



Geoestadística Funcional aplicada a la predicción de la precipitación en sitios no muestreados en el Valle del Cauca para el año 2011

Laura Andrea Orozco Cerón
Carlos Alberto Taimal Yepes

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2017

Geoestadística Funcional aplicada a la predicción de la precipitación en sitios no muestreados en el Valle del Cauca para el año 2011

Laura Andrea Orozco Cerón
Carlos Alberto Taimal Yepes

Trabajo de grado presentada(o) como requisito parcial para optar al título de:
Profesional en Estadística

Director(a):
Mercedes Andrade Bejarano, Ph.D.

Codirector(a):
Johann Ospina, M.Sc.

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2017

A mis padres y a mis hermanos, por el inmenso apoyo que he recibido incluso en las situaciones más adversas y por ser mi mayor fuente de inspiración.

Laura Andrea Orozco Cerón.

A mis padres y a mi hermano por ser la fortaleza y aliciente para continuar día a día.

Carlos Taimal Yepes.

Agradecimientos

En primer lugar queremos agradecer a la escuela de estadística por ser baluarte en la formación del conocimiento estadístico en la región y por permitirnos conocer tan maravillosa disciplina y tan maravillosos docentes que sembraron en nosotros las ansias del conocimiento.

A la profesora Mercedes Andrade PhD. por su gran disposición, acompañamiento y por guiarnos en la elaboración de este proyecto e impulsarnos cada día ser mejores estadísticos. Al profesor Johann Ospina MsC. por encaminar este proyecto, darnos las pautas necesarias para desarrollarlo apropiadamente y por proveernos las bases de datos completas que se usaron en este trabajo. Al profesor Javier Olaya PhD. por ser un soporte en el área de datos funcionales cuando fue necesario. Al profesor Ramón Giraldo PhD. por haber sentado las bases de este proyecto junto al profesor Johann Ospina y la Profesora Mercedes Andrade y además por compartir sus aportes en el área de Geoestadística Funcional. También, agradecemos a la CVC por suministrar los datos de la precipitación diaria del Valle del Cauca y finalmente, gracias a todos nuestros compañeros de carrera por el apoyo incondicional, por compartir su conocimiento y su gran compañía en este proceso.

Resumen

En el presente estudio se aborda la problemática de la predicción espacial de la precipitación restringido al Valle del Cauca, puesto que conocer el comportamiento de variables que pueden impactar la agricultura de una región resulta de vital importancia para anticipar eventos que puedan afectar su normal conducta en cuanto a agricultura y a economía se refiere. Así pues, para lograr dichos pronósticos es acertada la técnica de la modelación realizada a partir de técnicas asociadas a la estadística espacial o geoestadística, de esta manera caracterizar la variable climática de interés y monitorear el cambio climático en zonas donde no se tiene registros de dicha variable.

De esta manera, este estudio abarcará áreas relacionadas al medio ambiente, más específicamente temáticas meteorológicas, como lo es el estudio de la precipitación en una zona de interés (Valle del Cauca) en el periodo 2011, con la finalidad de presentar información relevante a los interesados en el área de agricultura en la región mencionada y en el periodo especificado.

Por último, es importante mencionar que los datos proporcionados para el trabajo que se presentará a continuación fueron proporcionados por la Corporación autónoma regional del Valle del Cauca más conocida por sus siglas (CVC), adicionalmente, este conjunto de datos fue completado por Ospina (2016), mediante las técnicas de imputación univariadas enmarcados en la línea de la geoestadística espacial a través de la metodología kriging, lo cual facilitó la construcción de precipitaciones pentadales y las posteriores curvas funcionales. Para obtener el modelo que pronostique la precipitación en el Valle del Cauca es necesario indagar sobre la estructura de correlación de la variable de acuerdo a la zona geográfica (estaciones meteorológicas) y por último, evaluar la precisión de las predicciones.

Palabras clave: precipitación, datos funcionales, correlación espacial.

Contenido

Agradecimientos	7
Resumen	9
1. Introducción	2
2. Planteamiento del problema	4
3. Justificación	6
4. Antecedentes	8
4.1. Distribuciones de Probabilidad	8
4.2. Procesos Estocásticos y Cadenas de Markov	9
4.3. Análisis de Series Temporales	9
4.4. Modelos Lineales Generalizados	10
4.5. Geoestadística	10
4.6. Análisis de Datos Funcionales	11
4.7. Geoestadística Funcional	12
5. Objetivos	14
6. Marco Teórico	15
6.1. Marco teórico conceptual	15
6.1.1. Lluvia	15
6.1.2. Valle del Cauca	16
6.2. Marco teórico estadístico	17
6.2.1. Geoestadística	17
6.2.2. Estimación del Variograma	21
6.2.3. Modelos Teóricos de Semivarianza	21
6.2.4. Ajuste del Semivariograma Experimental al Modelo Teórico	22
6.2.5. Predicción Espacial: Kriging	23
6.2.6. Datos Funcionales	25

6.2.7. Introducción	25
6.2.8. ¿Qué es un Dato Funcional?	25
6.2.9. Espacios Funcionales	26
6.2.10. Medidas resumen y Análisis Exploratorio para datos Funcionales . . .	27
6.2.11. Representación de datos mediante funciones Base	28
6.2.12. B-Splines	29
6.2.13. Ajuste de Datos a un Sistema de Bases mediante Mínimos Cuadrados Ordinarios	30
6.2.14. Varianza de un Dato Funcional	31
6.2.15. Elección del número de Funciones Bases K	31
6.3. Geoestadística Funcional	33
6.3.1. Kriging Funcional Ordinario	35
7. Metodología	38
7.1. Datos	38
7.2. Construcción de los Datos de Precipitación Pentadal	40
7.3. Análisis Exploratorio	42
7.4. Análisis de Datos Funcionales	42
7.4.1. Selección del número de funciones base K y el parámetro de penalización por Rugosidad	42
7.4.2. Ajuste de las funciones Base mediante Mínimos Cuadrados	46
7.4.3. Análisis Exploratorio Funcional	49
7.5. Geoestadística Funcional	49
7.5.1. Determinación de la estructura de Correlación Espacial	49
7.5.2. Kriging funcional ordinario y predicción espacial	50
8. Resultados y Discusión	52
8.1. Análisis Exploratorio	53
8.1.1. Construcción de los Pentadales	53
8.2. Suavización de las Curvas	57
8.3. Análisis Estructural	58
8.4. Kriging Funcional Ordinario Predicción Espacial	60
8.4.1. Análisis de Residuales	66
9. Conclusiones y Recomendaciones	68
9.1. Conclusiones	68
9.2. Recomendaciones	69
Bibliografía	70

10. Anexos	76
10.1. Código en R para la suavización de los Datos	76
10.2. Código en R para el Análisis Estructural	78
10.3. Código en R para la Validación del Modelo	79

Lista de Tablas

6.1.	Modelos Teóricos acotados de Semivarianza	22
6.2.	Tipos de kriging lineal. Fuente: Giraldo (2002)	23
7.1.	Validación Cruzada Generalizada para el número óptimo de funciones Base .	44
7.2.	Validación cruzada para el valor óptimo del parámetro de penalización . . .	45
8.1.	Resumen, Precipitación Pentadal Mensual	56
8.2.	Modelos teóricos	59
8.3.	Estaciones escogidas para ilustrar	61

Lista de Figuras

6.1.	Componentes del Semivariograma	20
7.1.	Zona de Estudio. Fuente: IGAC (2012)	38
7.2.	Mapa de las estaciones Climáticas	39
7.3.	Estaciones según altura	40
7.4.	Precipitación diaria para la estación la Balastera	41
7.5.	Precipitación pentadal para la estación la Balastera	41
7.6.	Resultados de la Validación Cruzada Generalizada	46
7.7.	Bases B-splines	47
8.1.	Precipitación diaria 2011	53
8.2.	Lluvia pentadal 2011.	54
8.3.	Lluvia pentadal por Estación	55
8.4.	Precipitación Pentadal Mensual.	56
8.5.	Curvas Suavizadas para el año 2011	57
8.6.	Curvas Suavizadas 2011, tras aplicar la transformación	58
8.7.	Traza-variograma teórica ajustada	59
8.8.	Modelos de Traza-variograma teóricos ajustados	60
8.9.	Estaciones escogidas	61
8.10.	Pesos asociados al predictor OKFD	64
8.11.	Predicción de curvas pentadales para las 92 estaciones	65
8.12.	Residuales de curvas pentadales para las 92 estaciones	66
8.13.	Análisis de los residuales	67

Capítulo 1

Introducción

La cantidad de precipitación depende de diversos factores meteorológicos, como la temperatura, la humedad, la presión atmosférica, entre otros Rodríguez et al. (2004), por lo cual, esta variable presenta alta volatilidad en sus observaciones, de manera que, el estudio de esta variable cobra importancia al revelarse la naturaleza de esta, la cual depende de diversos factores presentes en el medio y que inciden directamente sobre esta. Así pues, la modelación del comportamiento de la variable precipitación, se ha convertido en centro de análisis en diversos estudios en los últimos años (Jaramillo & Chaves, 2000). Las investigaciones que se tienen acerca de este fenómeno se concentran en el mecanismo físico que da lugar al mismo, sin embargo, la gran variabilidad de este proceso de condensación del agua ha incentivado a estudios que permitan la predicción de esta, con el fin de deducir datos faltantes de escurrimiento y así como ampliar su registro.

Diariamente se recolectan registros de precipitación, lo que permite pensar que esta es una variable de grandes dimensiones a nivel de observaciones, esto permite o puede dar lugar a un análisis de datos funcionales, debido a que este se caracteriza por analizar variables que evolucionan a lo largo del tiempo, de modo que, los valores que toman en general son funciones (Delgado, 2012). Se tienen registros de la precipitación en el Valle del Cauca para el periodo 2011, datos con los cuales se pretende hacer uso de las técnicas enmarcadas en el área de geoestadística la cual utiliza información de la variable en diferentes locaciones espaciales para desarrollar modelos que indiquen la dependencia entre las medidas de las diversas estaciones meteorológicas y que faciliten realizar la futura predicción espacial en sitios donde esta variable no haya sido medida (Giraldo, 2002).

