79%	1.613
Tixán	62%	1.623	No se relleno		No se relleno		69%	1.501
Tunshi	94%	0.811	No se relleno		80%	0.892	No se relleno	
Urbina	93%	0.662	No se relleno		No se relleno		95%	1.95

Realizado por: Checa G., Marisol C.,2020.

Aunque no se pudo rellenar todas las bases de datos de las 11 estaciones por año, debido a que el ADF (Análisis de Datos Funcionales) trabaja con curvas, se procedió a seleccionar de cada una de las estaciones no rellenadas los días completos para el posterior análisis.

3.4 Análisis de Datos Funcionales

3.4.1 Selección de la Base y número de funciones

El primer paso en el análisis de datos funcionales (ADF) es definir el método de base funcional que mejor ajuste y represente los datos observados (Gráfico 5-3), con el objetivo de obtener la función que da lugar a las observaciones discretas de la muestra y de eliminar el ruido producido por los sistemas de medición.

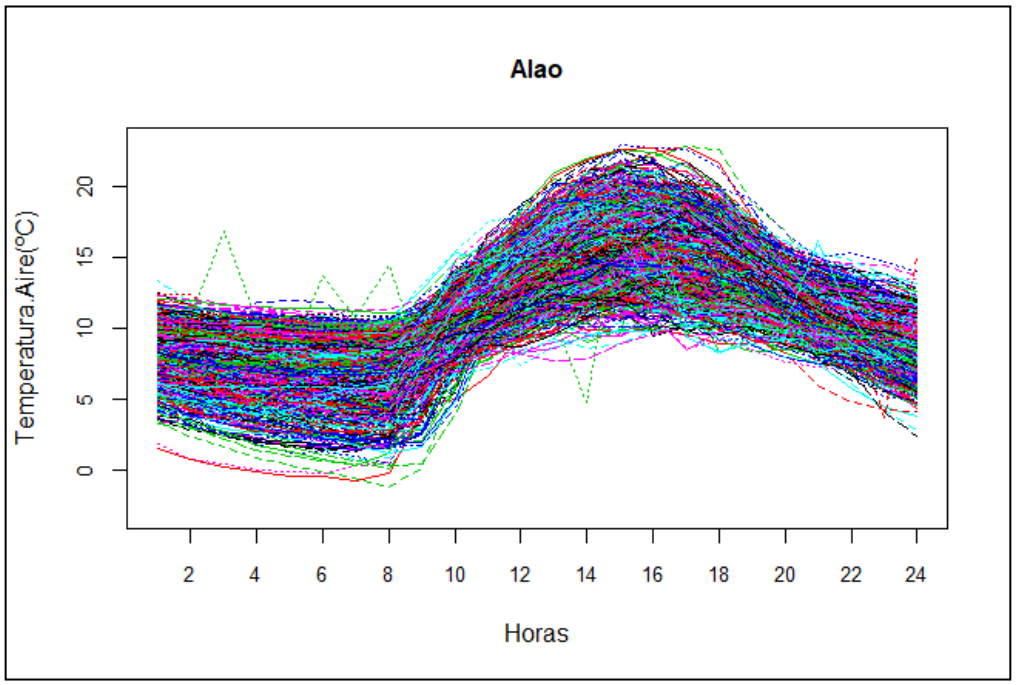


Gráfico 5-3: Curvas No suavizadas de temperatura diaria, Alao 2014-2017.

Realizado por: Checa G., Marisol C.,2020.

Para los datos de temperatura diaria se utilizó las bases de funciones B-Splines, debido a que es una de las más utilizadas para datos que no presentan patrones regulares de ciclo, pues únicamente se trabajó en un intervalo de tiempo de 24 horas. Además, que presentan un coste computacional muy bajo cuando se trabaja con grandes cantidades de datos.

Para la elección adecuada del número de bases *k* se utilizó la función $min.basis(\,)$ del paquete $fda.usc$ en el software R, la cual mediante Validación Cruzada Generalizada (VCG) y previamente definiendo el intervalo T = 1,…, 24; se obtuvo un número de bases de 13 y 15. Sin embargo considerando que la varianza residual (s^2) entre los datos reales y los suavizados tiende a disminuir considerablemente, la base funcional óptima fue de 15.

Con la ayuda de la función $plotfit.fd()$ se puede observar algunos ejemplos de la varianza residual de dicho ajuste:

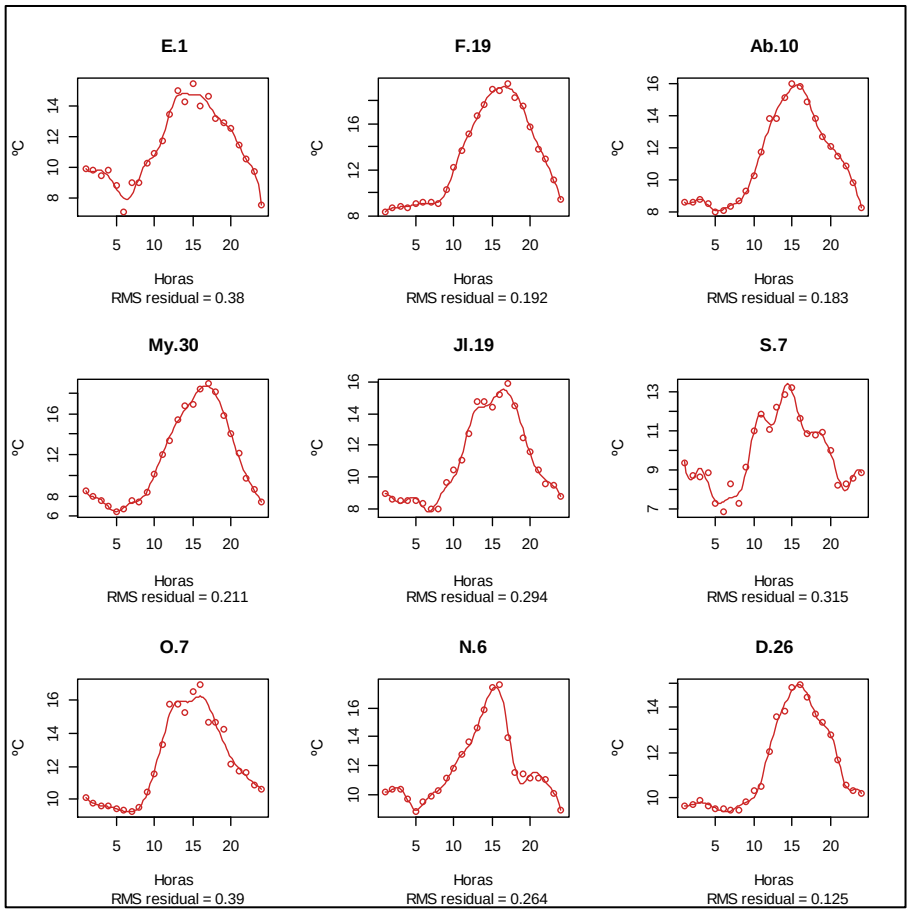


Gráfico 6-3: Ajuste de las curvas y la varianza residual de sus variaciones.

Realizado por: Checa G., Marisol C.,2020.

En el gráfico 6-3 con un ajuste B-Splines Cúbico de 15 bases, el promedio de la varianza residual entre los datos observados y suavizados RMS = 0.2382, siendo el más óptimo.

3.4.2 Suavización de las curvas

Para realizar el suavizado se utilizó el paquete fda y $fda.usc$ (Febrero-Bande y Oviedo de la Fuente, 2012; Ramsay et al., 2009), el cual se dio en dos pasos:

1. En primer lugar, se creó el sistema de bases, en este caso B-Splines mediante la función: $create.bspline.basis()$ con el número de funciones base que se determinó anteriormente $(nb = 15)$.
2. Se ajustó los datos a esta base creada mediante la instrucción $Data2fd()$.

Otra manera de suavizar los datos es mediante $fdata2fd()$, la cual automáticamente trabaja con base B-Splines y únicamente necesita el número de funciones base y una $fdata$. En este estudio se utilizó las dos maneras dependiendo de los resultados que se deseó obtener.

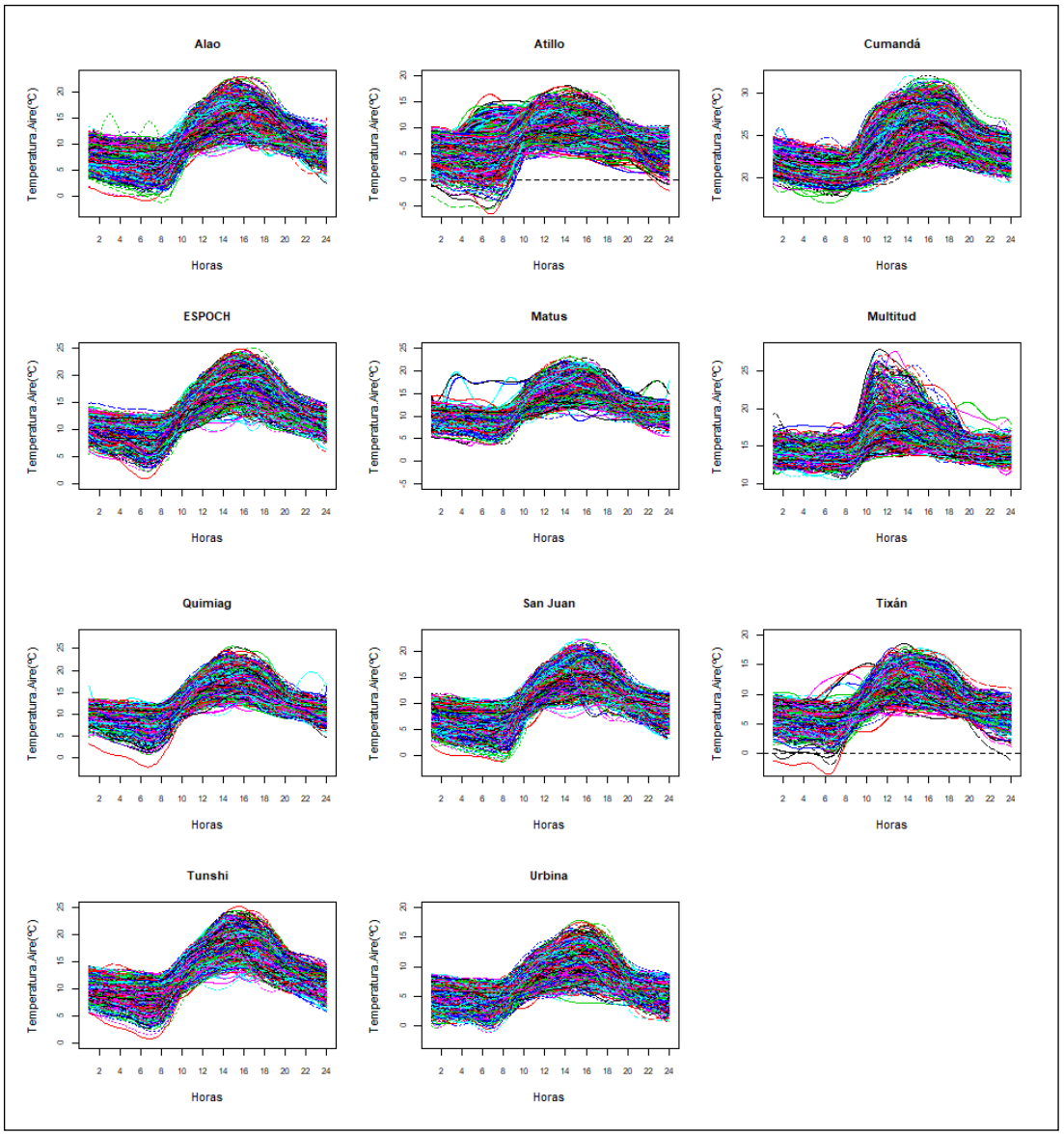


Gráfico 7-3: Curvas suavizadas de temperatura del aire de las 11 estaciones meteorológicas de Chimborazo, 2014-2017.

Realizado por: Checa G., Marisol C.,2020.

En el gráfico 7-3 se observa que cada uno de los datos suavizados representa curvas diarias de temperatura del aire (ºC) por hora. Además, se puede evidenciar que las estaciones de Matus, Multitud y Tixán presentan comportamientos extraños de temperatura, por lo que en un posterior análisis se realizó la detección de datos atípicos funcionales.

3.4.3 Análisis Descriptivo Funcional

Estación Alao

En la estación de Alao (Gráfico 8-3, (a)) se observó curvas medias de temperatura del aire (ºC) bajas en horas de la mañana y noche, y que creció significativamente en la tarde, mostrando un alcance de 22 ºC aproximadamente. La desviación estándar funcional (Gráfico 8-3, (b)) con respecto a su media funcional mostró variabilidad en todo el día, en especial entre las 09H00 y 17H00 donde se evidenció mayor inestabilidad.

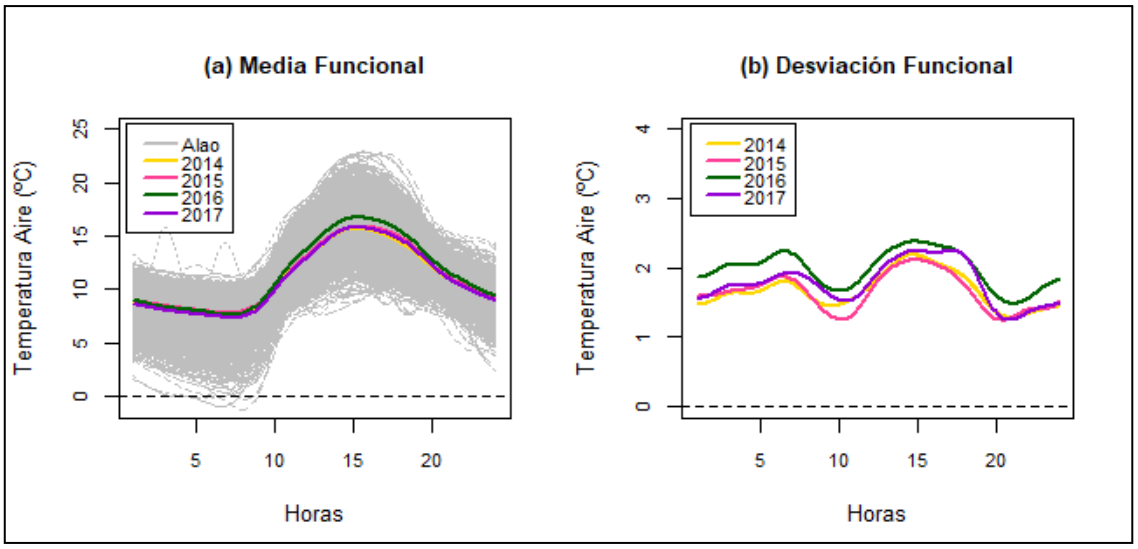


Gráfico 8-3: Media y Desviación Funcional de la temperatura, Estación Alao.

Realizado por: Checa G., Marisol C.,2020.

Estación Atillo

En la estación de Atillo (Gráfico 8-3, (a)) el comportamiento de temperatura en algunos casos estuvo por debajo de los 0 ºC en la mañana, mientras que en la tarde creció significativamente presentando un alcance de 15 ºC. El gráfico 8-3, (b) de desviación estándar funcional indicó una variabilidad en forma decreciente en todo el día. En el año 2014 las función media y desviación estándar de temperatura presentaron un comportamiento diferente al resto de años.

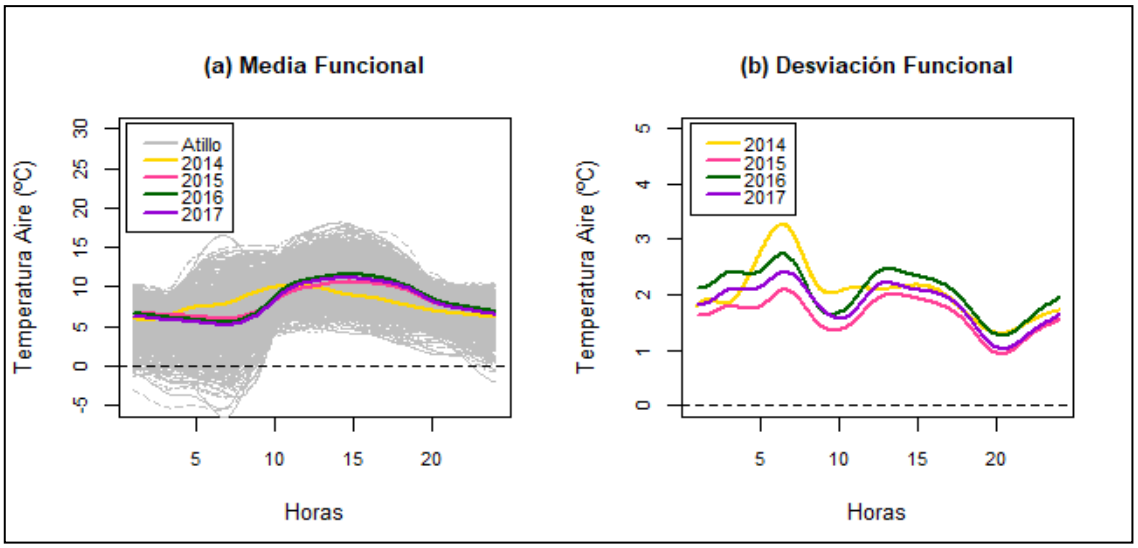


Gráfico 9-3: Media y Desviación Funcional de la temperatura, Estación Atillo.

Realizado por: Checa G., Marisol C.,2020.