El presente documento esta sujeto a la siguiente estructura: En el capítulo 1 se presenta la introducción del trabajo de grado. En el capítulo 2 se discute el plantamiento del problema. En el capítulo 3 se presenta la justificación asociada al capítulo anterior para el presente trabajo. En el capítulo 4 se realiza la revisión de literatura que corresponde a los antecedentes usados para la modelación de la precipitación y las diversas metodologías para abordar la misma. En el capítulo 5 se presentan los objetivos planteados para el trabajo de grado.

En el capítulo 6 se introducen los conceptos necesarios para implementar la metodología del trabajo de grado y contextualizar al lector. En el capítulo 7 se ilustra la metodología utilizada para el respectivo análisis de los datos. En los capítulos 8 y 9 se presenta la discusión de los resultados obtenidos, las conclusiones a las que se llegaron a partir de estas, y algunas recomendaciones para estudios posteriores.

Capítulo 2

Planteamiento del problema

En el departamento del Valle del Cauca se observa una gran variación respecto al clima, estas variaciones obedecen a factores geográficos de la región como lo son la latitud, la altitud, los relieves montañosos, etc (Castro & Carvajal, 2013). Geográficamente el pacífico es una región que presenta una humedad relativa promedio mensual de alrededor de 86 % a 87 %, se tiene información de que esta zona presenta una época de lluvias al año, la cual se da en el segundo semestre del año; de esta zona se han reportado precipitaciones de hasta 5.000 *mm* anuales, mientras que en el área plana del Valle del Cauca se presentan dos períodos de lluvias al año, los cuales se identifican cerca de los meses de Marzo a Mayo y de Septiembre a Noviembre respectivamente; por otro lado, los meses de Diciembre a Febrero y los meses de Junio a Agosto, se registran niveles de precipitación más bajos o inferiores a los que se reportan durante el año, por lo que se pueden caracterizar como período seco; por último el área montañosa presenta épocas lluviosas más prolongadas durante el año que van de Marzo a Junio y de Septiembre a Diciembre. En las áreas frías y de páramo, las lluvias son ligeramente superiores a los 1.500 *mm*; por encima de los 3.500 *mm* disminuyen hasta alrededor de 1.000 *mm* (Escobar, 2006).

Dada la inestabilidad entre periodos de un mismo año de este fenómeno, los cultivos realizados usando las lluvias como fuente de riego, estarán más propensos a cosechas arriesgadas dependiendo de la fluctuación de precipitación. Así mismo, debido a estas alteraciones en la precipitación, es fácil suponer que se registrarán escenarios inusuales, es decir, se generarán valores atípicos, los cuales pueden influenciar altamente la media de precipitaciones de una región; como consecuencia de estas irregularidades, se puede clasificar erróneamente una zona de cultivos. Resultaría entonces beneficioso para el agricultor, conocer el comportamiento de este factor climatológico para que los cultivos no sufran variaciones en su rendimiento (Schlegel et al., 1996). Sumado a esto, un inminente cambio climático que se ha venido dando en las últimas décadas, está afectando la oferta hídrica asociada a la lluvia. Consecuentemente la agricultura se verá perjudicada, se alterarán las condiciones ambientales normales en algunas zonas de cultivo, desestabilizando el equilibrio climático,

lo que puede provocar cambios en los procesos de siembra (Bedoya et al., 2010). Por esto y recalcando principalmente el cambio climático a nivel mundial, se hace necesario el poder “predecir” el comportamiento de dicho fenómeno.

Para predecir la precipitación, existen diversos métodos de interpolación como lo son el Polígono de Thiessen, el inverso de la distancia al cuadrado y el método isoyetal (Collado, 1988); sin embargo, dada la importancia que se le asigna a las relaciones espacio-temporales entre distintos puntos de medición, la geoestadística ha sido de vital importancia para llevar a cabo predicciones en puntos en los cuales dichas mediciones no se han tomado (Montero et al., 2015). Los métodos geoestadísticos disponibles para poder realizar predicciones que permitan determinar los comportamientos de dicho fenómeno, son diversos y han mejorado en los últimos años; los métodos geoestadísticos clásicos de interpolación como el kriging ofrecen una buena aproximación tal y como lo han demostrado estudios anteriores (Goovaerts, 2000); sumado a esto, el uso de datos funcionales, ha tenido un impacto resaltante en lo que se refiere a la implementación de las técnicas de interpolación geoestadísticas, dado que algunas mediciones poseen una estructura tanto funcional como espacial; el precedente más inmediato se remonta a Giraldo (2009), con la adaptación del método Kriging Ordinario a Datos Funcionales (Montero et al., 2015); este caso en particular, supone una generalización de un ejemplo propuesto por (Ramsay & Silverman, 2005) en donde se presentan datos correspondientes a temperaturas en Canadá para 35 estaciones climáticas; aquí, los autores introducen las coordenadas correspondientes al conjunto de datos funcionales original y proponen predecir la precipitación de un sitio no muestreado, nótese que a diferencia del kriging ordinario, no se busca predecir un valor particular de una variable unidimensional, sino por el contrario, predecir una función como tal; estando así, este concepto sumamente ligado al de predicción espacial multivariada (Giraldo et al., 2011).

Este estudio pretende ajustar un modelo de geoestadística funcional que permita la predicción de la precipitación en algunas zonas del Valle del Cauca en las cuales no se han tomado mediciones de esta variable, a partir de los datos registrados en las estaciones meteorológicas, partiendo de datos tomados en 92 estaciones alrededor del departamento del Valle del Cauca para el año 2011.

Capítulo 3

Justificación

El Valle del Cauca es una de las regiones más ricas y de mayor desarrollo de la República de Colombia. Está ubicado en el suroccidente del país y tiene costas sobre el Océano Pacífico, en una extensión de 200 kilómetros, donde se encuentra ubicado el puerto de Buenaventura, el más importante del país, por el inmenso movimiento de carga de importación y exportación que se registra por esa zona (Alcaldía de Santiago de Cali, 2016) constituyendo el 1.9 % del territorio nacional (Secretaría Departamental de Salud del Valle del Cauca, 2016).

Uno de los sectores que más genera ingresos en el departamento, es el sector agrícola, siendo la agricultura vallecaucana la más importante y variada del país, con cultivos tecnificados de caña de azúcar, café, algodón, soya y sorgo (DANE, 2013). Sin embargo, el clima y su variabilidad inciden en muchos ámbitos, pero el sector agrario presenta una mayor vulnerabilidad ante el cambio climático. Se preveen aumentos de la frecuencia y gravedad de acontecimientos extremos como inundaciones, tormentas de granizo o sequías, todos ellos son riesgos permanentes que ya sufre la agricultura y ganadería (Piqueras, 2007).

Dado que desde el punto de vista económico es conveniente que dichos cultivos presenten un rendimiento óptimo y generen ganancias significativas, conocer la distribución de la precipitación a lo largo del año o poder predecir su coincidencia con los ciclos agrícolas, permite entonces a los expertos en el sector agrícola planificar adecuadamente las labores del campo o detectar con suficiente antelación acontecimientos que pueden alterar las condiciones favorables de un cultivo o una cosecha para la proliferación de determinada plaga, puesto que éstas se pueden dar en ciertas condiciones climáticas idóneas (SIPSA, 2014).

De esta manera, modelar la precipitación se ha convertido en una tarea de vital importancia y es a partir de ello que se encamina el presente estudio, el cual pretende generar información sobre el comportamiento de este fenómeno a partir de las observaciones registradas en las estaciones meteorológicas, de manera que se pueda explorar a partir de estos registros las variaciones en la distribución de la precipitación durante el año debido a los factores externos (Euscátegui & Hurtado, 2011).

Es importante mencionar, que el comportamiento de la precipitación exhibe espacios de tiempo donde hay ausencia de la variable, lo que genera presencia de ceros en las bases de datos y su periodicidad (Hasan & Dunn, 2010, Dunn, 2004, Suhaila et al., 2011, Suhaila & Yusop, 2016, Ramsay & Silverman, 2005), la modelación de esta variable constituye también una problemática muy interesante desde el punto de vista climatológico y aún más, desde el punto de vista estadístico.

Así y con el objetivo de ajustar un modelo para predecir la precipitación en sitios donde no se tengan mediciones, se propone el siguiente proyecto enmarcado en el contexto de datos funcionales, más precisamente, datos funcionales con presencia de correlación espacial. En esta investigación se hará uso de la metodología de geoestadística funcional propuesta por Giraldo (2009).

A partir del análisis geoestadístico funcional es posible hacer interpolación espacial de curvas, es decir, predecir curvas (en vez de variables) en sitios donde no hay información. Esta metodología tiene como ventaja la disponibilidad de una predicción continua para el instante de tiempo que se requiera (Giraldo et al., 2011). Además, se puede resaltar que en el análisis de datos funcionales se busca una mejor representación de los datos para un mayor análisis, de manera que, los datos funcionales resaltan las características de la variable de interés, permitiendo estudiar patrones y variaciones de los mismos, los cuales no se abarcan en el análisis clásico (Ramsay & Silverman, 2005).

Capítulo 4

Antecedentes

Desde un enfoque estadístico, la modelación de la precipitación ha sido abordada a través de un sinnúmero de técnicas. Se destacan aquellas como las distribuciones de probabilidad, las cadenas de Markov, modelos mixtos, series temporales, geoestadística y más recientemente, Análisis de Datos Funcionales y Geoestadística funcional. Cada una de estas metodologías se abordará en detalle en este capítulo y se pretende sentar las bases necesarias para la implementación de la propuesta metodológica, junto con el marco teórico. La estructura de este capítulo esta dado de acuerdo al orden presentado previamente; se abordarán algunas técnicas clásicas para la modelación de la precipitación basadas en distribuciones de probabilidad, posteriormente se analizará el caso de los Procesos Estocásticos, Cadenas de Markov y Análisis de Series temporales y en las últimas tres secciones se entrará en materia en lo que concierne al análisis espacial para variables regioanalizadas; primero se estudiarán métodos geoestadísticos relevantes para modelar la variable, luego se estudiará como ha sido abordado el problema desde el Análisis de Datos Funcionales y finalmente, la última sección corresponde a la temática de Geoestadística Funcional, en donde se integrarán ambos métodos.