Estación Cumandá

En la estación de Cumandá la temperatura en horas de la tarde creció significativamente, presentando un clima cálido similar a la costa, ya que su temperatura está entre 17 ºC y 30 ºC aproximadamente, mientras que la desviación estándar funcional con respecto a su media funcional indicó poca variabilidad entre las 09H00 y 17H00 (Gráfico 10-3).

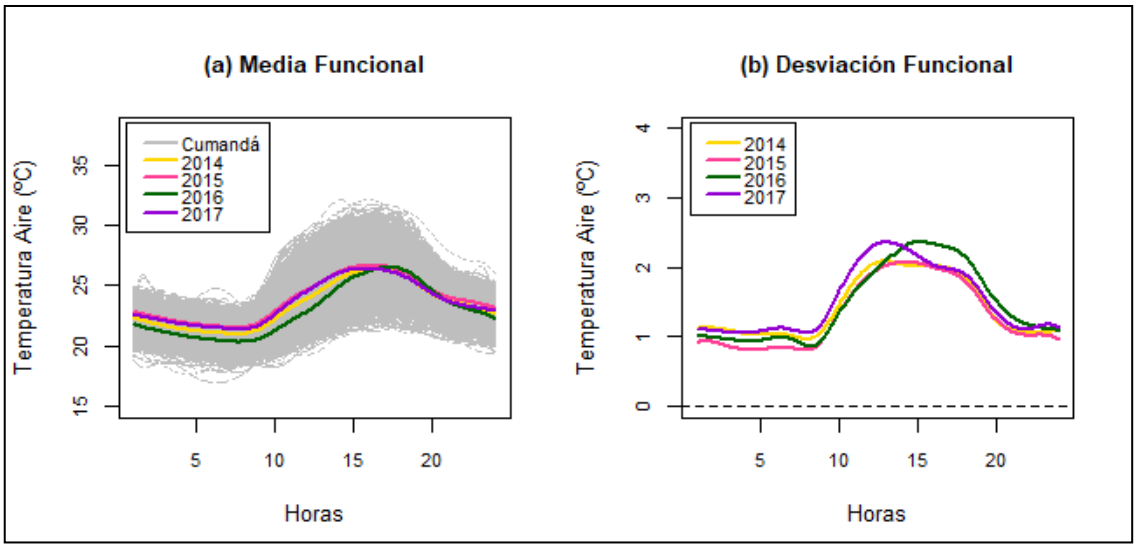


Gráfico 10-3: Media y Desviación Funcional de la temperatura, Estación Cumandá.

Realizado por: Checa G., Marisol C.,2020.

Estación ESPOCH

La estación ESPOCH presentó bajas temperaturas (ºC) en horas de la mañana y noche, mientras a medio día creció significativamente llegando hasta los 25ºC aproximadamente. La gráfica de desviación estándar funcional indicó variabilidad durante las 09H00 y 17H00 (Gráfico 11-3).

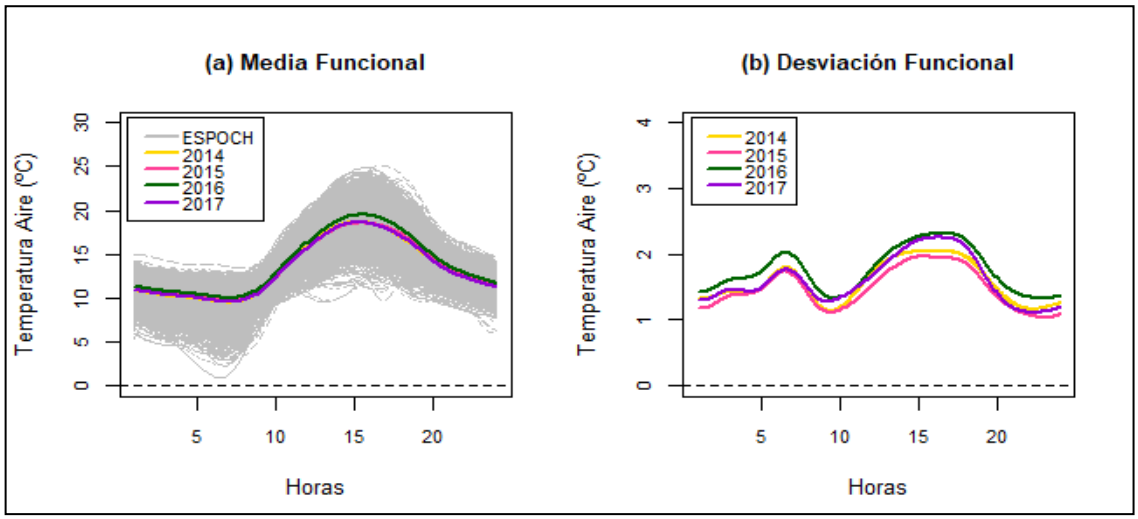


Gráfico 11-3: Media y Desviación Funcional de la temperatura, Estación ESPOCH.

Realizado por: Checa G., Marisol C.,2020.

Estación Matus

En la estación Matus, la temperatura creció significativamente en horas de la tarde. La gráfica de desviación estándar funcional indicó gran variabilidad en todo el día. En el 2014 la media y desviación funcional presentaron un comportamiento diferente al resto de años (Gráfico 12-3).

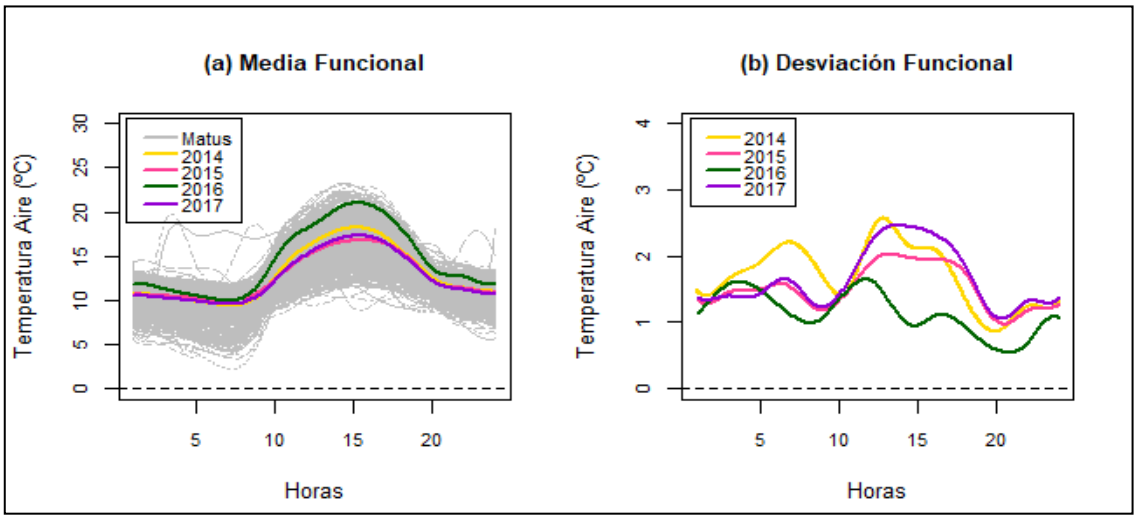


Gráfico 12-3: Media y Desviación Funcional de la temperatura, Estación Matus.

Realizado por: Checa G., Marisol C.,2020.

Estación Multitud

El gráfico 13-3, (a) indicó el comportamiento de la temperatura en la estación Multitud, la cual fue baja en la mañana y noche, y creció significativamente a medio día. La gráfica de desviación estándar funcional mostró variabilidad con respecto a su media funcional entre las 09H00 y 17H00. En el año 2016 la función media y desviación estándar presentaron un comportamiento diferente a los años 2014, 2015 y 2017. Esta estación meteorológica presentó temperaturas entre los 10º C y 28 ºC aproximadamente.

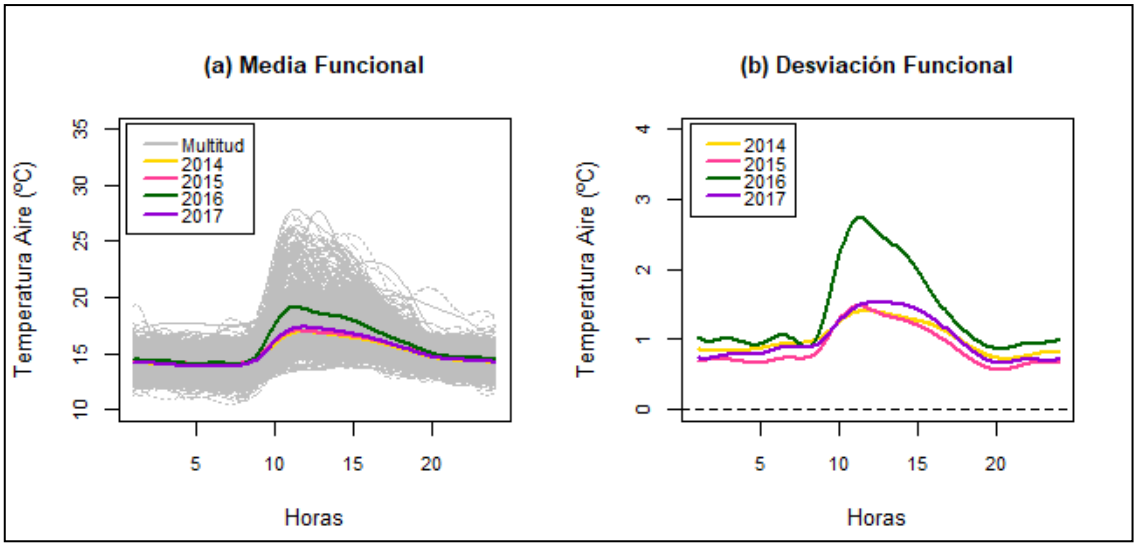


Gráfico 13-3: Media y Desviación Funcional de la temperatura, Estación Multitud.

Realizado por: Checa G., Marisol C.,2020.

Estación Quimiag

El gráfico 14-3, (a) presentó el comportamiento de temperatura en la estación Quimiag con un alcance de 25 ºC, temperaturas bajas en la mañana y noche, y altas en horas de la tarde. La gráfica 14-3, (b) de desviación estándar funcional con respecto a su media funcional muestra variabilidad en todo el día.

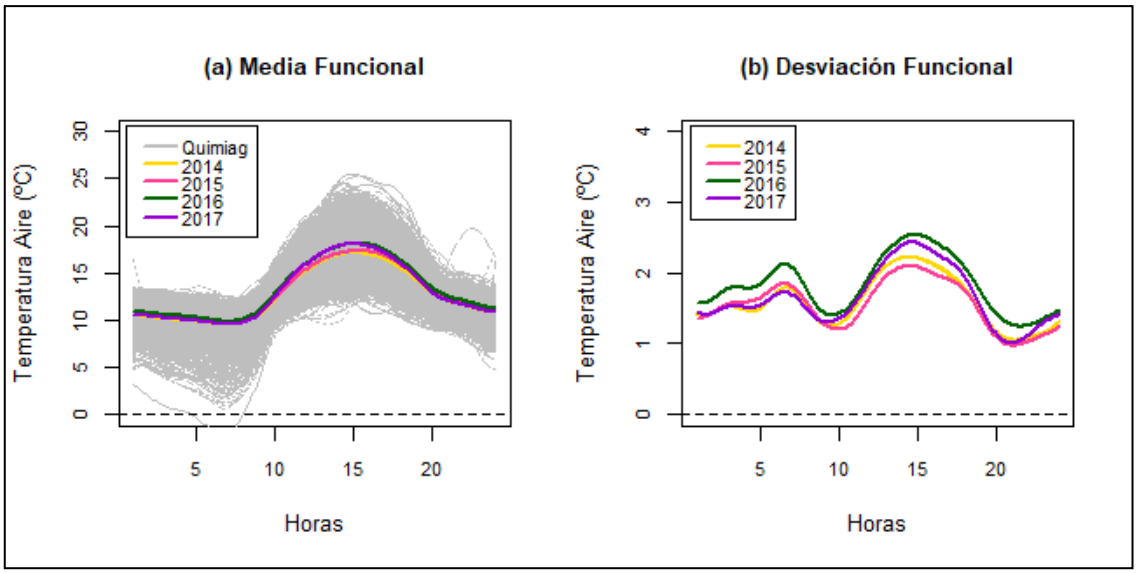


Gráfico 14-3: Media y Desviación Funcional de la temperatura, Estación Quimiag.

Realizado por: Checa G., Marisol C.,2020.

Estación San Juan

El gráfico 15-3, (a) indicó el comportamiento de la temperatura en la estación San Juan, la cual se encuentra entre -0 ºC y 22 ºC aproximadamente. La temperatura en horas de la tarde creció significativamente mientras que en la mañana y noche fue baja. La desviación estándar respecto a su media funcional mostró variabilidad en todo el día durante las 09H00 y 17H00.

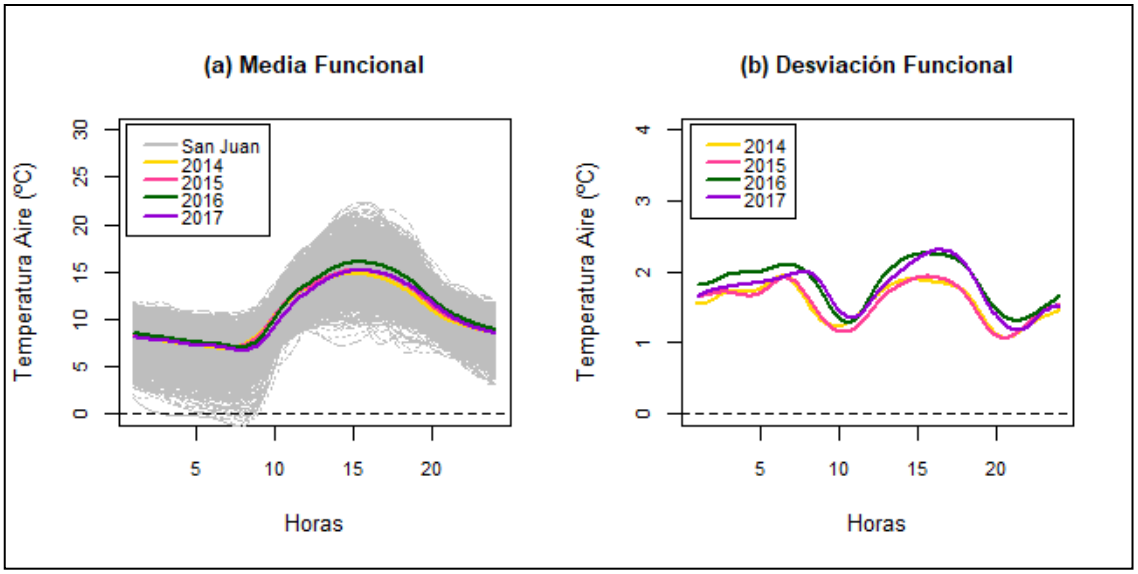


Gráfico 15-3: Media y Desviación Funcional de la temperatura, Estación San Juan.

Realizado por: Checa G., Marisol C.,2020.

Estación Tixán

En la estación Tixán el gráfico 16-3, (a) indicó el comportamiento de la temperatura en todo el día, donde en la mañana fue baja y a medio día creció significativamente. La desviación estándar funcional mostró variabilidad con respecto a su media funcional en todo el día.

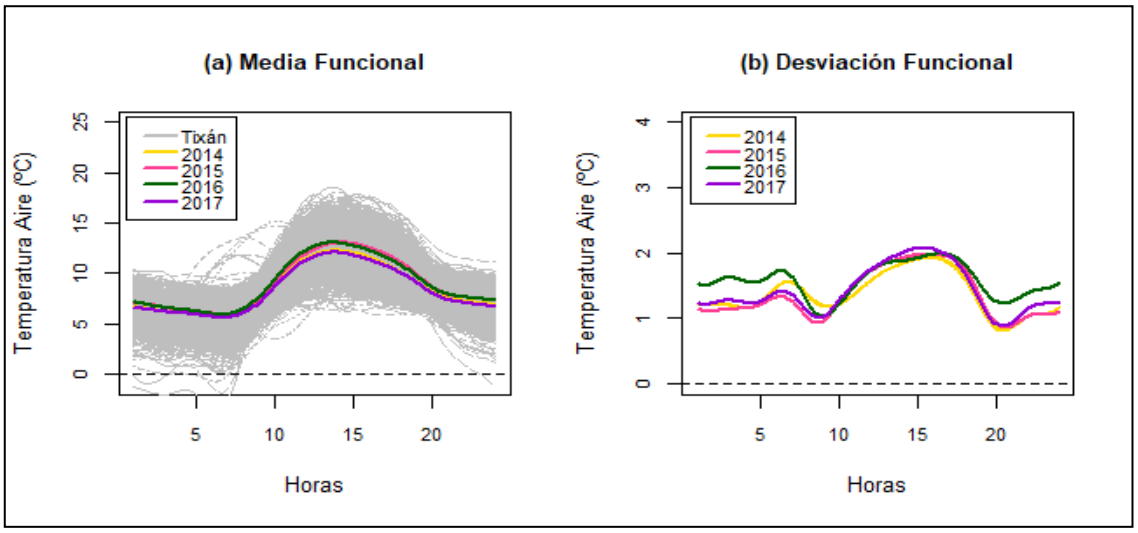


Gráfico 16-3: Media y Desviación Funcional de la temperatura, Estación Tixán.

Realizado por: Checa G., Marisol C.,2020.

Estación Tunshi

La estación de Tunshi en el gráfico 17-3, (a) indicó que al inicio y al final del día existen bajas temperaturas y a mediodía creció significativamente. En el gráfico 17-3, (b) se observó variabilidad en todo el día con respecto a su media funcional durante las 09H00 y 17H00.

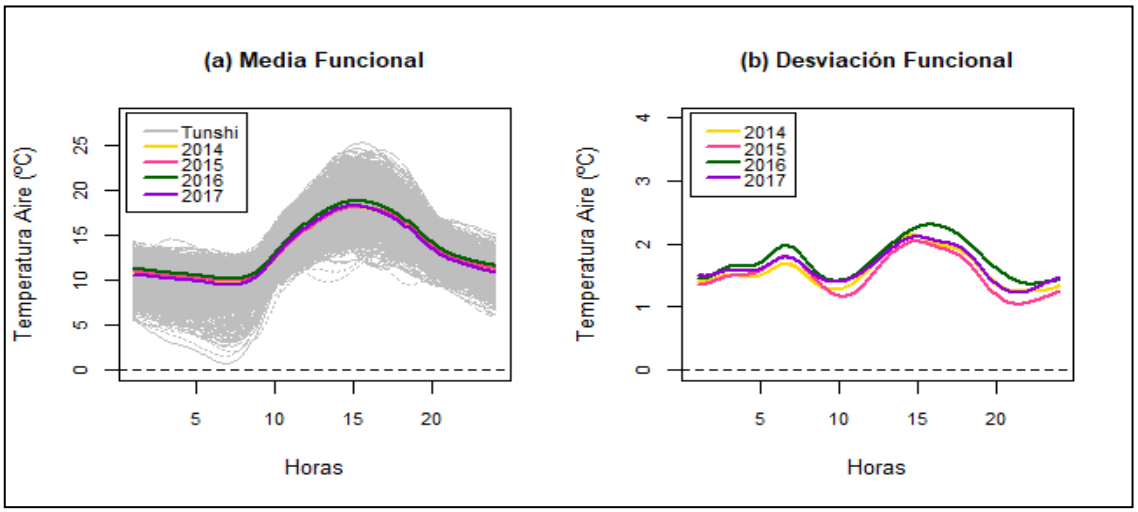


Gráfico 17-3: Media y Desviación Funcional de la temperatura, Estación Tunshi.

Realizado por: Checa G., Marisol C.,2020.

Estación Urbina

El gráfico 18-3, (a) indicó el comportamiento de la temperatura en la estación de Tunshi, cuya temperatura se encontró entre -0ºC y 18ºC aproximadamente, al inicio y al final del día existió bajas temperatura mientras que a medió día creció significativamente (10H00-17H00). En el gráfico 18-3, (b), se observó la desviación estándar funcional donde la variabilidad es significativa en todo el día respecto a su media funcional.

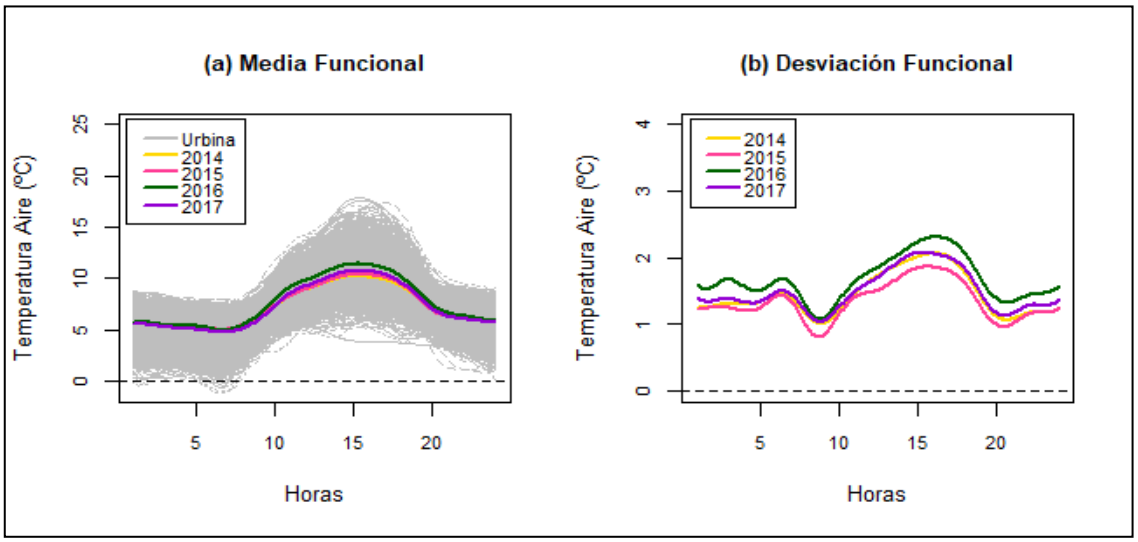


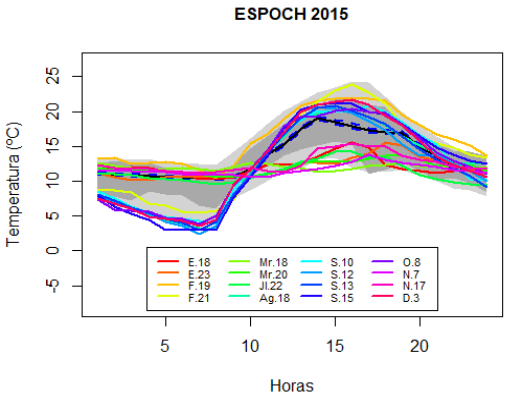
Gráfico 18-3: Media y Desviación Funcional de la temperatura, Estación Urbina.
Realizado por: Checa G., Marisol C.,2020.

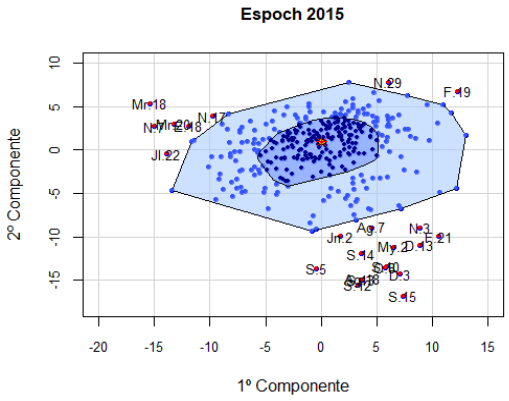
3.4.4 Detección de datos funcionales atípicos

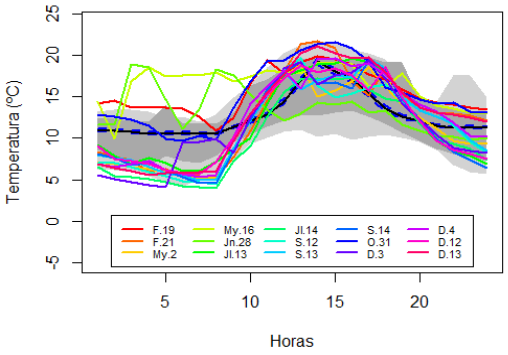
Los outliers o datos atípicos son un problema que alteran significativamente los resultados y las conclusiones que se derivan del análisis, por ello el siguiente paso fue la detección de curvas atípicas con medidas de profundidad.

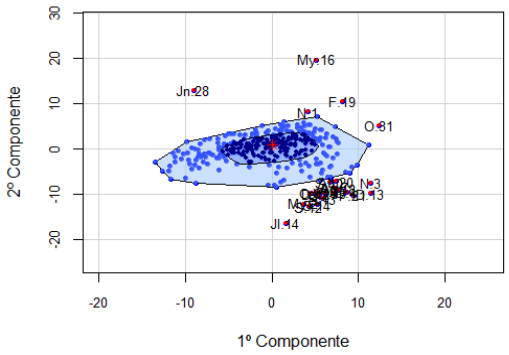
En este caso se utilizó una medida de profundidad basada en las dos primeras componentes principales de las curvas, transformando el problema funcional a bidimensional, esta profundidad es la profundidad de Tukey. Por ello se empleó la función $fboxplot()$ y $bagplot()$ que presentaron un gráfico visual de los datos funcionales atípicos (Hyndman y Shang, 2010).

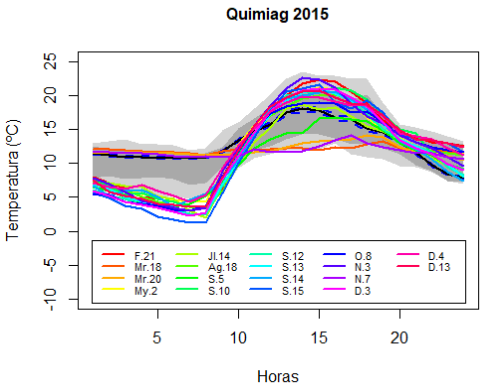
A continuación, se presenta bagplots funcionales y bidimensionales de algunas estaciones meteorológicas:

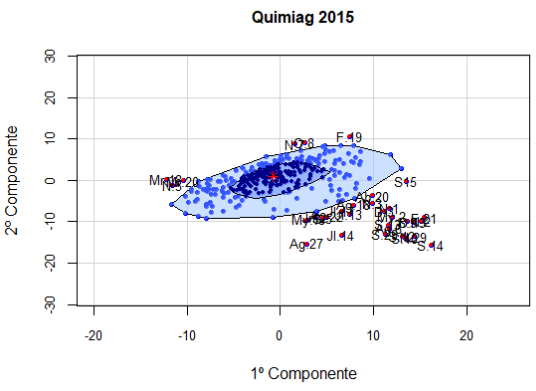

ESPOCH 2015
Temperatura (ºC)
Horas
E.18 E.23 F.19 F.21 Mr.18 Mr.20 Jl.22 Ag.18 S.10 S.12 S.13 S.15 O.8 N.7 N.17 D.3

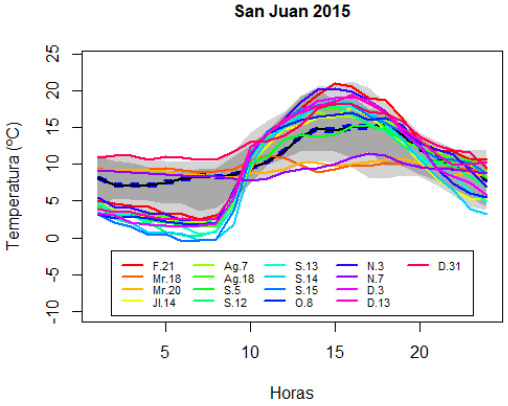

Espoch 2015
2º Componente
1º Componente

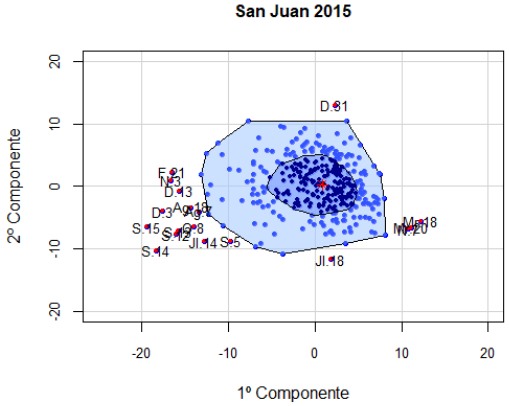

Matus 2015
Temperatura (ºC)
Horas
F.19 F.21 My.2 My.16 Jn.28 Jl.13 Jl.14 S.12 S.13 S.14 O.31 D.3 D.4 D.12 D.13


Matus 2015
2º Componente
1º Componente


Quimiag 2015
Temperatura (ºC)
Horas
F.21 Mr.18 Mr.20 My.2 Jl.14 Ag.18 S.5 S.10 S.12 S.13 S.14 S.15 O.8 N.3 N.7 D.3 D.4 D.13


Quimiag 2015
2º Componente
1º Componente


San Juan 2015
Temperatura (ºC)
Horas
F.21 Mr.18 Mr.20 Jl.14 Ag.7 Ag.18 S.5 S.12 S.13 S.14 S.15 O.8 N.3 N.7 D.3 D.13 D.31


San Juan 2015
2º Componente
1º Componente

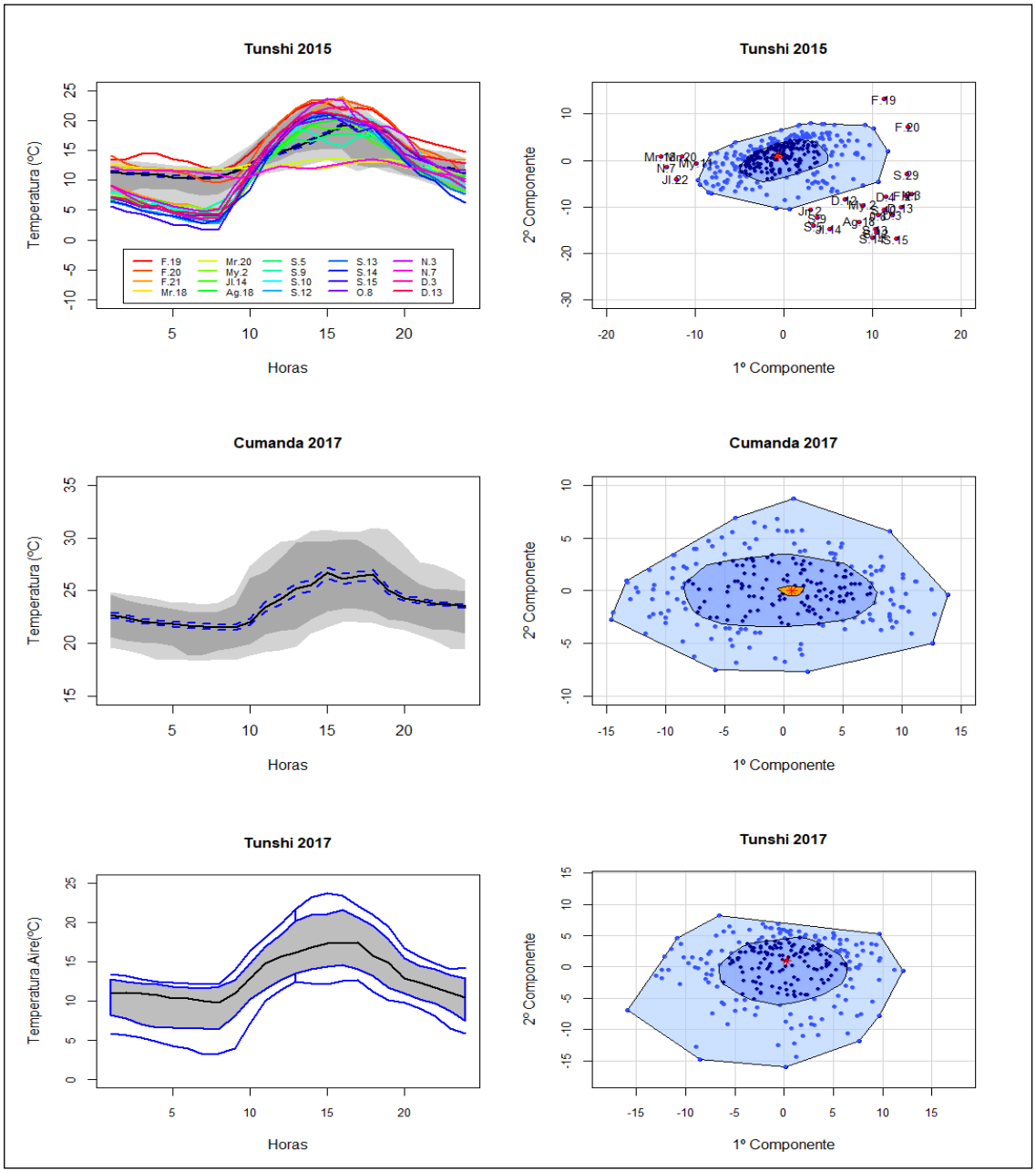


Gráfico 19-3: Bagplots funcional y bivariado de temperatura diaria.

Realizado por: Checa G., Marisol C.,2020.

Revisando los resultados que presenta el gráfico 19-3, se identificó que las estaciones que poseen la mayor cantidad de funciones atípicas (líneas de colores) fue ESPOCH (24), Matus (23), Quimiag (33), San Juan (18) y Tunshi (27) 2015, debido a que el comportamiento de la temperatura a lo largo del día fue diferente tanto en forma y magnitud del resto de las curvas, especialmente en los días del mes de Septiembre, donde la temperatura fue alta en horas de la tarde. Mientras que en las estaciones de Cumandá y Tunshi 2017 no se evidenció este problema.

Para evitar la influencia en los análisis posteriores como FANOVA y modelación geoestadística se separó las curvas atípicas presentes en cada una de las estaciones meteorológicas.

3.4.5 Análisis de la Varianza Funcional (FANOVA)

En esta sección se presenta el Análisis de la Varianza Funcional (FANOVA), con el objetivo de determinar si existen diferencias significativas entre las curvas diarias de temperatura por hora (Caso 1); además de curvas anuales por día (Caso 2), debido a que en el análisis geoestadístico de datos funcionales se trabajó con los dos casos. Este contraste se aplicó para los cuatro años de estudio en cada estación meteorológica

Se utilizó la función $anova.onefactor()$ del paquete $fda.usc$ para la obtención del valor p, valor observado (V_n) y crítico ($V_\propto$), para aplicar el contraste de igualdad de medias funcionales.

3.4.5.1 FANOVA para curvas diarias de temperatura por hora

El gráfico 20-3 presenta el contraste dado, donde se obtiene un gráfico de medias bootstrap de los 4 grupos (años) (Gráfico 20-3, (a)), uno de medias globales basadas en 100 remuestreos de bootstrap (Gráfico 20-3, (b)), finalmente el estadístico de contraste, donde el valor muestral se representa por una línea segmentada (Gráfico 20-3, (c)) para la estación Alao.