4.1. Distribuciones de Probabilidad

En primer lugar, Agud (2011) en el marco de un proyecto hidrológico modeló para España la precipitación intensa diaria y subdiaria usando dos modelos probabilísticos basados en una modificación de la distribución Gumbell y la log-Normal respectivamente. Se sometieron los modelos obtenidos a pruebas de bondad de ajuste para 108 observatorios climáticos, donde se probaron los parámetros y se obtuvo que “los modelos de probabilidad se ajustan mejor para los climas lluviosos que para los secos”. Se presenta como principal novedad en palabras del autor, el ajuste a los modelos de probabilidad que permiten corregir el error sistemático de las series temporales, permitiendo además, “rellenar huecos” de series incompletas o corregir discontinuidades.

4.2. Procesos Estocásticos y Cadenas de Markov

Stern & Coe (1984), ajustaron modelos para la precipitación diaria para registros de lluvia en Morogoro (Tanzania). Se ajustaron cadenas de Markov no estacionarias para la ocurrencia de la lluvia, mientras, se ajustó una distribución Gamma con parámetros que variaran según el periodo del año, para las cantidades de lluvia. No obstante y pese a predecir correctamente la probabilidad de lluvia y la cantidad y ser una alternativa viable para los no estadísticos, este enfoque de modelación requiere un cumplimiento considerable de supuestos restringiendo su aplicabilidad a cualquier conjunto de datos; asimismo, estos modelos no pueden ser estimados si existe un uso o tratamiento ineficiente de los datos, lo cual repercute en el error estándar de estos. Dunn (2004) señala que dado que la lluvia es una variable continua que toma valores de cero, supone uno de los limitantes en la modelación de la misma. Para la probabilidad de ocurrencia, se puede recurrir a cadenas de Markov de dos estados para días con lluvia (“Lluvia” y “No lluvia”). Otras modelos incluyen cadenas de orden híbrido en el cuál las cadenas de primer orden se usan para modelar temporadas húmedas pero la dependencia de orden alto pueden usarse para modelar temporadas secas. Los periodos secos y húmedos pueden modelarse usando una distribución geométrica. En vez de Cadenas de Markov, pueden usarse mezclas de distribuciones geométricas o binomiales negativas; finalmente el autor, propone la modelación de la ocurrencia de la lluvia mediante una familia de distribuciones denominadas *Tweedie*, entre las cuales se destacan las distribuciones Poisson-Gamma, recalcando el uso de la distribución Gamma para la modelación de la cantidad de lluvia.

4.3. Análisis de Series Temporales

Guenni et al. (2008), analizaron igualmente la tendencia y la estacionalidad de la precipitación anual para años recientes en Venezuela; para tal fin, los autores ajustaron modelos lineales con componente de tendencia, estacionalidad y errores autorregresivos mediante mínimos cuadrados generalizados. Tras obtener los modelos ajustados, se usó kriging para interpolar los valores de la tendencia a nivel nacional; se obtuvo así, que la parte norte costera del país presenta una tendencia negativa en la precipitación, al igual que la mayor parte de la zona andina y la parte norte del estado Bolívar y gran parte de los estados Anzoátegui y Monagas; además para la mayoría de las localidades analizadas, la tendencia no resultó significativa. Se concluyó entonces, que mensualmente se detectó un régimen bimodal en 39 de las 113 estaciones analizadas, ubicadas mayormente en la región de la cordillera de Los Andes y en el estado Zulia. En el resto de las estaciones (74), el régimen estacional fue unimodal.

4.4. Modelos Lineales Generalizados

Dada la flexibilidad que ofrece este tipo de modelos, su uso supone una alternativa atractiva en cuanto se refiere a la integración de componentes tanto determinísticos como aleatorios, se tienen trabajos a nivel del Valle del Cauca, que es la región de interés para el presente, se podría resaltar el trabajo de Cruz & Males (2012) quienes ajustaron un modelo lineal mixto generalizado para la modelación de la precipitación mensual para 155 estaciones durante el periodo 1971-2009, integrando una estructura de correlación espacial tipo Matérn. Se tuvo que la distribución de la variable seguía una distribución Gamma y se identificaron dos componentes, uno temporal (fenómeno del Niño y la Niña) y uno espacial relacionado a la altitud de la estaciones climáticas. Además, se encontró significativa la presencia de la variable altura y el cuadrado de esta, en la modelación de la variable en cuestión.

Similarmente, Portillo & Medina (2012) modelaron la precipitación diaria mediante el mismo método para el periodo 1971-2009 obteniendo resultados similares. Se obtuvo que la variable en cuestión presentaba un comportamiento descrito por una distribución Gamma o Log-Normal y una relación significativa entre “la precipitación y la altitud y la influencia de los fenómenos del Niño y la Niña, además de los periodos secos y húmedos en los sitios de estudio”. Se incluyeron además factores fijos como el Índice de Oscilación Sur, la altitud y como factores aleatorios el Tiempo(año) y la localización de las estaciones.

4.5. Geoestadística

Goovaerts (1999, 2000) usó el cokriging para llevar a cabo interpolaciones de la precipitación en la región de Algarve (Portugal), tomando como variable auxiliar, la elevación de las estaciones de medición, demostrando además mediante un estudio comparativo de distintas variaciones del Kriging, técnicas no estadísticas y el cokriging, la ventaja de este último respecto a técnicas de interpolación determinísticas como el Polígono de Thiessen, las isoyetas y la distancia media cuadrada y aún más, sobre el kriging como tal. Un estudio similar llevado a cabo por Ly et al. (2011), comparó los metodos determinísticos ya mencionados con distintas variaciones de Kriging en una zona de Bélgica; se introdujo la covariable elevación al Kriging con Deriva Externa y el Cokriging Ordinario sin mejora alguna en la predicción de la lluvia diaria; Cokriging Ordinario y la distancia inversa al cuadrado resultaron ser los mejores métodos de interpolación proveyendo el menor Error Cuadrático Medio Residual para casi todos los casos. Jewell & Gaussiat (2015) realizaron un estudio en el cual se llevó a cabo la implementación de Kriging con Deriva Externa, Kriging con Corrección por Radares, Kriging Ordinario y Kriging Multicuadrático. Los autores mezclaron datos de radar y pluviómetros. Se llegó a la conclusión que el Kriging Con Deriva Externa (KDE), presentó un mejor comportamiento en contraste con otras técnicas, pues, se logró una mayor precisión en cuanto a los resultados climatológicos y las condiciones meteorológicas, con el uso de un

modelo variograma paramétrico para periodos cortos de tiempo (15 minutos) y un modelo no paramétrico para periodos acumulados de mayor duración.

Verdin et al. (2015, 2016) utilizaron datos tanto de redes pluviométricas como de satélite para la predicción de la precipitación en sitios no muestreados para centro América (Guatemala, Belice, El Salvador, Honduras, Nicaragua, Costa Rica, Panamá), Colombia y Venezuela, usando la metodología de Kriging. Los datos fueron recolectados por la Red Histórica Climatológica Global en compañía del Grupo de Riesgos Climáticos (Climate Hazards Group o CHG) de la Universidad de California en Santa Bárbara; los datos se agregaron cada cinco días (precipitación pentadal) por la CHG con el fin de producir datos que concordaran en escala temporal con los datos satelitales. En el primer caso, los autores utilizaron un acercamiento bayesiano modelando los datos de los pluviómetros como funciones lineales de aquellos correspondientes a los satélites y hallando las distribuciones a priori de los parámetros mediante muestreo Monte Carlo para Cadenas de Markov (MCMC por sus siglas en inglés); luego, a partir de las distribuciones a posteriori, se hallaron los parámetros necesarios para llevar a cabo el proceso de predicción espacial mediante kriging. Se asumió que las observaciones son realizaciones de un proceso Gaussiano, mientras los errores se asumieron como un proceso Gaussiano de media cero y función de correlación espacial exponencial. Los residuos de la función lineal se modelaron mediante kriging. Los modelos se aplicaron al periodo Enero-Junio del 2009, obteniéndose buenos ajustes, al considerarse un nugget o efecto pepita igual a cero, esto con base a los resultados obtenidos en el análisis exploratorio. Las estimaciones mediante el método combinado propuesto, mostraron que estas son más altas que las observaciones satelitales para Nicaragua y Colombia, lo cual es consecuente con la topografía de las zonas.

4.6. Análisis de Datos Funcionales

Para el Análisis de Datos Funcionales, quizá el antecedente más remoto es el de Ramsay & Silverman (2005) quienes para el caso de la precipitación promedio durante los últimos 40 años en Vancouver exponen una metodología alternativa de suavización para dichas observaciones, garantizando que las curvas obtenidas no tomen valores negativos, dada la presencia de valores iguales a cero. Cabe destacar que los autores no profundizaron considerablemente en el ejemplo de estos datos, aunque hicieron especial énfasis en el comportamiento de la variable a la hora de construir las curvas para el análisis. En este ejemplo fueron utilizadas las técnicas de suavización restringida teniendo en cuenta las bases de Fourier, entre las cuales destacan la suavización positiva y la suavización monótona, las cuales resuelven el problema de calcular curvas negativas.