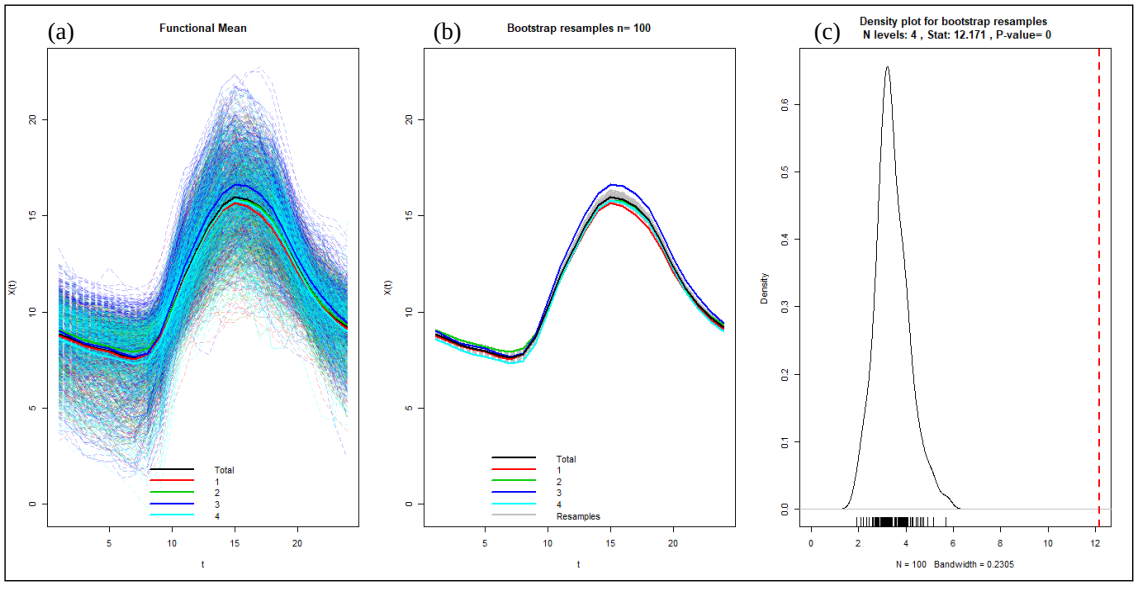


Gráfico 20-3: FANOVA de temperatura para la estación de Alao (Caso 1).

Realizado por: Checa G., Marisol C.,2020.

Para la estación de Alao, aplicando el FANOVA se obtuvo valores $V_n = 12.17$ y $V_\propto = 4.81$, donde es claro que $V_n > V_\propto$ y con un valor p igual a cero, se rechaza la hipótesis nula de igualdad de medias funcionales, concluyendo así que existe evidencia suficiente para indicar que el comportamiento medio de temperatura en Alao es diferente en los años 2014, 2015, 2016 y 2017.

Tabla 3-3: Resultados de la prueba FANOVA por estación (Caso 1).

Nº	Estaciones	V_n	$V_\propto$	P - value
1	Alao	12.171	4.806	0.00
2	Atillo	30.77043	5.287028	0.00
3	Cumandá	17.81972	4.824149	0.00
4	ESPOCH	10.39374	4.437562	0.00
5	Matus	37.85824	13.54087	0.00
6	Multitud	15.55654	3.047905	0.00
7	Quimiag	12.59154	4.631645	0.00
8	San Juan	15.01717	4.413837	0.00
9	Tixán	13.41498	3.889809	0.00
10	Tunshi	12.06286	4.595922	0.00
11	Urbina	11.52324	4.240449	0.00

Realizado por: Checa G., Marisol C.,2020.

Al igual que Alao, aplicando el contraste de FANOVA al resto de estaciones, se concluyó que el comportamiento medio de temperatura es significativamente diferente entre los años de estudio, dado su valor p = 0.00 (Tabla 3-3).

3.4.5.2 FANOVA para curvas anuales de temperatura por día

En este caso se utilizó bases Fourier ya que los datos presentan periodicidad, con un número de funciones igual a 365, debido a que el promedio de la varianza residual entre los datos observados y suavizados RMS = 0.047, siendo el más adecuado.

El gráfico 21-3 presenta el contraste dado, donde se obtiene un gráfico de medias bootstrap de los 4 grupos (años) (Gráfico 21-3, (a)), uno de medias globales basadas en 100 remuestreos de bootstrap (Gráfico 21-3, (b)), finalmente el estadístico de contraste, donde el valor muestral se representa por una línea segmentada, que permite afirmar de manera empírica que el comportamiento medio de la temperatura anual en Alao es significativamente igual en los años 2014, 2015, 2016 y 2017 (Gráfico 21-3, (c)).

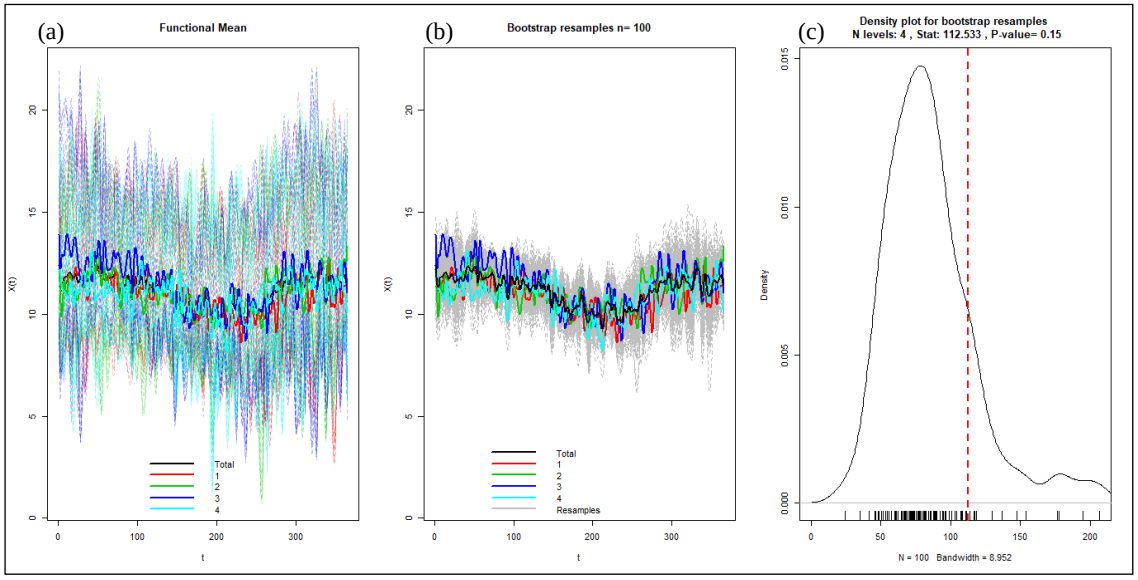


Gráfico 21-3: FANOVA de temperatura para la estación de Alao (Caso 2).

Realizado por: Checa G., Marisol C.,2020.

La tabla 4-3 muestra valores p mayores a un nivel de significancia del 5% en la mayoría de estaciones, por lo que no se rechaza la hipótesis nula, concluyendo así que no existe evidencia suficiente para indicar que el comportamiento medio de la temperatura anual entre los años 2014, 2015, 2016 y 2017 es diferente para las estaciones meteorológicas de Alao, Atillo, Cumandá, ESPOCH, Matus, Quimiag, San Juan, Tunshi y Urbina, al igual que Multitud y Tixán a un nivel del 1%.

Tabla 4-3: Resultados de la prueba de FANOVA por estación (Caso 2).

Nº	Estaciones	V_n	V_∞	p-value
1	Alao	112.53	147.59	0.15
2	Atillo	112.46	137.48	0.05
3	Cumandá	101.48	112.09	0.05
4	ESPOCH	123.3	146.92	0.16
5	Matus	164.82	175.66	0.05
6	Multitud	75.36	95.34	0.03
7	Quimiag	120.3	147.19	0.18
8	San Juan	114.84	143.51	0.22
9	Tixán	135.4	143.56	0.03
10	Tunshi	121.2	161.31	0.19
11	Urbina	101.96	104.17	0.06

Realizado por: Checa G., Marisol C.,2020.

3.5 Análisis descriptivo espacial

Antes de proceder con el análisis geoestadístico de datos funcionales mediante OKFD se realizó un análisis descriptivo espacial de la variable Temperatura del Aire (ºC), con el objetivo de evaluar el supuesto de estacionariedad.

Desde un punto de vista empírico la distribución espacial entre las 11 localizaciones de las estaciones meteorológicas fue estacionaria, es decir la media de la función de varianza es constante y la covarianza depende solo de la distancia entre los puntos muestreados (Gráfico 22-3, (a)), sin embargo, con el fin de tener continuidad espacial surgió la necesidad de generar puntos sistemáticos con un radio de 15 km que corresponde a cada grilla en base, dando como resultado 29, mismas que fueron obtenidos con la herramienta $fishnet$ del ArcGIS (licencia ESPOCH). En el gráfico 22-3, (b) se evidencia de forma más clara la estacionariedad, debido a la cantidad de puntos con los que se pretende trabajar.

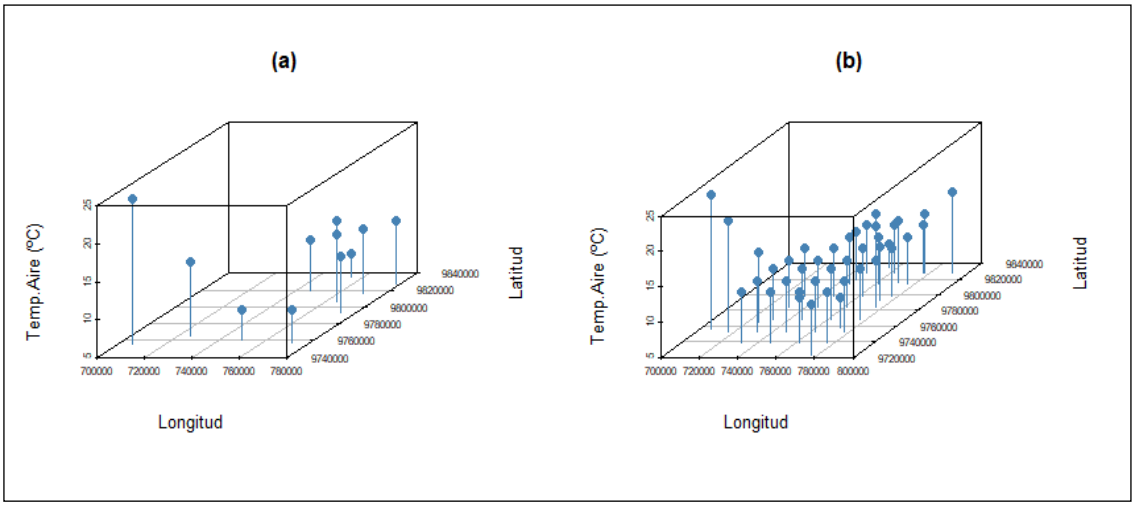


Gráfico 22-3: Dispersograma de temperatura para 11 (a) y 29 (b) sitios en Chimborazo.

Realizado por: Checa G., Marisol C.,2020.

3.6 Kriging Ordinario para datos funcionales de temperatura

La estimación espacial por medio de Kriging Ordinario para Datos Funcionales (OKFD) se llevó a cabo, tomando como base las curvas de temperaturas medias producidas con el conjunto de datos obtenidos de las estaciones meteorológicas de Chimborazo (Gráfico 23-3).

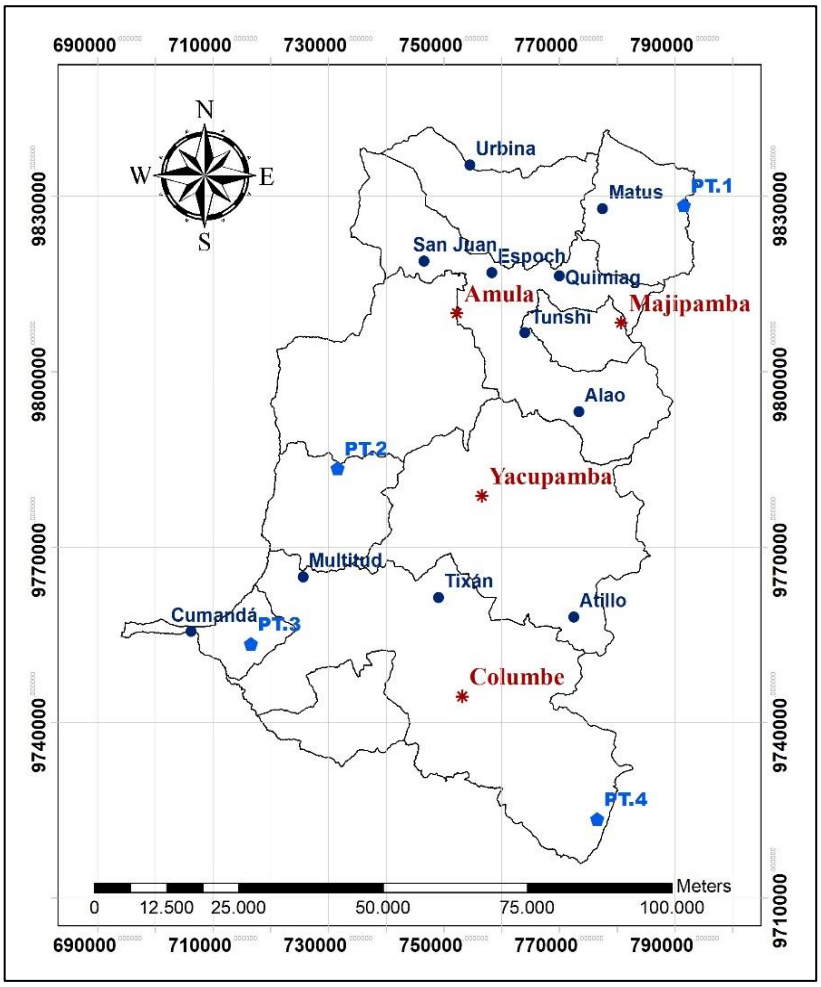


Gráfico 23-3: Localización de las estaciones meteorológicas, puntos sistemáticos y a estimar (rojo) en Chimborazo.

Fuente: (INEC, 2010).

Realizado por: Checa G., Marisol C.,2020.

En segunda instancia, se efectuó una estimación en un sitio no muestreado, usando el predictor de la ecuación 23.1 con el modelo de semivariograma seleccionado, luego con el fin de verificar la bondad del predictor se utilizó la técnica de VCF (dejar un dato fuera), y trabajar del siguiente modo: cada curva de temperatura se retiró del conjunto de datos funcionales y mediante una función de suavización, se realizó la estimación de la curva $\boldsymbol{\mathcal{X}}_{s0}(t)$ en el punto $\boldsymbol{s}_0$.

Este análisis geoestadístico de datos funcionales se realizó mediante los comandos del paquete $geoR$ y $geofd$ en R, los cuales proporcionan funciones para la selección del modelo de semivariograma y la estimación kriging (Giraldo et al., 2012).

3.6.1 OKFD para curvas medias de temperatura diaria por hora

En el gráfico 24-3 se puede observar las curvas medias de temperatura diaria por hora de las 11 estaciones durante los años 2014, 205, 2016 y 2017, donde Cumandá ubicada a 331 m s.n.m es la estación con mayor temperatura, a diferencia de Urbina que se localiza a 3646 m s.n.m y posee la temperatura más baja.

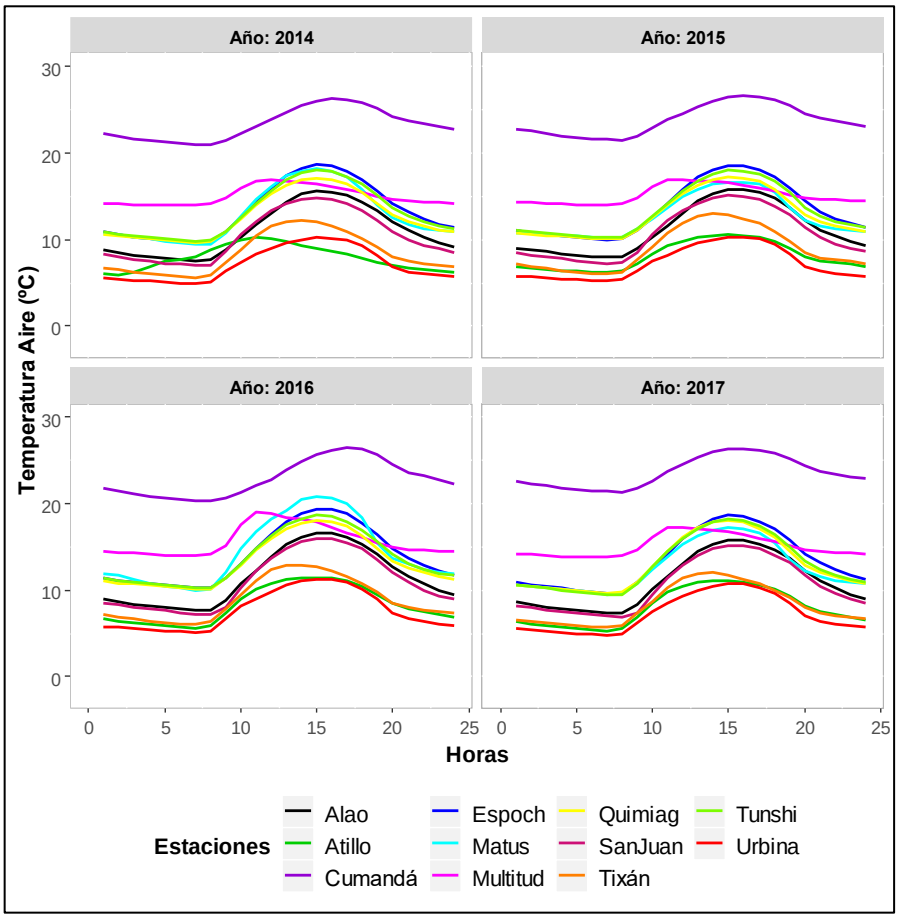


Gráfico 24-3: Curvas medias de temperatura diaria de las 11 estaciones de Chimborazo.

Realizado por: Checa G., Marisol C.,2020.

3.6.1.1 Análisis estructural

Para decidir qué modelo de semivariograma usar en la estimación por OKFD se probó mediante validación cruzada cuatro ejemplos de semivariogramas: esférico, exponencial, Gaussiano y Matérn (Gráfico 25-3), dejando Alao para la estimación como un sitio no muestreado.

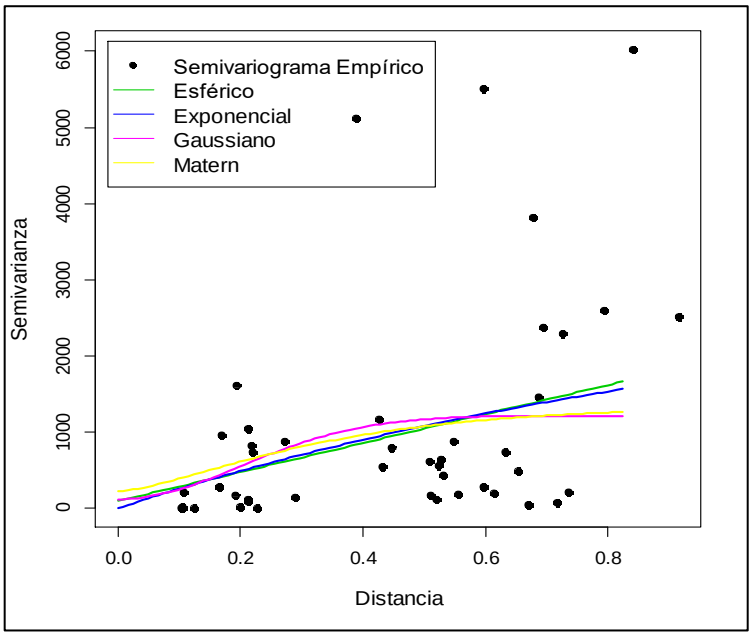


Gráfico 25-3: Semivariograma experimental, 2014.

Realizado por: Checa G., Marisol C.,2020.

En la tabla 5-3 se puede evidenciar que los modelos exponencial y Matérn presentan un comportamiento similar, siendo los más adecuados para la estimación dado que presentan los errores más bajos a diferencia del esférico y el Gaussiano.

Tabla 5-3: Principales tipos de kriging y sus propiedades (Caso 1).

Estadístico SSE	Esférico	Exponencial	Gaussiano	Matérn
Mínimo	0.07877	0.007168	0.02088	0.007168
1st Qu.	0.33792	0.243970	0.79835	0.243970
Mediana	3.10661	0.583821	4.51916	0.583821
Media	2.71453	0.774563	15.57019	0.774563
3nd Qu.	4.74890	1.324690	27.71617	1.324690
Máximo	6.46608	2.537926	56.37644	2.537926

Realizado por: Checa G., Marisol C.,2020.

Sin embargo, se utilizó el modelo exponencial, ya que el Matérn presentó resultados análogos.

3.6.1.2 Validación Cruzada Funcional

En el gráfico 26-3 se puede visualizar las curvas de temperatura diaria estimadas mediante VCF aplicando los cuatro modelos de semivariograma, donde se obtuvo la suma de cuadrados del error (SSE) para cada una de las estaciones durante las iteraciones que se llevó a cabo; con esto se ratificó que el modelo exponencial es el más adecuado para la estimación de temperatura diaria por hora.

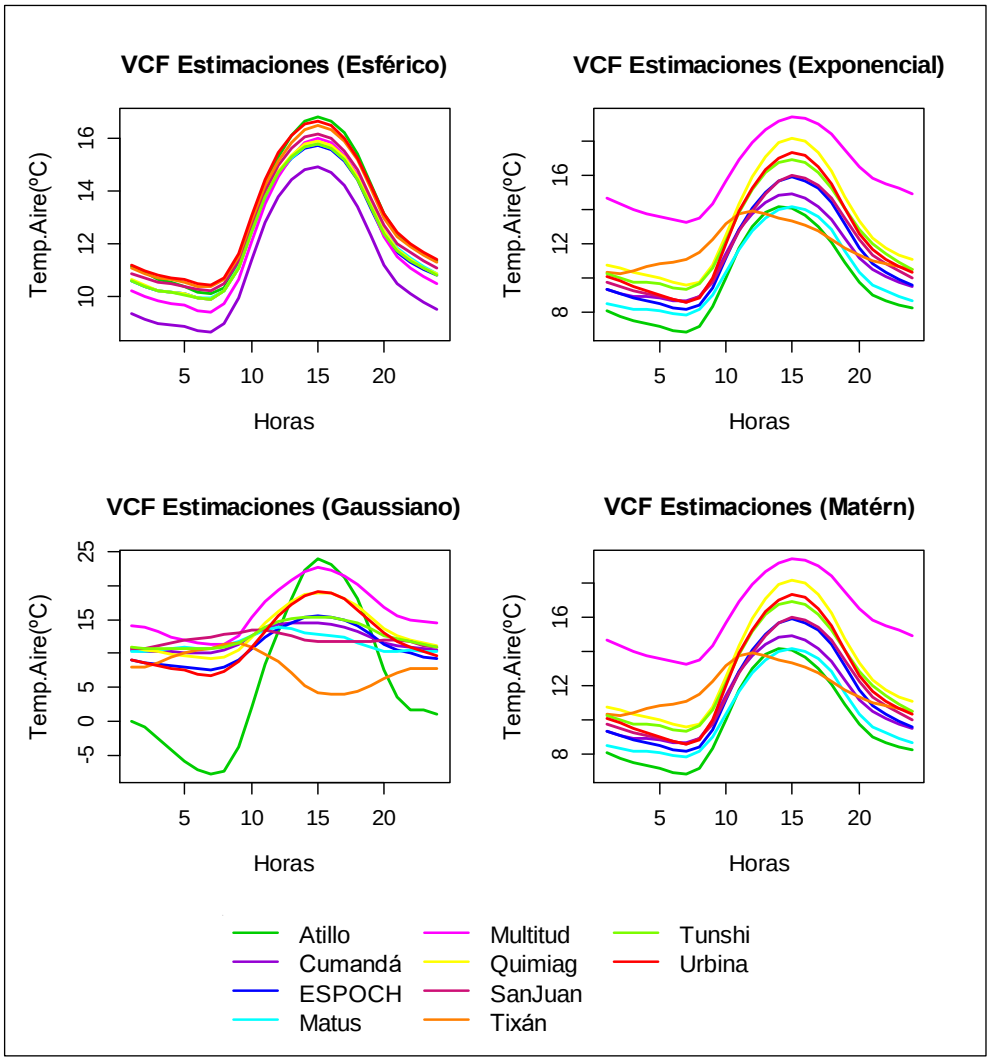


Gráfico 26-3: Curvas medias de temperatura diaria estimadas por VCF, 2014.

Realizado por: Checa G., Marisol C.,2020.

En el gráfico 27-3 se muestra los residuos obtenidos entre la temperatura estimada y original de cada una de las estaciones, en el cual se observa errores aproximadamente entre [-8, 14] aproximadamente.

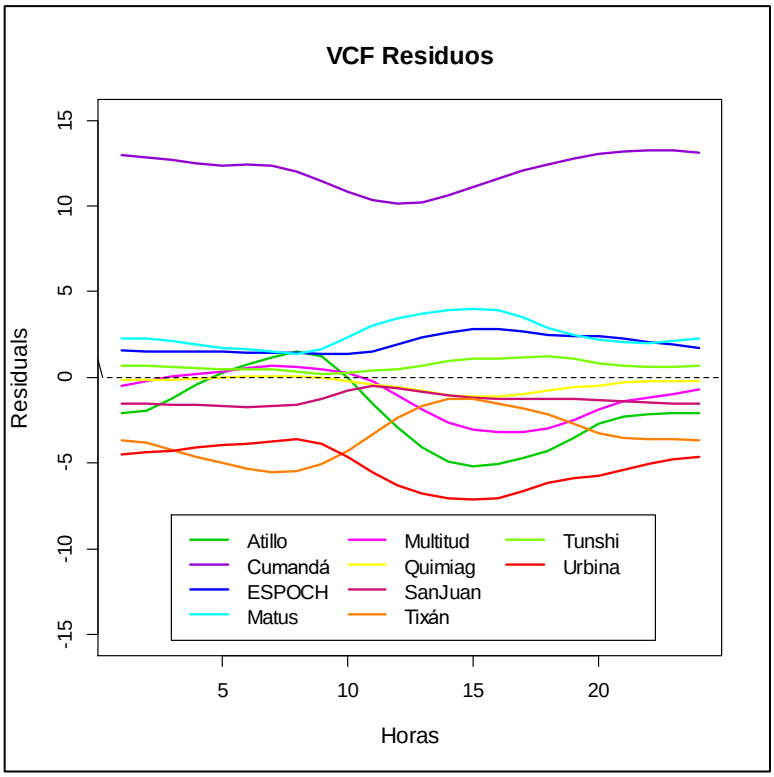


Gráfico 27-3: Residuos de VCF de las estaciones, 2014.

Realizado por: Checa G., Marisol C.,2020.

La tabla 6-3 indica el resumen de la SSE que se obtuvo en la evaluación del modelo exponencial a través de validación cruzada funcional para el OKFD, mostrando SSE medio de 403.25 ºC, y una desviación estándar de 836.76 ºC.

Tabla 6-3: Resumen de la SSE de validación cruzada para el OKFD, 2014.

Estadístico	OKFD
Mínimo	7.284
1st Qu.	30.819
Mediana	96.142
Media	403.253
3nd Qu.	244.508
Máximo	2855.082
Desviación Típica	836.7603
Suma	4435.781

Realizado por: Checa G., Marisol C.,2020.

3.6.1.3 Estimación

Finalmente, una vez determinado el modelo para la estimación de la temperatura diaria por hora para la provincia de Chimborazo, se halló las estimaciones para el sitio no muestreado.

En el gráfico 28-3 se observa la curva de temperatura estimada (rojo) para Alao comparada con la temperatura original (negra) de dicha estación.

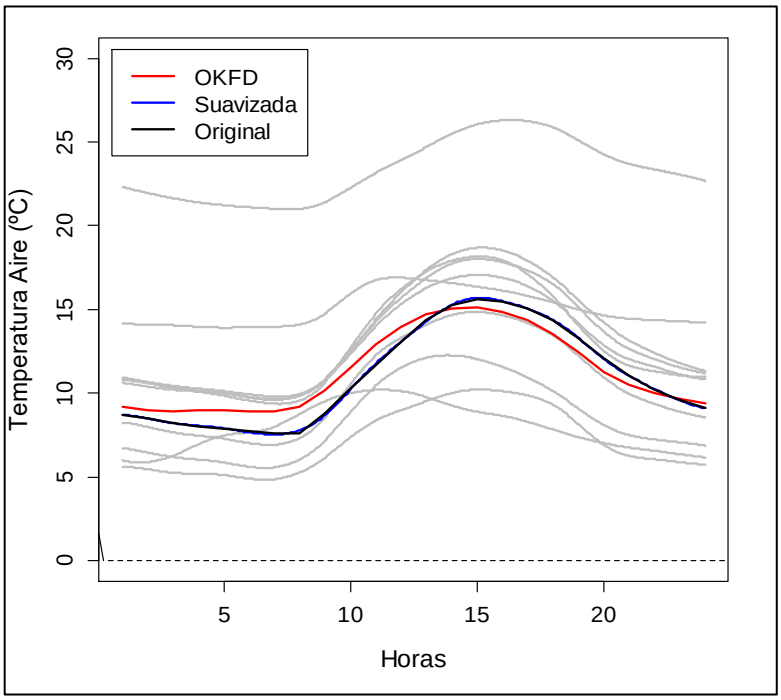


Gráfico 28-3: Estimación de temperatura diaria, Alao 2014.

Realizado por: Checa G., Marisol C.,2020.

La técnica de estimación kriging ordinario funcional se basa en el análisis del vecino más cercano, encontrando así que los resultados coinciden con dicha aseveración, puesto que los coeficientes que mayor peso tuvieron en la estimación de temperatura en Alao, correspondieron a las estaciones de Tunshi, Atillo y Quimiag, mismas que están localizados cerca al lugar de estimación (Tabla 7-3). La SSE para la estimación de temperatura de Alao fue de 18.58.

Tabla 7-3: Coeficientes del OKFD y distancias más representativas, Alao 2014.

	Tunshi	**Atillo**	**Quimiag**	**Cumandá**
$\boldsymbol{\lambda_i}$	0.580	0.261	0.150	-0.031
Distancia (m)	16472	35134	23489	77023

Realizado por: Checa G., Marisol C.,2020.

Se realizó el mismo análisis de estimación para las estacione de ESPOCH, Quimiag y Tunshi en los años 2015, 2016 y 2017 respectivamente (Gráfico 29-3).

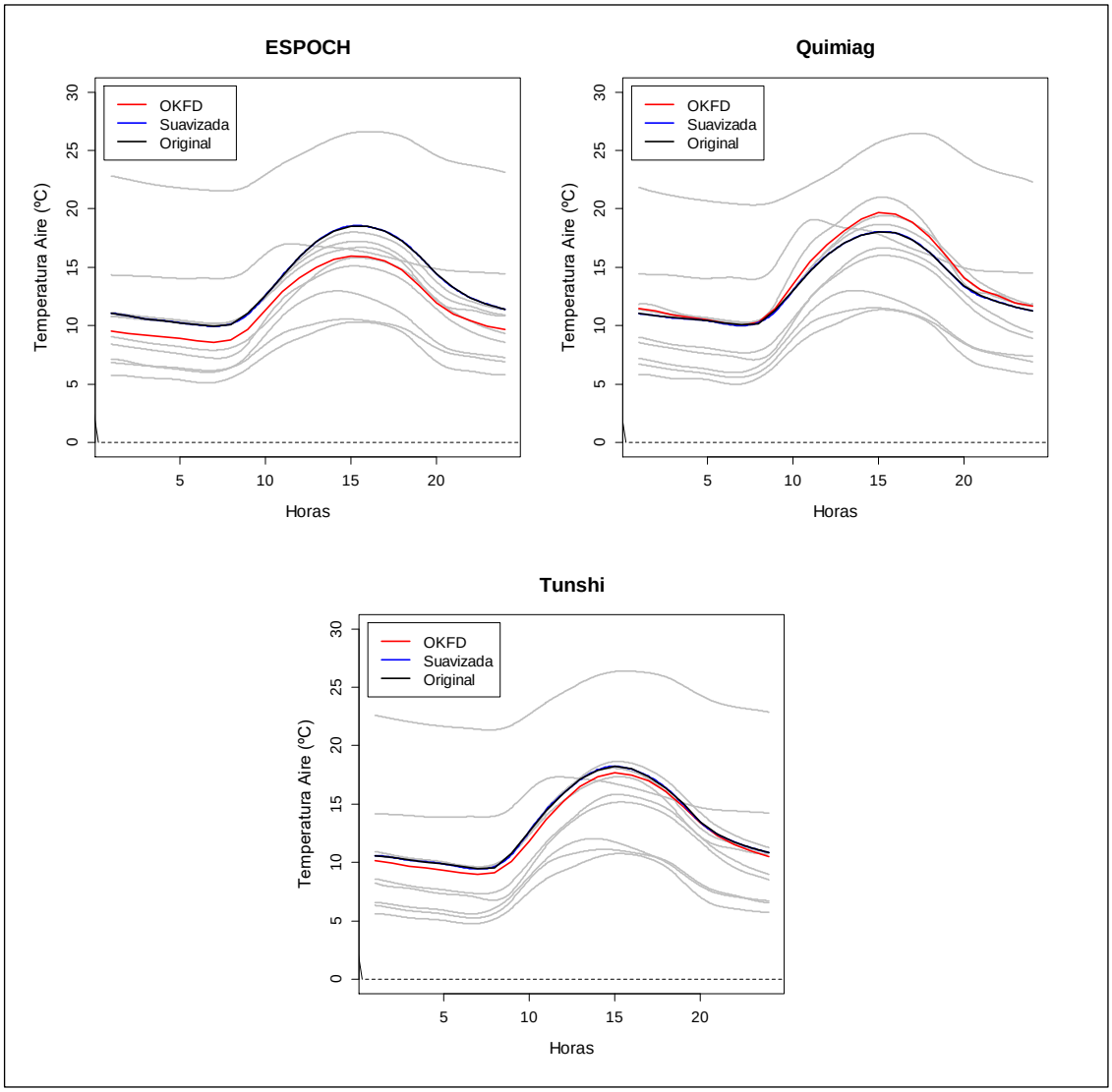


Gráfico 29-3: Estimación de temperatura diaria por horas.

Realizado por: Checa G., Marisol C.,2020.

Con una SSE para la estimación de temperatura del aire en las estaciones ESPOCH de 88.69, Quimiag de 17.17 y Tunshi de 6.

3.6.2 OKFD para curvas medias de temperatura anual por día

La estimación por OKFD y el procedimiento de VCF es mejor mientras mayor sea el tamaño de la muestra de puntos georreferenciados, por lo que, en principio se consideró las temperaturas descargadas de la página oficial de la NASA (https://power.larc.nasa.gov/) correspondientes a los 29 puntos sistemáticos, sin embargo, la estimación se realizó con tan solo 15 (11 estaciones y 4 sistemáticos), debido a que los 25 restantes no fueron significativos (temperatura iguales a los 4 puntos tomados) y sesgan la estimación.