Di Giuseppe (2013) mediante clústeres funcionales estudió la precipitación y la temperatura de 96 y 94 estaciones en Italia, respectivamente para un periodo de 30 años (1971-2000)

con el fin de “delinear zonas climáticas locales”; dado que se contaba con 5.36 % de datos faltantes, el autor estudió la imputación basada en promedios móviles y encontró que los clústeres funcionales son robustos en cuanto al método de imputación usado. Para reducir la variabilidad de la precipitación, utilizó la raíz cuadrada de acuerdo a la transformación de Box-Cox. El enfoque utilizado para la agrupación mediante clústers hace uso de los medoids en vez de las medias, ya que según expresa el autor, estos permiten una mejor identificación de los grupos y son más robustos a valores extremos y al ruido. Para el proceso de suavización se usaron bases B-Splines penalizadas, ya que según el autor, estas capturan mejor los comportamientos estacionales; el número de funciones base K se escogió mediante validación cruzada funcional; se ajustaron dichas bases de tal forma que hubiera un nodo de cuatro meses durante un periodo de 30 años. Para la precipitación, el total de nodos fue 90 y los polinomios ajustados fueron de orden cinco. Finalmente, se obtuvieron 9 clústeres los cuáles fueron consistentes con los patrones de la variable lluvia para la península italiana.

4.7. Geoestadística Funcional

Bel et al. (2011) ajustaron un modelo de regresión funcional con estructura de correlación espacial para modelar la diversidad genética de los árboles de Haya en Europa con la precipitación y temperatura como covariables.

El índice usado para medir la variabilidad fue el índice H el cual mide la probabilidad de que dos alelos escogidos al azar sean distintos. Los dos conjuntos de datos se escogieron al azar y sus localizaciones no coincidieron. Con el fin de lograr que los datos coincidieran espacialmente, se estimó una grilla regular de tiempo desde hace 15000 años hasta el presente, para las variables temperatura y precipitación. La interpolación se llevó a cabo mediante Kriging espacio-temporal asumiendo un modelo de correlación espacial exponencial separable. Tras ajustar el modelo asumiendo respuesta escalar, los residuales obtenidos presentaron correlación espacial; se ajustó un modelo de correlación exponencial a estos y se incorporan en el modelo y se reestimó nuevamente el modelo para comprobar su influencia en la estimación de los coeficientes de la temperatura y precipitación, así como su precipitación. Pese a que el modelo resumió adecuadamente una parte de la variabilidad, esta no fue suficiente para obtener un buen ajuste, ya que el R^2 fue de 0.31. Se encontró una mayor diversidad en el periodo de cambio climático, durante el Holoceno y menor cuando tanto la temperatura como la precipitación fueron considerablemente altas para todo el periodo de estudio. Finalmente, los autores propusieron mejorar el modelo mediante la inclusión pertinente de otras estructuras de correlación espacial y la correlación entre los predictores.

El referente más reciente y el cual servirá de base para la realización de este trabajo corresponde a Ospina (2016). En este, se ajusta un modelo de regresión funcional concurrente para la precipitación pentadal acumulada en el periodo 1993-2011; en la primera fase se ajustó el modelo lineal tomando como covariables la altura, el cuadrado de esta y la precipitación

pentadal acumulada satelital. La altura no supuso una mejora significativa en el modelo por lo cual se despreció a la hora de reajustar el modelo en la tercera fase; en la segunda fase, se modeló la estructura de correlación espacial de los residuos mediante geoestadística funcional; dicha estructura varió conforme al año de estudio y se introdujo en la tercera fase del modelo; en esta, se obtuvo como principal resultado una mejora en la explicación de la precipitación pentadal medida en estaciones al integrar dicha estructura, además, se presentó un mejor ajuste en periodos con episodios de fenómeno de la Niña. La gran mayoría de los residuales estudentizados presentaron valores cercanos a cero, no obstante, se presentó un comportamiento creciente para los periodos 1993, 1995, 1996 y 1998 el cual pudo causarse por “el proceso acumulativo de los datos” (Ospina, 2016). Se destacó además, una mejora significativa al modelo propuesto por Bel et al. (2011) tras incluir errores correlaciones espacialmente y al modelo propuesto por Verdin et al. (2015) en el cual no se consideró la variación de la variable en el tiempo.

Finalmente, en cuanto a otro tipo de variables se refiere, la modelación mediante geoestadística funcional ha presentado un gran auge, destacándose el trabajo de Giraldo et al. (2011, 2012) y Reyes et al. (2012) entre otros; se resalta el Kriging Funcional Ordinario propuesto por el Giraldo en el 2009.

Capítulo 5

Objetivos

Objetivo general

- Predecir mediante geoestadística funcional los valores correspondientes a la precipitación en el Valle del Cauca en sitios donde no se han tomado mediciones, a partir de registros recolectados en las estaciones meteorológicas para el año 2011.

Objetivos específicos

- Modelar la precipitación en el Valle del Cauca mediante Geoestadística Funcional.
- Realizar predicciones en sitios donde no exista información.

Capítulo 6

Marco Teórico

Este capítulo ofrece una visión del problema de interés, pues la base de datos con la cual se trabajará a lo largo de esta investigación corresponde a datos de precipitación, por lo cual se hace necesario realizar breves definiciones concernientes a la variable centro de este estudio; en concreto se dará una aproximación a las definiciones climatológicas de interés, correspondientes a la precipitación; adicionalmente, se exponen los conceptos cuya comprensión será de vital importancia para el entendimiento de la metodología planteada de manera subsecuente. La división correspondiente a estos conceptos se encuentra conformada por los siguientes temas: *Geoestadística*, *Análisis de Datos Funcionales* y *Geoestadística Funcional*.

6.1. Marco teórico conceptual

6.1.1. Lluvia

La precipitación puede ser descrita de acuerdo a la definición de la Organización Meteorológica Mundial (OMM) como la “la precipitación de gotas de agua que caen desde una nube” (Secretaría de la OMM, 1993). De manera que se pueden establecer diferentes tipos de lluvia dependiendo del diámetro y concentración de la misma, la cual varía conforme a la intensidad de la precipitación y su naturaleza (lluvia continua, chaparrón de lluvia, tempestad de lluvia, etc), además también es importante tener en cuenta si las nubes contienen un número anormal de finas partículas, por ejemplo, polvo o arena levantadas desde el suelo durante las tormentas que pueden ser transportadas al suelo con las gotas de lluvia (lluvia sucia) a menudo después de haber sido transportadas grandes distancias (Secretaría de la OMM, 1993).

Ahora bien, después de definida la precipitación es importante resaltar las unidades de medición de esta, la variable precipitación se encuentra medida en milímetros (mm), donde un milímetro es equivalente a L/m^2 , ya que un litro repartido en una zona de $1\ m^2$ origina

una lámina de agua de 1 *mm* (Sánchez, 2017). La intensidad de la lluvia se refiere a un instante, esto es, *Lluvia/tiempo*. Generalmente, esta suele ser *mm/hora*, en este estudio se considerará *mm/día* (Sánchez, 2017).

la lluvia suele medirse mediante el uso de distintos tipos de recipientes de paredes planas. A continuación, se indican los más importantes: **1) pluviómetro**, esta herramienta recoge el agua en un recipiente el cual evita que el líquido se escurra, se filtre o se evapore (Sánchez, 2017); la lectura se efectúa una vez al día y la precisión de la lectura del agua recogida es de $\pm 1\text{mm}$. Por otro lado la ubicación de este debe ser alejado de árboles o construcciones elevadas; en general a más del doble de la altura del obstáculo (Sánchez, 2017).

2) pluviógrafo: es un instrumento que mide la precipitación constantemente, este registra la precipitación en el tiempo, ya sea mediante tinta y papel o de manera digital. El gráfico obtenido en función del tiempo se denomina *pluviograma* y refleja la precipitación acumulada (Sánchez, 2017).

Por otra parte, se encuentran las **3) redes pluviométricas**: la cual representa una red de pluviómetros estos son diseñados teniendo en cuenta el relieve, la densidad de la población, el interés para obras hidráulicas, previsión de avenidas, etc. (Sánchez, 2017). También están las **Radiosondas** las cuales se usan con el fin de conocer el comportamiento climático en la tropósfera y estratósfera (primera y segunda capas más bajas de la atmósfera, con un grosor de 8 *km* sobre los polos y 6 *km* sobre el Ecuador, para la primera; la segunda se encuentra a 12 *km* sobre la atmósfera en el Ecuador y 50 *km* por encima de la atmósfera en los polos (Weather & Climate, 2017)).

Por otro lado, se encuentra los **Radares** los cuales se usan para monitorear la precipitación. Estos envían señales de microondas en un rayo angosto desde un transmisor en un periodo de tiempo muy corto (aproximadamente, una millonésima de segundo). Con base al tiempo en que la onda es recibida y enviada, la velocidad de la luz y el ángulo de la antena, los radares pueden encontrar las localizaciones de la lluvia en el espacio (Yu, 2011).

Por último, los **Satélites**; en general, los satélites miden la radiación proveniente de un número infinito de longitudes de onda, las cuales se dividen en unas pocas bandas. Con base a las radiaciones se capturan imágenes cuyos colores determinan un comportamiento específico de alguna variable meteorológica en concreto (Yu, 2011).

6.1.2. Valle del Cauca

El Valle del Cauca es un departamento colombiano ubicado en el suroccidente del país, en el centro de la región pacífica entre el flanco occidental de la cordillera central y el océano pacífico, ubicado geográficamente en latitud $4^{\circ} 35' 56.57''$ y longitud $74^{\circ} 4' 51.30''$. La extensión geográfica del departamento del Valle del Cauca es de 21.195 *km*² que representa el 1.9 % del territorio nacional (Secretaría Departamental de Salud del Valle del Cauca, 2016).

La temperatura promedio del departamento es de 23.5°C y presenta una humedad relativa que oscila entre el 65 % y el 75 %; al ser una región intertropical, presenta épocas secas de diciembre a febrero y de junio a septiembre, mientras las épocas lluviosas, se presentan de marzo a mayo y de octubre a noviembre; se encuentra conformado por 42 municipios y su capital es Santiago de Cali; Por otro lado, los municipios aledaños al río Cauca, presentan riesgo de inundaciones, los cercanos presentan riesgo de avenidas torrenciales y todos los municipios del departamento presentan riesgo de tormentas eléctricas y vendavales (Secretaría Departamental de Salud del Valle del Cauca, 2016).

Por otro lado, la hidrografía del Valle del Cauca se encuentra compuesta por muchas corrientes que se forman en dos vertientes, la del pacífico y la del río Magdalena que surten de agua al río Cauca (Valle del Cauca: Colombia, guía Turística, 2017). Entre los principales cursos de agua se encuentran los ríos (Secretaría Departamental de Salud del Valle del Cauca, 2016): Cauca, San Juan, Calima, Bugalagrande, La Soledad, Rasposo, Anchicayá, Dagua, Agua Sucia, La Vieja, Naya, Yurumanguí, Guapi, Tuluá, Rosario, Cofre (proviene del anterior), las Vueltas, Fraile, Jamundí, la Quebrada Aguaclara y la Quebrada los Alpes. Entre otros cuerpos de agua y accidentes costeros, se encuentran: el Lago Calima, conformado por la represa homónima; Laguna El Sonso, Bahía Málaga y la Bahía de Buenaventura (Secretaría Departamental de Salud del Valle del Cauca, 2016).

6.2. Marco teórico estadístico

6.2.1. Geoestadística

La geoestadística es el estudio de fenómenos regionalizados (Matheron, 1969), es decir, aquellos que se desplazan en el espacio y presentan cierta estructura, la cuál es de especial interés para poder inferir aspectos que no han sido medidos u observados (Bivand et al., 2008); esto se logra mediante la adaptación de técnicas propias de la regresión clásica que emplean información de la continuidad espacial (Isaaks & Srivastava, 1989).

Cuando el objetivo es el de realizar predicciones, la geoestadística, opera en dos etapas; estas son: el *análisis estructural* y la predicción en sitios no muestrados mediante la técnica del *kriging* (Giraldo, 2002). A continuación, se precisan otras ideas básicas como lo son las variables regionalizadas y se describen en detalle las etapas mencionadas anteriormente.

Variable regionalizada

Es una variable aleatoria Z definida en un punto s que se distribuye espacialmente y exhibe una estructura de correlación; formalmente, se puede definir como un procesos aleatorio o

estocástico contenido en un espacio euclidiano d -dimensional (Giraldo, 2002); esto es:

$$\{Z(s) : s \in D\} \quad (6.1)$$

Donde s hace referencia a la ubicación en la cuál se tomó la medición y D es fijo y se encuentra en el dominio $\mathbb{R}^d$ (d es comúnmente 2) (Cressie, 1993).

Momentos de una variable regionalizada

A continuación se definirán los momentos de la variable regionalizada (Giraldo, 2002)

- a) **Media.** $E[Z(s_i)] = \mu(s_i)$
- b) **Varianza.** $V[Z(s_i)] = E[(Z(s_i) - \mu(s_i))^2] = \sigma_i^2$
- c) **Autocovarianza.** $Cov(Z(s_i), Z(s_j)) = E[(Z(s_i) - \mu(s_i))(Z(s_j) - \mu(s_j))]$
- d) **Función de semivarianza o semivariograma.** $\gamma(s_i) = \frac{1}{2}Var[Z(s_i) - Z(s_j)]$

Estacionariedad

Se entiende por estacionariedad, la propiedad por la cual, la distribución de un proceso aleatorio tiene ciertos atributos que son los mismos en cualquier sitio (Webster & Oliver, 2007). Formalmente, esta condición consiste en que la función conjunta de la variable regionalizada no presente variación dada una traslación.

Definición 6.2.1. Se dice que un proceso $\{Z(s)\}$ es estacionario en segundo orden o estacionario en covarianza si posee segundos momentos finitos (Parzen, 1999), esto es,

- La función de media existe y no depende de la localización $\mu(s_i) = \mu \forall s_i \in D, i = 1, 2, \dots, n$
- La función de covarianza existe y sólo depende de las diferencias entre las localizaciones involucradas, esto es, $C(s_i, s_j) = C(h) \forall s_i, s_j, i \neq j$ donde $h = s_i - s_j$ es el vector de distancia entre las localizaciones (Melo, 2012).

Cabe resaltar que la estacionariedad considerada en series temporales toma en cuenta una sola dirección, en contraste, en el campo espacial, se debe suponer que la función es estacionaria en todas las direcciones (Giraldo, 2002).

Definición 6.2.2. Se dice que un proceso es intrínsecamente estacionario si (Melo, 2012)

- “La función de media existe y no depende de la localización”, $\mu(s_i) = \mu, \forall s_i \in D, i = 1, 2, \dots, n$
- La varianza de la diferencia del proceso en dos localizaciones distintas depende únicamente de la distancia entre estas. Esto es $Var[Z(s_i) - Z(s_j)] = 2\gamma(h) \forall s_i, s_j \in D$ donde $h = s_i - s_j$

La función $2\gamma(\cdot)$ recibe el nombre de *variograma* mientras, $\gamma(\cdot)$ es el *semivariograma*.

Definición 6.2.3. Un proceso se dice isotrópico si la dependencia entre $Z(s)$ y $Z(s + h)$ depende solamente de la distancia entre estos (Cressie (1993), Melo (2012)).

Correlación espacial

Con el fin de determinar alguna dependencia espacial entre las observaciones, se debe llevar a cabo un estudio de la correlación espacial entre estas; esta fase es comúnmente conocida como *análisis estructural*. Para llevar a cabo dicho análisis, se recurren a tres funciones que utilizan la información muestral, las cuales son: el *semivariograma*, el *covariograma* y el *correlograma* (Giraldo, 2002).

- **Covariograma y correlograma:** Se define la función de *covariograma* muestral para parejas de observaciones s_i, s_j a una distancia h de manera similar a la *covarianza* muestral, de manera que la estructura del covariograma esta dado por:

$$c(h) = Cov[Z(s + h), z(x)] = \frac{\sum_{i=1}^n (Z(s + h)Z(s))}{n} - \mu^2 \quad (6.2)$$

Ahora, para un proceso $Z(s)$ estacionario de segundo orden y con función de covarianza $C(h)$, se tiene que $C(0) = Cov[Z(s), Z(s)] = Var[Z(s)]$, por lo cual $C(0) > 0$. Se define así, el correlograma o función de autocorrelación (Melo, 2012):

$$\rho(h) = \frac{C(h)}{C(0)} \quad (6.3)$$

Se tiene además, que $\rho(h) = \rho(-h)$ y $\rho(0) = 1$.

- **Variograma y semivariograma:** El variograma es la función crucial de la geoestadística (Cressie, 1993). A diferencia de las otras funciones mencionadas previamente, su estimación no requiere del conocimiento de otros parámetros adicionales. Bajo el supuesto de estacionariedad intrínseca del proceso, los incrementos $Z(s + h) - Z(s)$ son estacionarios, con media cero y su varianza depende solamente de

la distancia entre los sitios de medición de estos, h (Melo (2012), Giraldo (2002)); se define de esta forma el variograma teórico como:

$$2\gamma(h) = \text{Var}[Z(s_i) - Z(s_j)] \quad (6.4)$$

donde $s_i, s_j \in D$ y $h = s_i - s_j$, es el vector de distancia entre las localizaciones (Melo, 2012). Se puede observar de (6.4) que $2\gamma(h) = 2\gamma(-h)$ y $2\gamma(0) = 0$.

La manera de interpretar esta función recae sobre la distancia mínima, pues a menor distancia entre las locaciones mayor similitud o correlación espacial habrá entre las observaciones (Giraldo, 2002)

Componentes del Semivariograma

Se definen los componentes del modelo teórico de semivarianza para contextualizar definiciones posteriores (Giraldo, 2002).

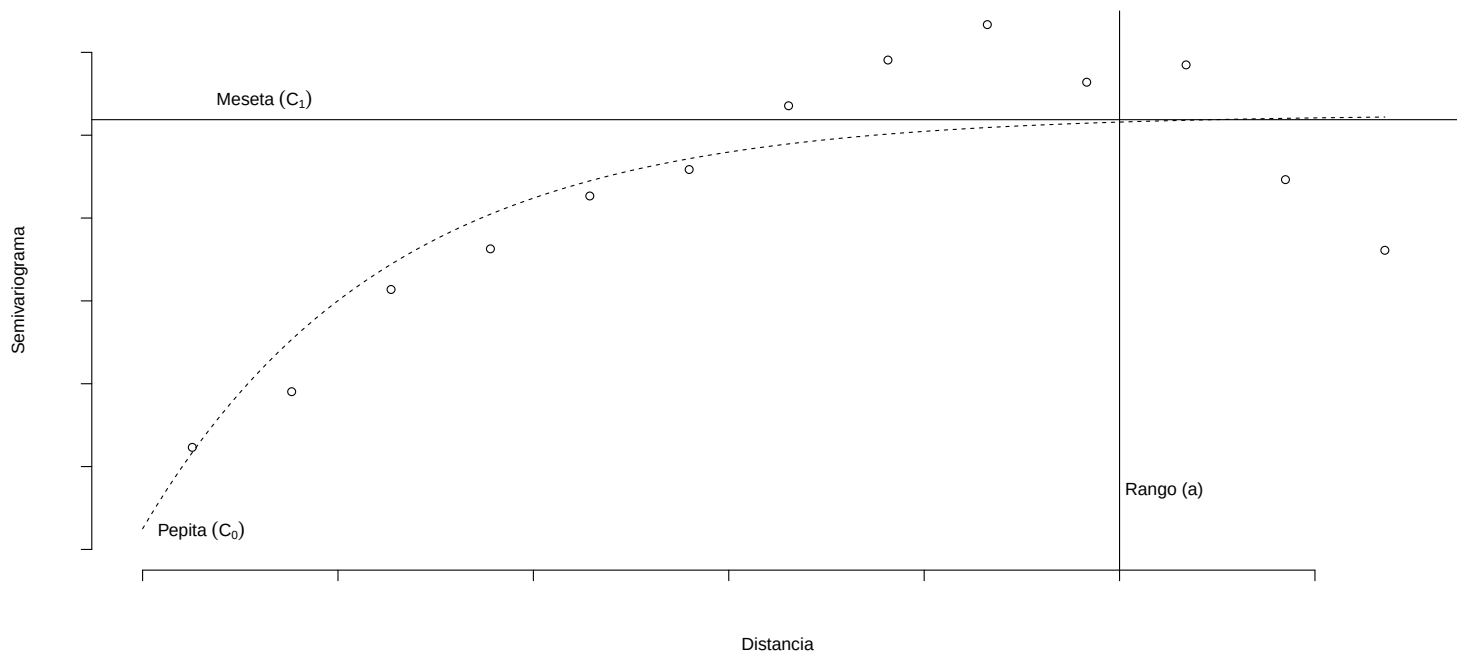


Figura 6.1: Componentes del Semivariograma

A partir de la Figura 6.1 es posible visualizar la **Pepita o Nugget**: la cual se denota por C_0 y representa la discontinuidad espacial del semivariograma respecto al origen. Ahora bien, la **Meseta o Sill**: es la cota superior del semivariograma; en otras palabras,

es el límite del semivariograma cuando h tiende a infinito. Es de resaltar que cuando el semivariograma cuenta con una meseta finita, el proceso es fuertemente estacionario de lo contrario es intrínsecamente estacionario. Se denota por C_1 . Finalmente, el **Rango**: Corresponde a la distancia en la cual dos observaciones son independientes entre sí. Entre más pequeño sea el rango, más cerca se encuentra del modelo de independencia espacial. Se denota por a .