Para realizar la estimación por OKFD se realizó previamente un análisis estructural y validación cruzada funcional del modelo seleccionado. A continuación, se muestra las estimaciones considerando la temperatura media anual de las 11 (estaciones) y 15 datos georreferenciados, con el fin de identificar diferencias significativas entre los dos tipos de análisis (Gráfico 30-3).

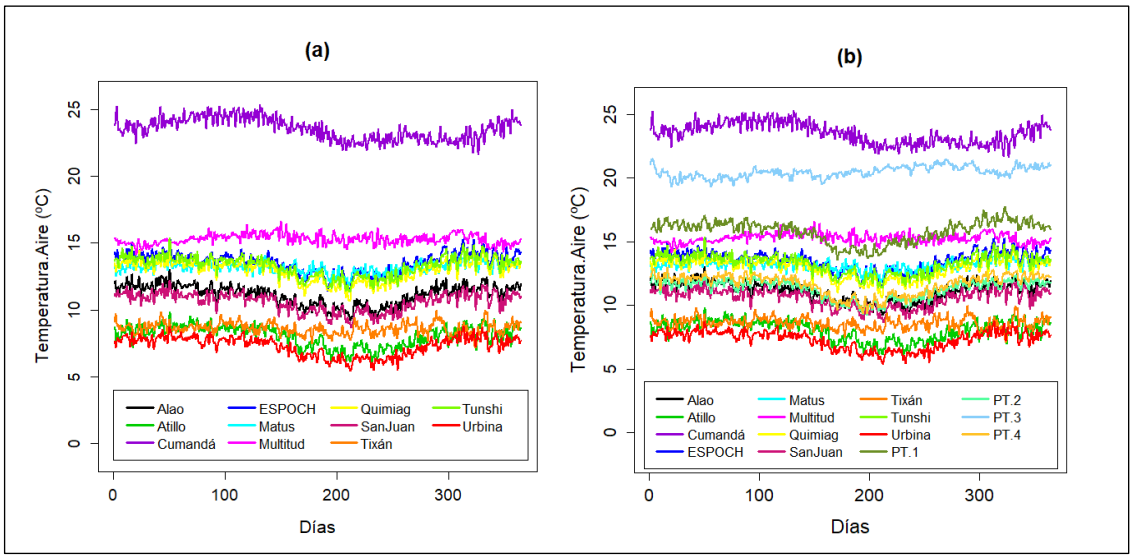


Gráfico 30-3: Curvas medias de temperatura anual en 11 (a) y 15 sitios (b) de Chimborazo.

Realizado por: Checa G., Marisol C.,2020.

No se realizó análisis con datos diarios por hora, debido a que la página oficial de la NASA no proporciona información con ese detalle.

3.6.2.1 Análisis estructural

Para decidir qué modelo de semivariograma usar para la estimación con las 11 y 15 localizaciones, se probó cuatro ejemplos de semivariogramas teóricos: Esférico, Exponencial, Gaussiano y Matérn (Gráfico 31-3) para la estimación por OKFD, al igual que en el caso anterior.

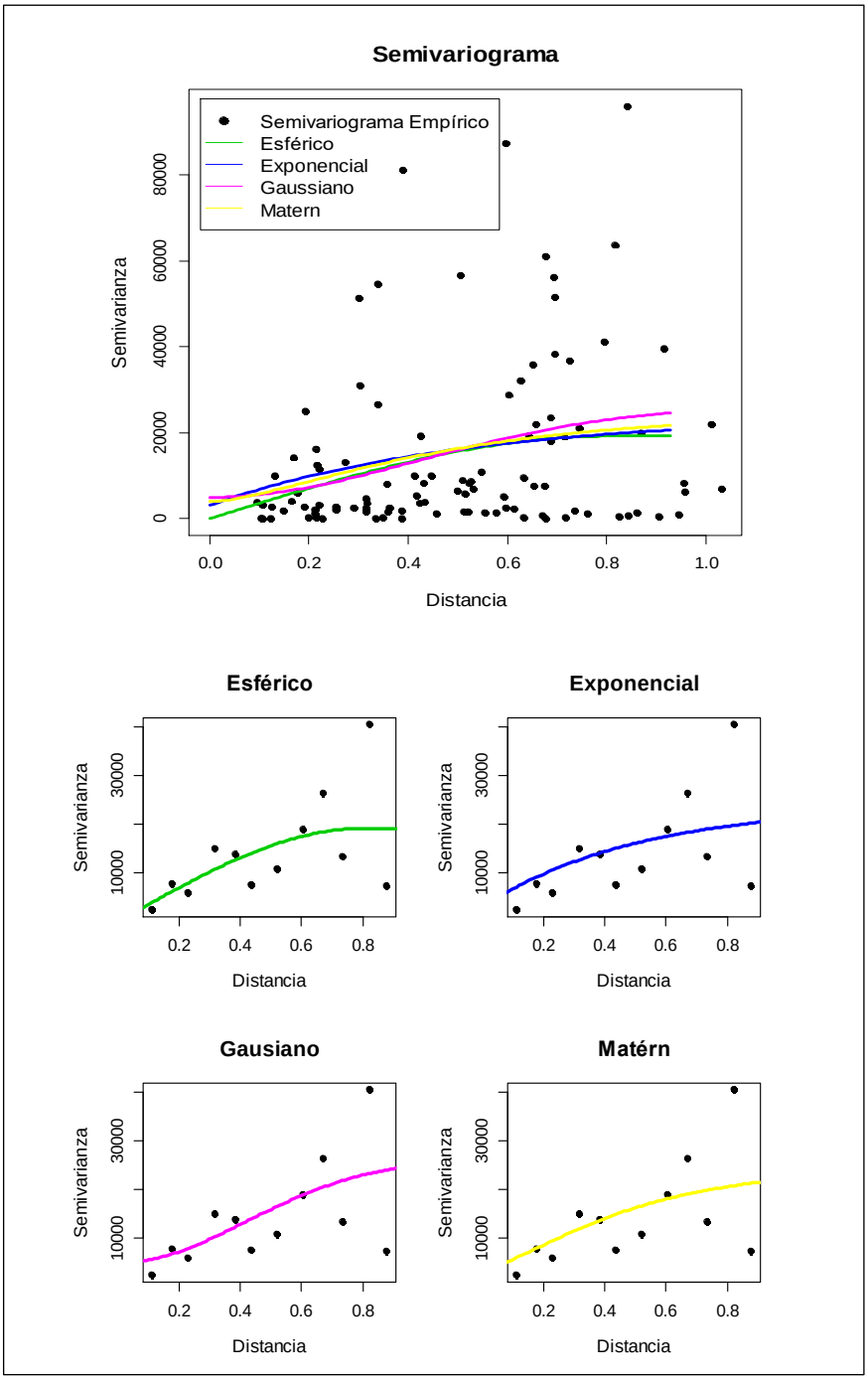


Gráfico 31-3: Semivariograma experimental y modelos teóricos ajustados.

Realizado por: Checa G., Marisol C.,2020.

En este caso, donde se evalúa la temperatura anual por día en las 11 y 15 localizaciones de Chimborazo, se puede evidenciar que los modelos esféricos, exponencial y Matérn son los más adecuados, debido a que presentan errores bajos para la estimación a diferencia del Gaussiano (Tabla 8-3).

Tabla 8-3: Principales tipos de kriging y propiedades (Caso 2).

	Con 11 estaciones				Con 15 sitios			
Estadístico	**Esférico**	**Exponencial**	**Gaussiano**	**Matérn**	**Esférico**	**Exponencial**	**Gaussiano**	**Matérn**
Mínimo	0.01	0.01	1.53	0.01	0.00	0.00	0.89	0.00
1st Qu.	1.05	1.05	5.61	1.05	0.50	0.63	4.06	0.63
Mediana	1.58	1.58	6.86	1.58	0.86	1.03	5.03	1.03
Media	1.72	1.72	7.11	1.72	0.97	1.14	5.22	1.14
3nd Qu.	2.22	2.22	8.50	2.22	1.29	1.50	6.17	1.50
Máximo	5.69	5.69	14.92	5.69	3.58	3.89	10.42	3.89

Realizado por: Checa G., Marisol C.,2020.

Sin embargo, para fines prácticos se seleccionó el modelo esférico, ya que los resultados fueron análogos con los otros dos modelos.

3.6.2.2 Validación Cruzada Funcional

Una modelación geoestadística exige una validación a posteriori de sus resultados, por ello para evaluar la bondad de ajuste del modelo de semivariograma seleccionado para la estimación por kriging en sitios no muestreados, se empleó VCF para comparar las curvas estimadas y observadas de temperatura para los 11 y 15 puntos georreferenciados.

En el gráfico 32-3 se muestra las curvas de temperatura estimadas mediante VCF aplicando los 4 modelos de semivariograma, se obtuvo la suma de cuadrados del error en cada una de las iteraciones dadas, cabe mencionar que el análisis con 15 sitios mejora notablemente las estimaciones de Cumandá, mostrando estimaciones de temperatura con comportamientos similar a las curvas originales en las diferentes estaciones (Gráfico 30-3, (b)); con esto se ratificó que el modelo esférico es el más óptimo para este proyecto de investigación.

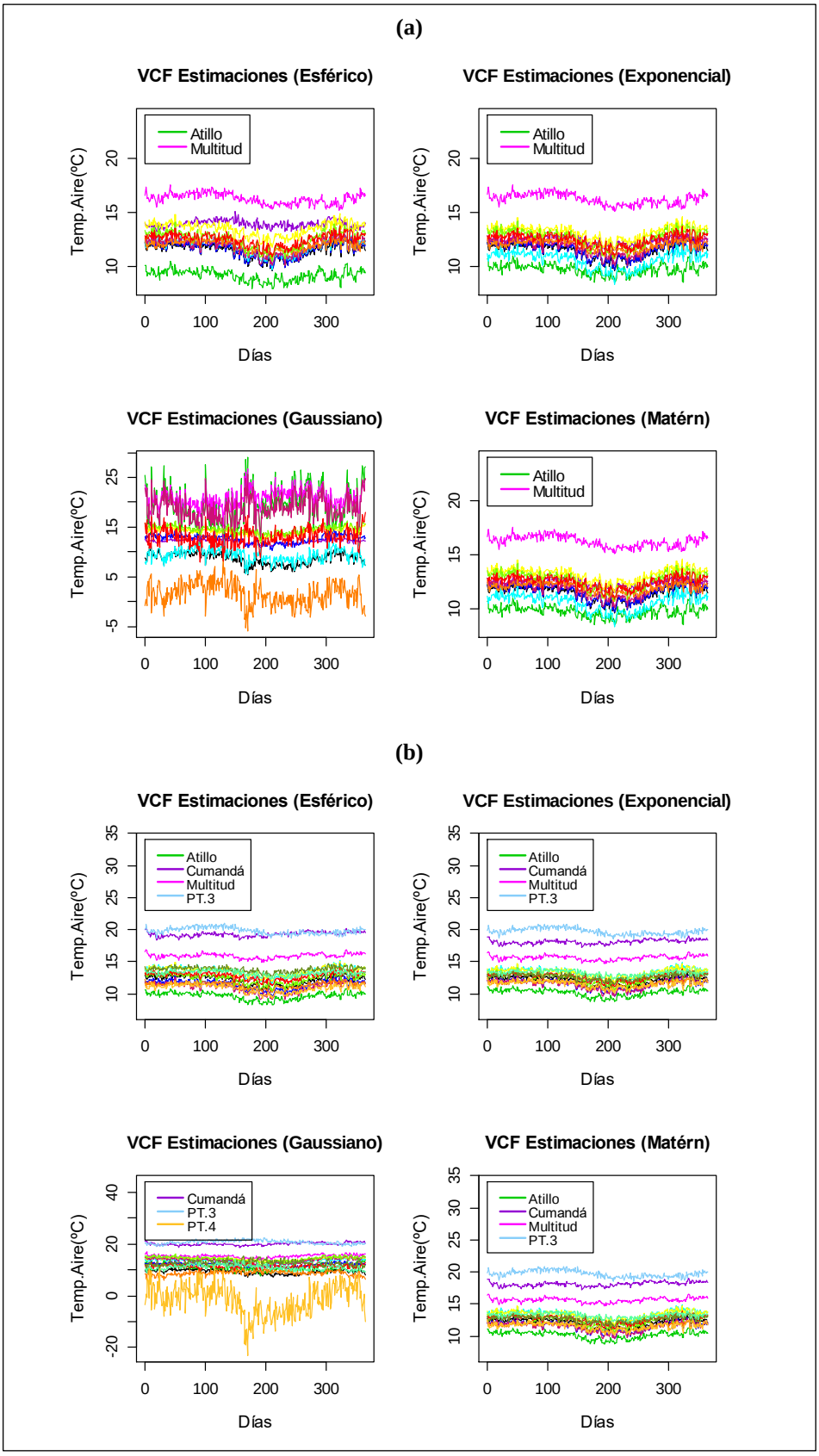


Gráfico 32-3: Curvas medias de temperatura anual estimadas por VCF para 11 (a) y 15 (b) sitios en Chimborazo.

Realizado por: Checa G., Marisol C.,2020.

El gráfico 33-3 muestra los residuos obtenidos con 11 y 15 puntos georreferenciados, en el cual se observó errores aproximadamente entre [-6, 13] y [-6, 6] respectivamente, concluyendo así que con 15 datos espaciales la estimación es mejor.

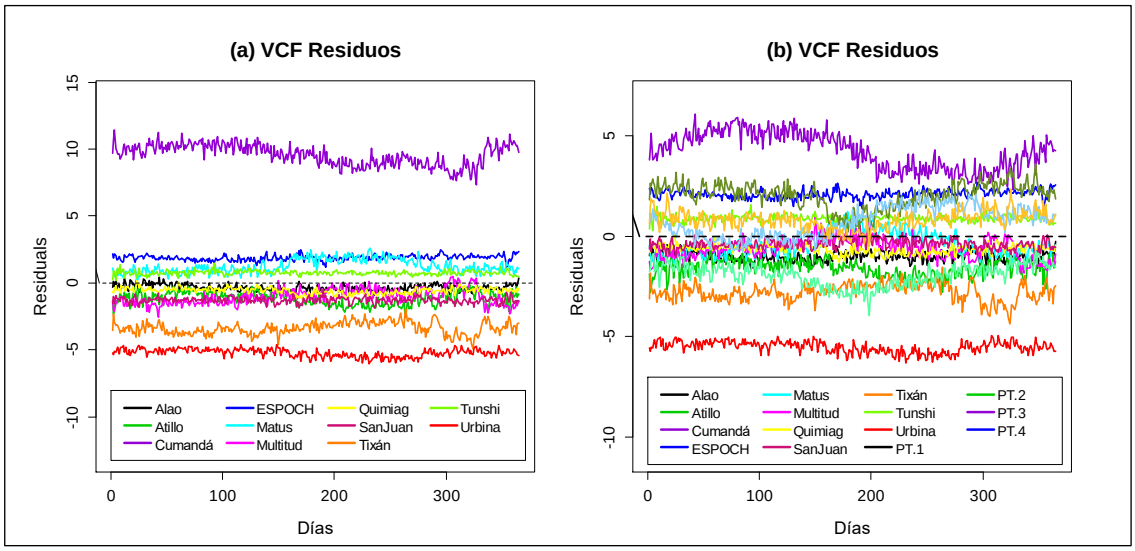


Gráfico 33-3: Residuos de VCF para 11 (a) y 15 (b) sitios en Chimborazo.

Realizado por: Checa G., Marisol C.,2020.

La tabla 9-3 muestra el resumen de la SSE que se obtuvo en la evaluación del modelo esférico mediante validación cruzada funcional para el Kriging Ordinario para los 11 y 15 puntos de referencia, mostrando SSE medio de 4718.87 ºC y 1913.51ºC, y desviación estándar de 10045.26 ºC y 3102.62 ºC respectivamente, con dicho análisis se ratifica que los 4 puntos sistemáticos agregados si mejoran la estimación.

Tabla 9-3: Resumen de la SSE de la validación cruzada para el OKFD.

Estadísticos	OKFD (11) ºC	OKFD (15) ºC
Mínimo	66.92	76.89
1st Qu.	357.63	264.8
Mediana	613.43	403.280
Media	4718.87	1913.51
3nd Qu.	2760.85	1628.11
Máximo	33678.15	11161.87
Desviación Típica	10045.36	3102.62
Suma	51906.52	28702.71

Realizado por: Checa G., Marisol C.,2020.

3.6.2.3 Estimación

Una vez identificado el mejor modelo de estimación de la temperatura para la provincia de Chimborazo, se halló las estimaciones en sitios no muestreados, mismos que fueron proporcionados por el GEAA y que corresponden a sembríos de Quinua en: Amulá Casaloma (Cacha), Majipamba (Chambo), San Pedro de Yacupamba (Guamote) y Columbe Grande (Alausí) (Gráfico 23-3), cuyos sectores se caracterizan por la siembra y cosecha de quinua, alimento ancestral y que constituye un aporte para la economía de la provincia.

En el gráfico 34-3 se presenta la curva de temperatura estimada (rojo) para Amulá Casaloma, misma que fue comparada con la temperatura descargada de la página oficial de la NASA (negra).