6.2.2. Estimación del Variograma

A continuación se presenta la estimación del semivariograma teórico. Se estudiará el semivariograma clásico propuesto por Matheron (1969). La importancia del semivariograma empírico radica en la necesidad de hallar una estructura de correlación espacial teórica que mejor se ajuste, cuyos parámetros se utilizarán para llevar a cabo el proceso de predicción.

Estimador Clásico

Bajo el supuesto de media constante, el estimador clásico obtenido por el método de los momentos propuesto por Matheron (1969) está dado por (Cressie, 1993):

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(s_i) - Z(s_j))^2, \mathbf{h} \in \mathbb{R}^d \quad (6.5)$$

Donde el promedio en (6.5) se toma sobre $N(\mathbf{h}) = \{(s_i, s_j) : s_i - s_j = \mathbf{h}\}$ y $|N(\mathbf{h})|$ es el número de distintos elementos en $N(\mathbf{h})$. En el caso de datos que no son regularmente espaciados, $N(\mathbf{h})$ se modifica a $\{(s_i, s_j) : s_i - s_j \in T(\mathbf{h})\}$, donde $T(\mathbf{h})$ es una región de tolerancia en $\mathbb{R}^d$ alrededor de $\mathbf{h}$ (Cressie, 1993). Cressie (1993) detalla el *estimador robusto*, el cual es una modificación del anterior que no es susceptible a valores extremos.

6.2.3. Modelos Teóricos de Semivarianza

Los modelos isotrópicos de semivarianza pueden clasificarse en *no acotados* (lineal, logarítmico y potencial) y *acotados* (Esféricos, Gaussianos, Matérn y exponencial); los últimos aseguran que la covarianza del proceso es finita, por lo cual, en la práctica estos son ampliamente utilizados Giraldo et al. (2011).

Con el semivariograma experimental hallado, se procede a ajustar el modelo teórico que mejor describa la forma del proceso espacial (Giraldo, 2002).

Tabla 6.1: Modelos Teóricos acotados de Semivarianza

Modelo	Expresión
Esférico	$\gamma(h) = \begin{cases} C_0 + C_1 \left(\frac{3}{2} \left(\frac{h}{a} \right) - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right) & \text{si } h \leq a, \\ C_0 + C_1, & \text{si } h > a \end{cases}$
Exponencial	$\gamma(h) = C_0 + C_1 \left(1 - \exp\left(\frac{-3h}{a}\right) \right)$
Gaussiano	$\gamma(h) = C_0 + C_1 \left(1 - \exp\left(\frac{-h}{a^2}\right) \right)$
Matérn	$\gamma(h) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{h}{a} \right)^{\kappa} K_{\kappa} \left(\frac{h}{a} \right)$

Modelo Esférico

Este modelo es válido sólo en $\mathbb{R}^2$ y $\mathbb{R}^3$. Presenta un comportamiento aproximadamente lineal en el origen y alcanza su sill o meseta en $h = a$ (Chilès & Delfiner, 2012).

Modelo Exponencial

Es más simple en forma que el modelo esférico y es válido para cualquier d (Serban (2008), Chilès & Delfiner (2012)). No obstante, este modelo no posee rango finito (Serban, 2008).

Modelo Gaussiano

También válido para todo d . Alcanza su rango al rededor de $1.73 a$ (Chilès & Delfiner, 2012).

Modelo Matérn

Es una función positiva e integrable en $\mathbb{R}^d$ para todo d . Este modelo varía de acuerdo al parámetro κ y cuando $\kappa = 0.5$, coincide con el modelo exponencial (Giraldo, 2002), mientras, el modelo Gaussiano es un caso extremo de la función Matérn cuando $\kappa \rightarrow \infty$. La función K_{κ} es una función de Bessel modificada de segunda clase de orden κ (Chilès & Delfiner, 2012).

6.2.4. Ajuste del Semivariograma Experimental al Modelo Teórico

El siguiente paso en el análisis estructural, es ajustar el modelo teórico al semivariograma experimental. Esto se logra mediante alguno de los métodos de estimación ya conocidos como Mínimos Cuadrados Ordinarios (*OLS* por sus siglas en inglés), Mínimos Cuadrados Ponderados (*WLS* por sus siglas en inglés), Mínimos Cuadrados Generalizados (*GLS* por sus siglas en inglés) o Máxima Verosimilitud Giraldo (2002). Para este trabajo, se llevará a cabo el ajuste mediante WLS.

- **Mínimos Cuadrados Ponderados (WLS):** en este caso se escoge θ tal que se minimice

$$(\hat{\gamma} - \gamma(\theta))^T W(\theta) (\hat{\gamma} - \gamma(\theta)) \quad (6.6)$$

Donde $W(\theta)$ es una matriz diagonal cuyas entradas son las varianzas de los elementos de $\hat{\gamma}(\cdot)$.

Los estimadores (6.6) tienden a ser más eficientes que aquellos hallados por OLS, no obstante, estos requieren de la especificación de las matrices $V(\theta)$ y $W(\theta)$ respectivamente, limitando la implementación de los mismos (Serban, 2008).

Criterios para la Selección del Mejor Modelo

Para determinar la eficiencia entre distintos modelos, se hace uso de diversos criterios; dicha eficiencia se mide en términos de la Suma de Cuadrados de los Errores. Algunos de los criterios más importantes para la selección de modelos son (Solera, 2000): Coeficiente de determinación (R^2), Coeficiente de Determinación Ajustado (R^2 ajustado), Suma de los cuadrados de los errores (SCE), Criterio de información Akaike (AIC) y Criterio de Información Bayesiana (BIC).

6.2.5. Predicción Espacial: Kriging

Este término encierra diversas técnicas de predicción espacial, donde su principal objetivo es la reducción del error cuadrado de la predicción. Además, esta técnica tiene más aplicaciones diferentes a la predicción, como lo es la simulación y el diseño de redes óptimas de muestreo (Giraldo, 2002).

Existen diversos tipos de kriging, de los cuales se destacan los siguientes (Tabla 6.2)

Tabla 6.2: Tipos de kriging lineal. Fuente: Giraldo (2002)

Nombre	Propiedades
Simple	Son óptimos si hay normalidad multivariada
Ordinario	independiente de la distribución
Universal	predictores linealmente insesgados

Para este trabajo solo se discutirá a fondo el kriging ordinario. Los otros tipos de kriging lineales pueden consultarse en los trabajos de Cressie (1993), Chilès & Delfiner (2012), Webster & Oliver (2007) y Giraldo (2002).

Kriging Ordinario

Se está interesado en estudiar una variable Z , en los puntos s_i , con $i = 1, 2, \dots, n$, de la región de estudio, y se desea predecir $Z(s_0)$, es decir, la variable de interés en un lugar donde no se realizó medición. Para llevar a cabo esto, esta técnica propone predecir el valor de la variable como una combinación lineal de las n variables aleatorias observadas; así:

$$\hat{Z}(s_0) = \sum_{i=1}^n \lambda_i Z(s_i) \quad (6.7)$$

donde $\lambda_i, i = 1, 2, \dots, n$ es la ponderación o pesos de las variables observadas y $\sum_{i=1}^n \lambda_i = 1$ garantiza que la predicción sea insesgada y que la esperanza de la predicción será igual a la esperanza de la variable; esto es comúnmente conocido como el requisito de insesgamiento (Giraldo, 2002). Así pues, el interés, es hallar los pesos que minimicen la suma de cuadrados de la predicción y que garanticen la optimalidad de la misma.

Se asume que el proceso subyacente tiene la siguiente forma:

$$Z(s) = \mu + \delta(s), s \in D, \mu \in \mathbb{R}, \text{ y } \mu \text{ desconocido} \quad (6.8)$$

Se dice que el predictor $\hat{Z}(s_0)$ es el mejor estimador lineal, dado a que los pesos se han obtenido de manera que minimizan la varianza del error de la predicción, así pues, se puede resaltar que el método kriging en comparación con otros métodos de interpolación garantizan mínima varianza en la predicción.

En concreto, se busca la minimización de

$$E \left[(Z(s_0) - \sum_{i=1}^n \lambda_i Z(s_i))^2 \right] - 2m \left(\sum_{i=1}^n \lambda_i - 1 \right) \quad (6.9)$$

Con respecto a $\lambda_1, \lambda_2, \dots, \lambda_n$ y m que es el multiplicador de Lagrange, el cual asegura que $\sum_{i=1}^n \lambda_i = 1$ (Cressie, 1993).