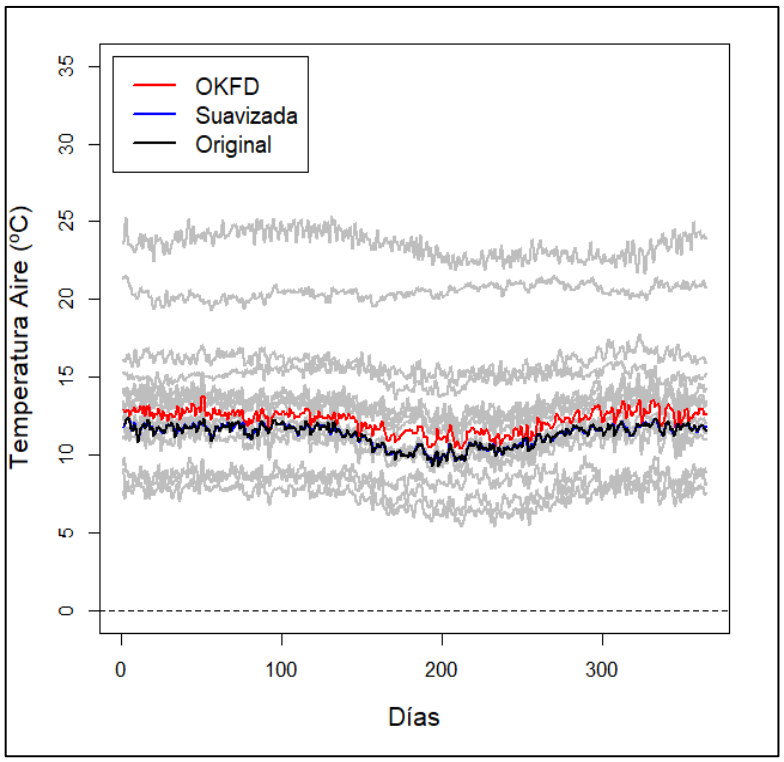


Gráfico 34-3: Estimación de temperatura anual, Amulá Casaloma.

Realizado por: Checa G., Marisol C.,2020.

La técnica de estimación kriging ordinario funcional se basa en el análisis del vecino más cercano, encontrando así que los resultados coinciden con dicha aseveración, puesto que los coeficientes del kriging que mayor peso tuvieron para la estimación de temperatura en Amulá correspondieron a las estaciones de San Juan y ESPOCH, mismos que geográficamente son los más cercanos a Amulá (Tabla 10-3). La SSE para la estimación de temperatura de Amulá fue de 355.13.

Tabla 10-3: Coeficientes del OKFD y distancias más representativas, Amulá Casaloma.

	San Juan	ESPOCH	Tunshi	PT.4
λ_i	0.379	0.346	0.307	-0.008
Distancia (m)	10578	9229	12187	89882

Realizado por: Checa G., Marisol C.,2020.

Se realizó el mismo análisis de estimación para las zonas no muestreadas: Majipamba, San Pedro de Yacupamba y Columbe Grande (Gráfico 35-3) en la provincia de Chimborazo.

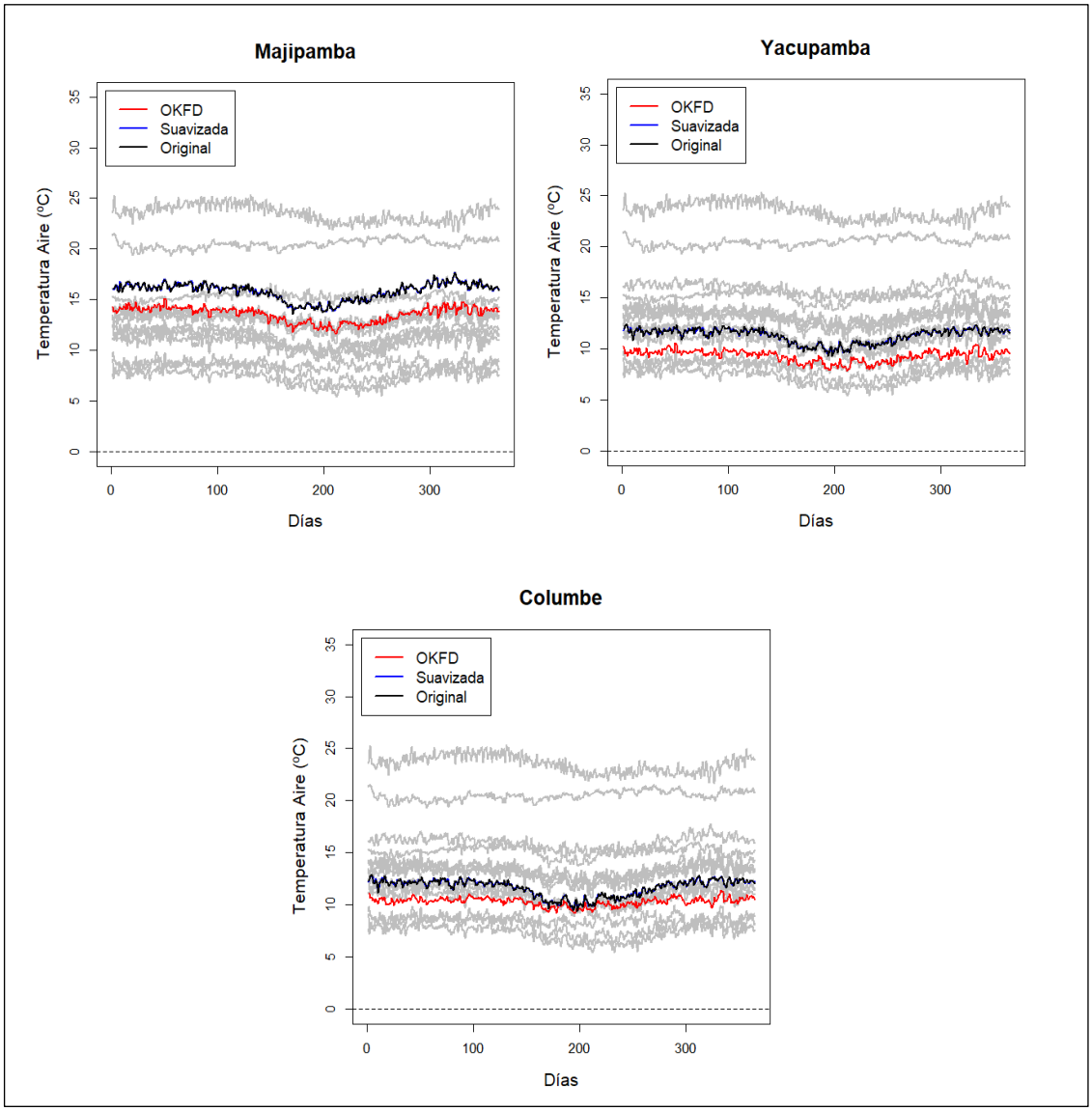


Gráfico 35-3: Estimación de temperatura anual en sitios no muestreados.

Realizado por: Checa G., Marisol C.,2020.

Con una SSE de 1878.12, 1465.88 y 765.05 respectivamente.

Se realizó mapas de temperatura promedio del aire (2014-2017) en horas del día (07:00 – 19:00) y la noche para observar los cambios de temperatura que surgen en estos lapsos de tiempo, ya que su comportamiento es esencial, por ejemplo, para la realización de estimaciones de evapotranspiración (ET) junto con otras variables meteorológicas (radiación solar, velocidad de viento y presión de vapor), para analizar un escenario de cambio climático y que sirva como guía para el desarrollo de futuros proyectos (Doorenbos y Pruitt, 1997; citado en Goyal, 2004).

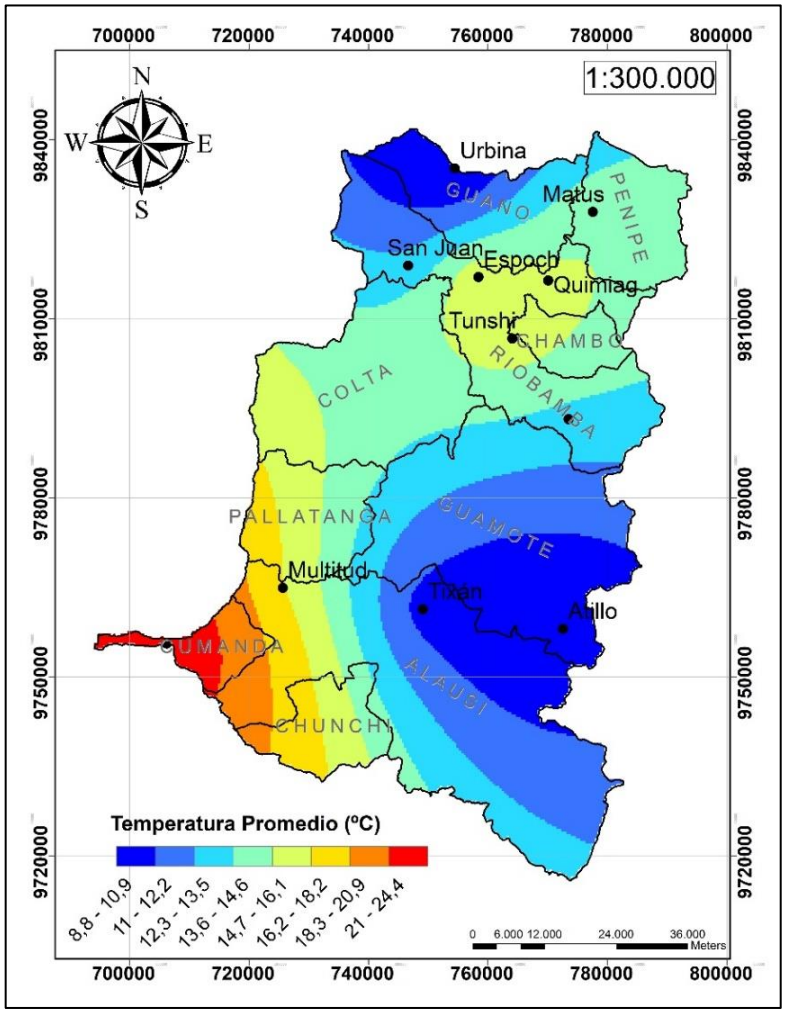


Gráfico 36-3: Mapa de temperatura promedio anual en horas de la mañana en la provincia de Chimborazo, 2014-2017.

Fuente: (INEC, 2010).

Realizado por: Checa G., Marisol C.,2020.

El gráfico 36-3, muestra temperaturas altas en el cantón Cumandá donde se ubicó la estación meteorológica que lleva el mismo nombre, con un promedio de temperatura anual en el día entre 21 ºC y 24.4 ºC, mientras que en los cantones de Guano, Guamote y Alausí donde se encuentran las estaciones de Urbina, Atillo y Tixán respectivamente registran valores bajos que varían entre 8.8ºC y 10.6 ºC.

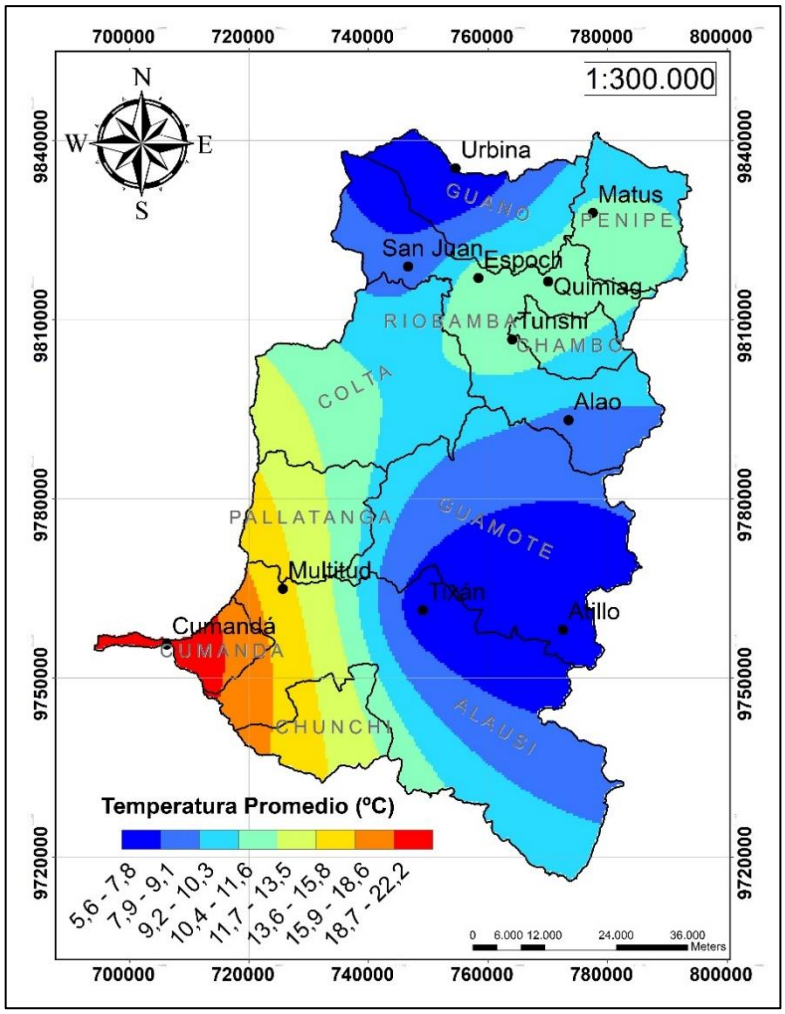


Gráfico 37-3: Mapa de temperatura promedio anual en horas de la noche en la provincia de Chimborazo, 2014-2017.

Fuente: (INEC, 2010).

Realizado por: Checa G., Marisol C.,2020.

El gráfico 37-3 muestra que existió valores altos de temperatura promedio anual en la noche entre 18.7ºC y 22.2 ºC en Cumandá, mientras que en la noche en los cantones: Guano, Guamote y Alausí la temperatura se encontró entre 5.6 ºC y 7.8 ºC

Al comparar los dos mapas de temperatura promedio anual (día y noche) se evidenció que la temperatura decreció, especialmente en la ciudad de Riobamba de 15.8ºC a 11.6ºC aproximadamente.

CONCLUSIONES

Se validó los datos de temperatura mediante las especificaciones que presenta la OMM (Organización Mundial de Meteorología), identificando atípicos en las estaciones de Atillo 2014 y Matus 2015, mismo que fueron separados del análisis. Las estaciones que no superaron el 20% de faltantes fueron imputados mediante el enfoque semi-paramétrico de ajuste de la media predictiva (pmm).

Se consideró como dato funcional la temperatura diaria por horas y anual por días. El suavizado de las curvas diarias se realizó mediante B-Splines Cúbico con 15 bases, mientras que en el segundo caso se utilizó la base Fourier con 365 bases debido a la presencia de periodicidad. El número de bases se determinó mediante el comando $min.basis(\,)$ de R y el criterio de validación cruzada generalizada cuya varianza residual entre los datos observados y suavizados fue de 0.2382 y 0.047 respectivamente.

Se realizó un análisis exploratorio funcional de las curvas de temperatura del aire (ºC) donde se encontró un total de 101 funciones atípicas pertenecientes a ESPOCH, Matus, Quimiag, San Juan y Tunshi en el año 2015, los cuales fueron separados para el análisis. Se cálculo la función media y desviación estándar, y mediante el FANOVA se determinó que las curvas medias por hora entre los años 2014, 2015, 2016 y 2017 de cada una de las estaciones meteorológicas son significativamente diferentes por lo que se realizó el análisis por año, mientras que no existe diferencia significativa en las curvas medias por día, por lo que se utilizó el promedio de los 4 años para continuar con el análisis.

La modelación OKFD es mejor mientras más datos muestrales se disponga, motivo por el cual aparte de las 11 estaciones meteorológicas existentes se generó 29 puntos sistemáticos mediante $Create\ Fishnet$ de ArcGIS en un radio de 15 km., sin embargo, solo 4 fueron representativos; motivo por el cual se contó con 15 datos con distribución espacial estacionaria requisito indispensable para el análisis geoestadístico de datos funcionales, a más de un mejor ajuste en la estimación de la temperatura del aire.

La modelación de la temperatura del aire (ºC) se realizó mediante el análisis estructural por validación cruzada funcional (VCF) utilizando cuatro modelos de semivariograma: esférico, exponencial, Gaussiano y Matérn, los cuales permitieron analizar el comportamiento espacial de la variable en estudio sobre el área definida. El modelo con mejor ajuste para las curvas medias de temperatura fue el esférico.

La estimación de temperatura se realizó en zonas de cultivo de quinua como: Amulá Casaloma (Cacha-Riobamba), Majipamba (Chambo), San Pedro de Yacupamba (Guano) y Columbe Grande (Alausí), puesto que en esos sitios no existe estaciones meteorológicas. Las estimaciones fueron comparadas con las temperaturas descargadas de la NASA, cuya suma de cuadrados del error fueron: 355.13, 1878.12, 1465.88 y 765.05 respectivamente. Cabe mencionar que a medida que aumente la distancia entre la zona no muestreada y alguna estación meteorológica el SSE será mayor.

RECOMENDACIONES

Verificar el buen funcionamiento de los diferentes dispositivos de medición de las variables meteorológicas que se maneja en las estaciones, para evitar la existencia de datos faltantes y atípicos, que causen investigaciones con resultados sesgados.

Buscar otras fuentes de información que proporcionen datos de variables meteorológicas con detalle en la provincia de Chimborazo, además de la página oficial de la NASA, como punto de referencia para las estimaciones proporcionadas por esta investigación, y por ende mejorar la calidad de resultados de temperatura del aire mediante técnicas geoestadísticas de datos funcionales.

Utilizar la metodología descrita en este trabajo de investigación para las variables meteorológicas como: humedad relativa, presión barométrica, radiación solar difusa y global, temperatura del suelo, dirección y velocidad del viento.

Los docentes e investigadores de la carrera de Estadística promuevan el análisis de datos funcionales (ADF) sobre todo aplicado a la geoestadística, ya que va de la mano con el avance de las aplicaciones informáticas y dispositivos que permiten almacenar grandes cantidades de datos conservando muy bien la información de los mismos.