Tras realizar el debido proceso de optimización y aplicar valor esperado, se obtiene el siguiente sistema de ecuaciones de manera matricial:

$$\begin{bmatrix} C_{11} & \dots & C_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{n1} & \dots & C_{nn} & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -m \end{bmatrix} = \begin{bmatrix} \gamma(s_0 - s_1) \\ \vdots \\ \gamma(s_0 - s_n) \\ 1 \end{bmatrix}$$

Esto es, los pesos que minimizan el error de la predicción se calculan a partir del covariograma, $C_{ij} = \gamma(s_i - s_j)$, de manera que, $\lambda = [C^{-1}]_{ij}C_{i0}$. Ya encontrados los pesos, se procede a calcular el valor de la predicción para el punto s_0 y así sucesivamente para cualquier punto en el que se quiera predecir.

Finalmente, la varianza del estimador kriging esta dado por:

$$\sigma_k^2 = \sigma^2 + \sum_{i=1} \lambda_i C_{i0} - \mu$$

6.2.6. Datos Funcionales

6.2.7. Introducción

El Análisis de Datos Funcionales o FDA (por sus siglas en inglés) constituye una de las ramas más modernas de la estadística cuya expansión y campos de aplicaciones se han diversificado en los últimos años (Delicado et al., 2010).

Desde el trabajo pionero de Deville (1974), Cuevas (2014) destaca como claves los trabajos de Ramsay & Silverman (2005), Ramsay & Silverman (2002) y Ferraty & Vieu (2006), donde los últimos, pertenecen a la llamada “escuela francesa” de datos funcionales asociada al grupo de investigación STAPH (STAPH, 2017), cuya contribución más notable al tema es tal vez, el trabajo pionero en Análisis de Componentes Principales para Datos Funcionales propuesto por Dauxois et al. (1982); más recientemente, se destacan los trabajos de Horváth & Kokoszka (2012) donde se aborda extensivamente el proceso inferencial para datos funcionales, Zhang (2013) el cual presenta formalmente las ideas concernientes al Análisis de Varianza Funcional, Hsing & Eubank (2015) quien ofrece una extensiva revisión teórica de la teoría matemática subyacente en el FDA y Kokoszka & Reimherr (2017) cuyo trabajo presenta de manera introductoria el FDA así como todos los posibles campos de aplicación a los cuales se ha extendido como lo son el Análisis de Regresión Funcional, los Modelos Lineales Generalizados para Datos Funcionales, el Análisis de Series de Tiempo Funcional y el Análisis de Datos Espaciales Funcional.

6.2.8. ¿Qué es un Dato Funcional?

Levitin et al. (2007), definen a un dato funcional como “un conjunto de mediciones a lo largo de un intervalo continuo (generalmente el tiempo) que en conjunto deben considerarse como una entidad, curva o imagen”; este tipo de datos, es especialmente útil cuando se trata de un conjunto de observaciones muy “denso”.

Aún más, dado que en la práctica los datos son recolectados por dispositivos que toman muestras de una determinada variable aleatoria en distintos momentos de tiempo en un rango específico (t_{min}, t_{max}) una observación puede expresarse por la familia aleatoria $\{X(t_j)\}_{j=1, \dots, J}$. Al ser más cercanas las observaciones, los datos pueden considerarse como

una sola observación de la familia continua $\chi = \{\chi(t) : t \in (t_{min}, t_{max})\}$ (Ferraty & Vieu, 2006). Se define así, el concepto de variable funcional.

Definición 6.2.4. Variable Funcional: Una variable $\chi(t)$ se denomina *variable funcional* (v.f. o f.v., por sus siglas en inglés) si toma valores en un espacio funcional. Una observación χ de $\chi(t)$ se denomina *dato funcional*.

6.2.9. Espacios Funcionales

Definición 6.2.5. Espacio Funcional Un espacio funcional es un espacio compuesto por funciones donde cada una de estas puede pensarse como un punto en dicho espacio (Tomaso Poggio & Frogner, 2010). Algunos espacios funcionales son:

- I) $C[a, b]$, el conjunto de todas las funciones continuas que toman valores en los reales en el intervalo $[a, b]$
- II) $L_1[a, b]$, el conjunto de todas las funciones que toman valores en los reales cuyo valor absoluto es integrable en el intervalo $[a, b]$
- III) $L_2[a, b]$, el conjunto de todas las funciones cuadrado-integrables que toman valores en los reales en el intervalo $[a, b]$

Definición 6.2.6. Espacios L_2

Uno de los espacios más importantes en el FDA, es el espacio L_2 o espacio de funciones cuadrado integrables. En general, para un intervalo $I \in \mathbb{R}$, el espacio vectorial $L_2(I)$ se define como (Navarro (2004), Belton (2006)):

$$L_2(I) := \{f : I \in \mathbb{R} \mapsto \mathbb{R} : \int_{\mathbb{R}} f^2 dt < \infty\} \quad (6.10)$$

Definición 6.2.7. De igual manera, para f y $g \in L_2$, se define el producto escalar euclidiano como (Navarro, 2004):

$$\langle f; g \rangle := f'g = \int f(t) \cdot g(t) dt \quad (6.11)$$

Cumpliendo todas las propiedades correspondientes a un espacio vectorial.

Definición 6.2.8. La norma- L_2 de $f(t)$ se define como:

$$\|f\| = \sqrt{\langle f; f \rangle} = \sqrt{\|f\|^2} = \sqrt{\int f(t)f(t)dt} \quad (6.12)$$

Definición 6.2.9. Las funciones $f(t)$ y $g(t)$ en L_2 son ortogonales si:

$$\langle f; g \rangle = \int f(t)g(t)dt = 0$$

6.2.10. Medidas resumen y Análisis Exploratorio para datos Funcionales

Para dos funciones $f(t), g(t) \in L_2$ y suponiendo $t \in [0, T]$, se definen las siguientes medidas poblacionales (Navarro, 2004):

- **Media funcional:** $\mu_{f(t)} = \frac{1}{T} \langle f; 1 \rangle = \frac{1}{\int_0^T 1(t)^2 dt} \int_0^T f(t)1(t)dt$
Donde $1(t) = 1 \forall t \in [0, T]$
- **Varianza Funcional:** $\sigma_{f(t)} = \frac{1}{T} \langle f - \mu_{f(t)} \cdot 1; f - \mu_{f(t)} \cdot 1 \rangle = \frac{1}{T} \int_0^T (f(t) - \mu_{f(t)} \cdot 1)^2$
- **Covarianza Funcional:** la covarianza funcional entre dos funciones $f(t), g(t)$ con $t \in [0, T]$ denotada por $\gamma_{F(t), G(t)}$ se define como:

$$\gamma_{f(t), g(t)} = \frac{1}{\int_0^T 1(t)dt} \int_0^T (f - \mu_{f(t)} \cdot 1(t))(g - \mu_{g(t)} \cdot 1(t))dt$$

En cuanto estadísticos muestrales, análogo al análisis de datos univariado para N réplicas puntuales de la función f (Ramsay & Silverman, 2005) se define:

- **Media Funcional Muestral:** $\bar{f}(t) = N^{-1} \sum_{i=1}^N f_i(t) = \frac{1}{N} \langle f; 1 \rangle$
- **Varianza Funcional Muestral:** $var_f(t) = (N - 1)^{-1} \sum_{i=1}^N [f_i(t) - \bar{f}(t)]^2$
Donde además, $\sqrt{var_f(t)}$ es la función de desviación estándar.
- **Funciones de covarianzas y Correlaciones:** “la función de covarianza resume la dependencia de registros a través de los distintos valores de los argumentos y se calcula para todo t_1 y t_2 mediante” (Ramsay & Silverman, 2005):

$$Cov_f(t_1, t_2) = (N - 1)^{-1} \sum_{i=1}^N \{f_i(t_1) - \bar{f}(t_1)\} \{f_i(t_2) - \bar{f}(t_2)\} \quad (6.13)$$

La función de correlación asociada es entonces:

$$Corr_f(t_1, t_2) = \frac{Cov_f(t_1, t_2)}{\sqrt{var_f(t_1)}\sqrt{var_f(t_2)}}$$

- **Covarianza cruzada y correlación cruzada:** ahora, si se dispone de parejas de observaciones (f_i, g_i) de las funciones $f(t)$ y $g(t)$, la forma en la cual estas dependen una de la otra puede cuantificarse por la función de correlación cruzada:

$$Cov_{f,g}(t_1, t_2) = (N - 1)^{-1} \sum_{i=1}^N \{f_i(t) - \bar{f}(t)\} \{g_i(t) - \bar{g}(t)\}$$

Y así mismo,

$$Corr_{f,g}(t_1, t_2) = \frac{Cov_{f,g}(t_1, t_2)}{\sqrt{var_f(t_1)} \sqrt{var_g(t_2)}}$$

$$Corr_{f,g}(t_1, t_2) = Corr_{g,f}(t_1, t_2)$$

6.2.11. Representación de datos mediante funciones Base

Para la implementación del FDA, es necesario definir los datos discretos o “brutos” en términos de funciones (Levitin et al., 2007); estas se pueden expresar como combinaciones lineales de funciones conocidas denominadas “*funciones base*” para representarlas.

Definición 6.2.10. Sistema de Funciones Base

Un sistema de funciones bases, es un conjunto de funciones conocidas, matemáticamente independientes una de otra y cuya propiedad principal es poder representar cualquier función como una suma ponderada o combinación lineal de un número adecuado de estas (Ramsay & Silverman, 2005). Por ejemplo, para una función $x(t)$ una representación en funciones base de esta mediante una expansión lineal está dada por

$$x(t) = \sum_{k=1}^K c_k \phi_k(t) \tag{6.14}$$

Dónde K es el número de *funciones base*.

Entre los sistemas de bases a partir de los cuales se pueden expresar una función, se destacan:

- **Bases de Fourier.** Estas se definen como la expansión de una serie de potencias de Fourier,

$$1, \text{sen}(wt), \cos(wt), \text{sen}(2wt), \cos(2wt), \dots, \text{sen}(kwt), \cos(kwt), \dots \tag{6.15}$$

Este tipo de base es especialmente útil cuando los datos son periódicos; en caso de no serlo y cuando las curvas son tan complejas que un conjunto de polinomios de orden pequeño no representan las características de las curvas, una alternativa es el uso de *B-splines*.