En la carrera de Estadística se realice cursos y mantenga la electiva de Análisis Estadístico de Datos Funcionales, debido a su aplicabilidad en varias áreas como: agrícola, energética, financiera, etc. sin perder gran parte de la información que se obtiene para el análisis.

GLOSARIO

Análisis de Datos Funcionales: Es aquella parte de la estadística que estudia y analiza información contenida en curvas o cualquier otro elemento que generalmente varía en el tiempo (Torrecilla, 2010, p. 5).

Base de funciones: Es un conjunto de funciones conocidas $\phi_k(t)$ tales que son matemáticamente independientes y al construir combinaciones lineales de ellas con un número *K* suficientemente grande de términos permiten aproximar la forma de cualquier curva (Ramsay y Silverman, 2005).

Bases de B-Splines: Son trozos de polinomio de grado *p* conectados entre sí, usados generalmente para datos no periódicos cuya función es localmente suave (Aguilera, 2009; Torrecilla, 2010; Escudero, 2016).

Bases de Fourier: Es una de las bases más antiguas, formada por funciones seno y coseno, utilizada para series temporales largas que muestran cierto tipo de periodicidad, es decir para funciones estables sin grandes variaciones y con curvatura más o menos constante (Ramsay y Silverman, 2005).

Dato Funcional: Es una curva que procede de la realización de un proceso estocástico en tiempo continuo (Aguilera, 2011, p. 4).

Estadística espacial: Es la reunión de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios de una región (Giraldo, 2002, p.8).

Geoestadística: Es una rama de la estadística que aborda el problema de predicción en una región con continuidad espacial, aplicada a diferentes disciplinas como: hidrología, meteorología, control ambiental, etc. (Caballero, 2011; Cardona, 2015).

Kriging Ordinario para Datos Funcionales: Procedimiento de kriging funcional donde la curva a predecir es una combinación lineal de las curvas observadas y los coeficientes son números reales (Ginzo, 2011, p. 53).

Meteorología: Es una ciencia que permite estudiar y predecir numerosos fenómenos que se producen en la atmósfera en espacio y tiempo, basada en el conocimiento de una serie de magnitudes o variables meteorológicas como: la temperatura, la radiación solar, la presión atmosférica o la humedad (Rodríguez et al., 2014, p. 6).

Semivariograma: Es la herramienta central de la geoestadística, que permite analizar el comportamiento espacial de una variable sobre un área definida (Giraldo, 2002; Ginzo, 2011; Cardona 2015).

Suavizado: Técnica para convertir los datos recogidos en forma discreta a una función y que trata de ajustarlos mediante una base de funciones, permitiendo eliminar el ruido registrado al obtener las observaciones (Aguilera, 2009; Millán, 2017).

Temperatura del aire: Es una de las variables meteorológicas más utilizadas para describir el estado de la atmósfera, la cual varía entre el día y la noche, entre una estación y otra, entre una ubicación geográfica y otra, debido a que está sometida a numerosas oscilaciones, está condicionada por la longitud, latitud y altura s.n.m (Rodríguez et al., 2014, pp. 12-15).

Validación Cruzada: Proceso iterativo en el que cada vez se excluye un dato de la muestra y se estima con el resto de los datos el modelo de semivariograma escogido, para predecir vía kriging el valor de la variable en estudio en el sitio del punto que se excluyó (Ginzo, 2011, p. 24).

BIBLIOGRAFÍA

AGUILERA MORILLO, M.C. *Estimación penalizada con datos funcionales* [En línea]. (Trabajo de Titulación). (Maestría). Universidad de Granada. España. 2009. pp. 1-76. [Consulta: 9 noviembre 2018]. Disponible en: https://masteres.ugr.es/moea/pages/tfm0809/estimacin-penalizada-con-datos-funcionales.

AGUILERA MORILLO, M.C. *Penalized estimation methods in functional data analysis* [En línea]. (Trabajo de Titulación). (Doctoral). Universidad de Granada. España. 2013. pp. 1-215. [Consulta: 9 noviembre 2018]. Disponible en: https://dialnet.unirioja.es/servlet/tesis?codigo=59229.

CALAFATI, R.O. *Estrategias para el tratamiento de datos faltantes («missing data») en estudios con datos longitudinales* [En línea]. (Trabajo de Titulación). (Maestría). Universidad Oberta de Catalunya. Barcelona - España. 2017. pp. 1-75. [Consulta: 21 noviembre 2018]. Disponible en: http://openaccess.uoc.edu/webapps/o2/handle/10609/64085.

CARDONA RÍOS, J.A. *Generación de superficies climáticas usando datos funcionales de temperatura y precipitación por medio de métodos geoestadísticos para el Valle del Río Cauca, Colombia* [En línea]. (Trabajo de Titulación). (Maestría). Universidad San Francisco de Quito. Quito-Ecuador. 2015. pp. 1-114. [Consulta: 29 octubre 2018]. Disponible en: http://repositorio. usfq.edu.ec /handle/23000/4096.

CASTRO CACABELOS, M. *Imputación de datos faltantes en un modelo de tiempo de fallo acelerado* [En línea]. (Trabajo de Titulación). (Maestría). Universidad de Vigo. Pontevedra, España. 2014. pp.1-53. [Consulta: 20 enero 2018]. Disponible en: http://eio.usc.es/pub/mte /descargas/ProyectosFin Master/Proyecto_940.pdf.

CUEVAS, A.; et al. An anova test for functional data. *Computational statistics & data analysis* [En línea], 2004, vol. 47, nº. 1, pp. 111–122. [Consulta: 21 noviembre 2018]. Disponible en: https://scholar.google.es/citations?user=f9Cg66kAAAAJ&hl=nl.

DÍAZ VIERA, M.A. *Geoestadística aplicada* [En línea]. Universidad Nacional Autónoma de México, Instituto de Geofísica y Astronomía. México. 2002. pp. 1-144. [Consulta: 2 octubre 2019]. Disponible en: https://www.academia.edu /23486534/Geoestadistica_Aplicada.

DUEÑAS HERRERA, M.P. *Análisis geoestadístico multivariado a través de métodos funcionales y curvas de Andrews* [En línea]. (Trabajo de Titulación). (Maestría). Universidad Nacional de Colombia. Bogotá-Colombia. 2017. pp. 1-75. [Consulta: 26 septiembre 2019]. Disponible en: http://bdigital.unal.edu.co/60946/.

ESCUDERO VILLA, A.I. *Modelos funcionales para el tratamiento de datos de Radiación Solar Global* [En línea]. (Trabajo de Titulación). (Maestría). Universidad de Granada. España. 2016. pp. 1-128. [Consulta: 10 junio 2019]. Disponible en: https://masteres.ugr.es/moea/pages/curso 201516 /tfm 1516/TFM_Escudero_Villa.

FEBRERO-BANDE, M.; & OVIEDO DE LA FUENTE, M. Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software* [En línea], 2012, vol. 51, nº. 1, pp. 1-28. [Consulta: 3 diciembre 2019]. ISSN 1548-7660. Disponible en: https://www. jstatsoft.org/index.php/jss/article/view/v051i04.

FERRATY, F.; & VIEU, P. *Nonparametric Functional Data Analysis: Theory and Practice* [En línea]. New York-EEUU: Springer Science & Business Media. 2006. ISBN 978-0-387-36620-3. [Consulta: 20 octubre 2018]. Disponible en: https://www.springer.com/la/book/9780387303697.

GALVÁN, M.; & MEDINA, F. *Imputación de Datos: Teoría y Práctica* [En línea]. Santiago de Chile: United Nations Publications. 2007. ISBN 978-92-1-323101-2. [Consulta: 11 febrero 2019]. Disponible en: https://repositorio.cepal.org//handle/11362/4755.

GINZO VILLAMAYOR, M.J. *Análisis geoestadístico de datos funcionales* [En línea]. (Trabajo de Titulación). (Maestría). Universidad Da Coruña. España. 2011. pp. 1-85. [Consulta: 7 julio 2018]. Disponible en: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_ 388.

GIRALDO HENAO, R. *Introducción a la Geoestadística. Teoría y aplicación* [En línea]. Bogotá-Colombia: Facultad de Ciencias, Departamento de Estadística, Universidad Nacional de Colombia. 2002. [Consulta: 5 enero 2019]. Disponible en: https://geoinnova.org/blog-territorio/ wp-content/uploads/2015/05/libro_-de-_geoestadistica-r-Giraldo.pdf.

GIRALDO HENAO, R. Análisis exploratorio de variables regionalizadas con métodos funcionales. *Revista Colombiana de Estadística* [En línea], 2007, vol. 30, nº. 1, pp. 115–127. [Consulta: 25 diciembre 2018]. Disponible en: http://www.bdigital.unal.edu.co/30496/1/29326-105305-1-PB.pdf.

GIRALDO HENAO, R. *Geostatistical analysis of functional data* [En línea]. (Trabajo de Titulación). (Doctoral). Universitat Politécnica de Catalunya (UPC). Barcelona-España. 2009. pp. 1-122. [Consulta: 1 octubre 2018]. Disponible en: https://dialnet.unirioja.es/servlet/tesis?cod=21518.

GIRALDO, R.; et al. Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics* [En línea], 2011, vol. 18, nº. 3, pp. 411-426. [Consulta: 1 octubre 2018]. ISSN 1573-3009. Disponible en: https://doi.org/10.1007/s10651-010-0143-y.

GIRALDO, R.; et al. geofd: An R Package for Function-Valued Geostatistical Prediction. *Revista Colombiana de Estadística* [En línea], 2012, vol. 35, nº. 3, pp. 385-407. [Consulta: 6 noviembre 2019]. ISSN 0120-1751. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0120-17512012000300004&lng=en&nrm=iso&tlng=en.

GOULARD, M.; & VOLTZ, M. Geostatistical Interpolation of Curves: A Case Study in Soil Science. En: A. SOARES (ed.), *Geostatistics Tróia '92* [En línea]. Dordrecht. 1993. Springer Netherlands, pp. 805-816. [Consulta: 5 noviembre 2018]. ISBN 978-0-7923-2157-6. Disponible en: http://link.springer.com/10.1007/978-94-011-1739-5_64.

GOYAL, R.K. Sensitivity of evapotranspiration to global warming: a case study of arid zone of Rajasthan (India). *Agricultural Water Management* [En línea], 2004, vol. 69, nº. 1, pp. 1-11. [Consulta: 10 enero 2020]. ISSN 0378-3774. Disponible en: http://www.sciencedirect.com/science/article/pii/S0378377404001441.

HYNDMAN, R.J.; & SHANG, H.L. Rainbow Plots, Bagplots, and Boxplots for Functional Data. *Journal of Computational and Graphical Statistics*, 2010, vol. 19, nº. 1, pp. 29-45. ISSN 1061-8600. DOI 10.1198 / jcgs.2009.08158.

INAMHI. Instituto Nacional de Meteorología e Hidrología. [En línea]. [Consulta: 15 octubre 2019]. Disponible en: http://www.serviciometeorologico.gob.ec.

INEC. Instituto Nacional de Estadística y Censos. [En línea]. 2010. [Consulta: 1 diciembre 2019]. Disponible en: https://www.ecuadorencifras.gob.ec.

ITURRALDE, M.L. Predicciones del tiempo y matemáticas. *Sigma: revista de matemática* [En línea], 2003, nº. 23, pp. 22. Disponible en: https://www.researchgate.net/profile/Mikel_Lezaun/publication/28184977_Predicciones_del_tiempo_y_matematicas/links/0f31752d6686ac1c8e000000/Predicciones-del-tiempo-y-matematicas.pdf.

MATEU, J.; & ROMANO, E. Advances in spatial functional statistics. *Stochastic Environmental Research and Risk Assessment* [En línea], 2017, vol. 31, nº. 1, pp. 1-6. [Consulta: 1 octubre 2018]. ISSN 1436-3259. Disponible en: https://link.springer.com/article/10.1007/s00477-016-1346-z.

MILLÁN ROURES, L. *Outliers de datos funcionales para la detección de caudales anómalos en el sector hidráulico* [En línea]. (Trabajo de Titulación). (Maestría). Universitat Jaume I. Castellón-España. 2017. pp. 1-111. [Consulta: 23 noviembre 2018]. Disponible en: http://repositori.uji.es/xmlui /handle/10234/174477.

MORAL GARCÍA, F.J. Aplicación de la geoestadística en las ciencias ambientales. *Revista Ecosistemas* [En línea], 2004, vol. 13, nº. 1. [Consulta: 25 noviembre 2019]. ISSN 1697-2473. Disponible en: https://revistaecosistemas.net/index.php/ecosistemas/article/view/582.

NARVÁEZ, R.P. Evaluación preliminar de la temperatura media en superficie del Ecuador para el año 2010, obtenida mediante el modelo Weather Research Forecasting (WRF). *ACI Avances en Ciencias e Ingenierías* [En línea], 2012, vol. 4, nº. 2. [Consulta: 15 octubre 2019]. ISSN 2528-7788. Disponible en: https://revistas.usfq.edu.ec/index.php/avances/article/view/110.

OMM. Organización Meteorológica Mundial. [En línea]. [Consulta: 19 marzo 2018]. Disponible en: https://public.wmo.int/en.

PÉREZ MONTILLA, A. *Métodos Avanzados de Análisis de Datos Funcionales* [En línea]. (Trabajo de Titulación). (Maestría). Universidad de Cádiz. España. 2018. pp. 1-66. [Consulta: 11 diciembre 2018]. Disponible en: http://rodin.uca.es:80/xmlui/handle/10498/20583.

PETITGAS, P. Geostatistics and their applications to fisheries survey data. En: B.A. MEGREY y E. MOKSNESS (eds.), *Computers in Fisheries Research* [En línea]. Dordrecht-Países Bajos. 1996. Springer Netherlands, pp. 113-142. [Consulta: 25 noviembre 2019]. ISBN 978-94-015-8598-9. Disponible en: https://doi.org/10.1007/978-94-015-8598-9_5.

***Plan de Desarrollo y Ordenamiento Territorial de Chimborazo*.** [En línea], Chimborazo-Ecuador: 2018. [Consulta: 2 enero 2019]. Disponible en: http://www.chimborazo.gob.ec.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. [En línea]. 2018. [Consulta: 14 septiembre 2018]. Disponible en: https://www.r-project.org/.

RAMSAY, J.; et al. *Functional Data Analysis with R and MATLAB*. New York - EEUU: Springer Science & Business Media. 2009. ISBN 978-0-387-98185-7.

RAMSAY, J.; & SILVERMAN, B.W. *Functional Data Analysis*. 2a ed. New York - EEUU: Springer-Verlag. Springer Series in Statistics, 2005. ISBN 978-0-387-40080-8.

RAMSAY, J.O.; & DALZELL, C.J. Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society. Series B*. 1991, vol. 53, nº. 3, pp. 539-572. ISSN 0035-9246.

RODRÍGUEZ JIMÉNEZ, R.M.; et al. *Meteorología y climatología* [En línea]. España: s.n. 2014. [Consulta: 18 enero 2018]. Disponible en: https://www.tysmagazine.com/libro-gratuito-meteorologia-climatologia/.

SALMERÓN GÓMEZ, R. *Análisis estadístico de datos espacio-temporales mediante modelos funcionales de series temporales* [En línea]. (Trabajo de Titulación). (Doctoral). Universidad de Granada. España. 2008. pp. 1-161. [Consulta: 9 enero 2018]. Disponible en: https://dialnet.unirioja.es/servlet/tesis?codigo=21049.

SANTOFIMIA NAVARRO, M.E. *Predicción de picos de polen mediante regresión logística funcional* [En línea]. (Trabajo de Titulación). (Maestría). Universidad de Granada. España. 2011. pp. 1-59. [Consulta: 22 noviembre 2018]. Disponible en: https://masteres.ugr.es/moea/pages /tfm1011/predicciondepicosdepolenmedianteregresionlogisticafuncional.

TARRÍO, J.; & NAYA, S. Influencia de la adición de nano y microsílice en la estabilidad térmica de una resina epoxi. Aplicaciones del ANOVA funcional. *Revista Colombiana de Estadística* [En línea], 2011, vol. 34, nº. 2, pp. 211-230. [Consulta: 31 octubre 2019]. ISSN 0120-1751. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0120175120110002000001&lng=en&nrm=iso&tlng=es.

Tipos de estaciones meteorológicas. [En línea]. [Consulta: 11 febrero 2019]. Disponible en: http://www.guiaspracticas.com/estaciones-meteorologicas/tipos-de-estaciones-meteorologicas.

TORECILLA NOGUERALES, L.J. *Análisis de datos funcionales, clasificación y selección de variables* [En línea]. (Trabajo de Titulación). (Maestría). Universidad Autónoma de Madrid. España. 2010. pp. 1-75. [Consulta: 21 noviembre 2018]. Disponible en: https:// repositorio.uam.es/handle/10486/12556.

URIBE OPAZO, M.A. *Modelos espaciales lineales gausianos en el estudio de la variabilidad espacial* [En línea]. (Trabajo de Titulación). (Pregrado). Universidad Nacional Mayor San Marco. Lima-Perú. 2015. [Consulta: 11 febrero 2020]. Disponible en: http://cybertesis.unmsm.edu.pe.

USECHE CASTRO, L.M.; & MESA ÁVILA, D.M. Una introducción a la Imputación de Valores Perdidos. *Terra. Nueva Etapa* [En línea], 2006, vol. 22, nº. 31, pp. 127-151. [Consulta: 11 febrero 2020]. ISSN 1012-7089. Disponible en: https://www.redalyc.org/articulo.oa?id=72103106

VAN BUUREN, S.; & GROOTHUIS-OUDSHOORN, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* [En línea], 2011, vol. 45, nº. 3, pp. 1-67. [Consulta: 20 noviembre 2017]. ISSN 1548-7660. Disponible en: https://www.jstatsoft.org/article/view/v045i03.