- **Bases B-splines.** Los B-splines son esencialmente polinomios unidos entre ellos en un intervalo de límites llamados nodos. Esta estructura permite diferenciar cantidades suaves o ásperas en varios lugares de la curva. Los B-splines se caracterizan por su orden el cual es uno más el grado de los polinomios que lo conforman. En general, se suelen usar B-splines de orden 4 (Levitin et al., 2007).

Otros sistemas destacables son (Ramsay & Silverman, 2005)

- **Wavelets**
- **Bases polinomiales**
- **Bases Exponenciales y de potencias**
- **Bases de polígonos**

A continuación, se ampliará más la discusión respecto a la representación en bases B-Splines, los cuáles se usarán en el presente trabajo.

6.2.12. B-Splines

Definición 6.2.11. Un spline es un objeto matemático que replica flexibilidad y adaptabilidad. Esta es una función que se construye por partes a través de funciones polinómicas (Olaya, 2012).

Los B-Splines son una técnica de estimación de curvas a través de sumas de funciones de base que pertenecen a un espacio de funciones cuadrado integrable, los cuales son una generalización de las curvas de Beziér; estas técnicas son mucho más versátiles que los splines debido a que estas se componen de muchas más funciones polinómicas (Olaya, 2012).

Este método consiste en construir n funciones base y a partir de la suma ellas y de la multiplicación de cada una por una constante, obtener nuevas funciones. Por ejemplo, se tienen las funciones $B_1(x)$, $B_2(x)$ y $B_3(x)$ y se tienen las constantes β_1 , β_2 y β_3 , entonces, la nueva función quedaría expresada como:

$$B(x) = \beta_1 B_1(x) + \beta_2 B_2(x) + \beta_3 B_3(x)$$

Por lo que se dice que las funciones $B(x)$ generadas a partir de las funciones $B_i(x)$ pertenecen al subespacio generado por las i -funciones. A manera de notación las funciones de la base se pueden reescribir como $B_{i,n}(x)$ donde n indica el grado del polinomio (Olaya, 2012).

Estos polinomios se construyen en un intervalo $[t_0, t_{N+1}]$, también conocido como valores extremos. Adicionalmente, se supone que se tienen N puntos en el intervalo $[t_0, t_{N+1}]$, los cuales serían $t_1, t_2, \dots, t_N$, los puntos en este intervalo se denominan nodos, estos pueden ser o no equidistantes y satisfacen $t_0 \leq t_1 \leq t_2 \leq \dots \leq t_N \leq t_{N+1}$.

En adelante, se usará la notación $B_k(t, \tau)$ para indicar el valor de la función base B-spline definida por la secuencia de puntos de corte τ . (Ramsay & Silverman, 2005).

6.2.13. Ajuste de Datos a un Sistema de Bases mediante Mínimos Cuadrados Ordinarios

El objetivo es ajustar las observaciones $y_j, j = 1, 2, \dots, n$ usando el modelo $y_j = x(t_j) + \varepsilon_j$, “donde ε_j contribuye a la ‘aspereza’ de los datos brutos” (Ramsay & Silverman, 2005) y usando para $x(t)$ la expansión de la forma (6.14).

Un suavizador simple se obtiene determinando los coeficientes que minimicen la Suma de Cuadrados de los Errores (Ramsay & Silverman, 2005):

$$\text{SMSSE}(y|c) = \sum_{j=1}^n [y_j - \sum_k^K c_k \phi_k(t)]^2 \quad (6.16)$$

Matricialmente, (6.16) se puede expresar como:

$$\text{SMSSE}(y|c) = (y - \Phi c)'(y - \Phi c) \quad (6.17)$$

El lado derecho de la ecuación (6.17) se puede expresar funcionalmente como $\|y - \Phi c\|^2$.

Derivando $\text{SMSSE}(y|c)$ respecto a c se tiene la siguiente ecuación

$$2\Phi'\Phi c - 2\Phi'y = 0 \quad (6.18)$$

y resolviendo para c se obtiene el estimador $\hat{c}$ que minimiza los mínimos cuadrados (Ramsay & Silverman, 2005)

$$\hat{c} = (\Phi'\Phi)^{-1}\Phi'y \quad (6.19)$$

Luego, el vector de valores ajustados $\hat{y}$ es:

$$\hat{y} = \Phi\hat{c} = \Phi(\Phi'\Phi)^{-1}\Phi'y \quad (6.20)$$

El ajuste por Mínimos Cuadrados Ordinarios es preferible, análogo al caso univariado, cuando se tiene que los errores son independientes e idénticamente distribuidos con media 0 y varianza constante σ^2 .

6.2.14. Varianza de un Dato Funcional

Asumiendo $\Sigma_\varepsilon = \sigma^2 \mathbf{I}$ y dado un dato funcional de la forma (6.20), su varianza puede estimarse de la siguiente forma (Ramsay & Silverman, 2005)

$$\text{var}[\hat{y}] = \sigma^2 \Phi (\Phi' \Phi)^{-1} \Phi' \quad (6.21)$$

6.2.15. Elección del número de Funciones Bases K

La elección del tamaño de funciones de la expansión es de vital importancia y uno de los problemas clave en el proceso de suavización no paramétrica. Si se escoge un K lo suficientemente grande se logra ajustar todos los datos, pero por el contrario, si se escoge un K pequeño, se pueden perder aspectos cruciales de la función suave x que son de especial interés y que se desean estimar (Ramsay & Silverman, 2005).

Así, con base a Ramsay & Silverman (2005) se define el intercambio entre varianza y sesgo,

$$MSE[\hat{x}(t)] = E[\{\hat{x}(t) - x(t)\}^2] \quad (6.22)$$

También llamada *función de pérdida* o *Error Cuadrático Medio*; se da a entender entonces, que es aceptable un poco de sesgo siempre y cuando, la reducción de la varianza sea significativa; posteriormente, la escogencia de K se hace con base a la minimización de (6.22). El proceso para escoger K puede realizarse manualmente y determinando el mejor valor de este que a juicio del investigador proporcione ya sea un ajuste visual óptimo o una representación adecuada de la información disponible (Delicado, 2008).

Validación Cruzada

La validación cruzada se ha usado en muchos campos con el fin de elegir los parámetros que proporcionen equilibrio entre precisión y variabilidad o entre bondad de ajuste y capacidad predictiva cuando no hay disponibilidad de una muestra de prueba (Delicado, 2008). A grandes rasgos el procedimiento consiste en “sacar de la muestra consecutivamente cada una de las observaciones x_i , estimar el modelo con los datos restantes (sea $\hat{x}_i$ el estimador obtenido), predecir el dato ausente con ese estimador (así se está haciendo de hecho predicción fuera de la muestra) y finalmente, comparar esa predicción con el dato real” para cada una

de los de K y posteriormente se construye la función

$$ECMP_{CV}(K) = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (6.23)$$

donde (6.23) mide el error de predicción fuera de la muestra para cada valor de K . Finalmente, el valor K que minimice dicha función, será el parámetro de suavización elegido. En pocas palabras la validación cruzada evalúa que tan buena es la precisión de la predicción en un conjunto hipotético de datos, ésta prueba optimiza los parámetros del modelo para garantizar un buen ajuste.

Validación Cruzada Generalizada

Para el estimador de la función de regresión $\mathbf{y} = f(\mathbf{X}) + \varepsilon$,

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{x}) \quad (6.24)$$

La predicción de regresión de cada valor observado x_i puede verse como:

$$\hat{y}_i = \sum_{j=1}^n w^*(x_i, x_j) y_j$$

Matricialmente, $\hat{Y} = SY$, donde $\hat{Y}$ y Y tienen elementos $\hat{y}_i$ y y_i respectivamente, y S es la matriz de suavización cuyo elemento (i, j) es $s_{ij} = w^*(x_i, x_j)$; esta, es análoga a la matriz *HAT*, por lo que la expresión puede verse también como $\hat{Y} = HY$ (Delicado, 2008). Se puede demostrar qué:

$$ECMP_{CV}(K) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - s_{ii}} \right)^2 \quad (6.25)$$

Mediante una modificación del criterio de validación cruzada se obtiene la *Validación Cruzada Generalizada*. Esta modificación consiste en sustituir en la fórmula (6.25) los valores s_{ii} de la diagonal de S por su promedio:

$$ECMP_{GCV}(K) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - \nu/n} \right)^2 \quad (6.26)$$

Donde $\nu = \text{Traza}(S) = \sum_{i=1}^n s_{ii}$ y se denomina *número de parámetros efectivos* del estimador

no paramétrico correspondiente a la matriz S (Delicado, 2008).
Tras manipular (6.26), se tiene

$$ECMP_{GCV}(K) = \frac{n\hat{\sigma}_\varepsilon^2}{n - \nu}, \quad (6.27)$$

Donde

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n - \nu} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6.28)$$

Es un estimador de las varianzas de los errores del modelo. Finalmente, el valor K que minimiza dicha función, es el parámetro de suavización elegido por el criterio de validación cruzada generalizada (Delicado, 2008).

6.3. Geoestadística Funcional

Siendo el FDA una de las ramas más recientes en la estadística, la literatura existente en lo que concierne a las aplicaciones de este para datos espaciales es muy escasa, sin dejar de ser uno de los posibles usos más atractivos para los datos funcionales (Delicado et al., 2010). El trabajo de Giraldo (2009) es el referente más extenso y detallado del tema, así como los trabajos de Delicado et al. (2010), Giraldo et al. (2011) y más recientemente, Kokoszka & Reimherr (2017) quienes ofrecen un acercamiento al tema en un capítulo de su libro.

A continuación, se abordarán los aspectos claves de la geoestadística funcional los cuales serán las bases para la solución de los problemas planteados al inicio del documento y servirán de apoyo para alcanzar los objetivos planteados.

Uno de los antecedentes primarios en lo que se refiere a la geoestadística funcional citados por Giraldo (2009) es el de Goulard & Voltz (1993). En este trabajo los autores propusieron un método para interpolar curvas de resistencia a la penetración en sitios no muestreados. Los autores, tomaron un conjunto de M puntos observados $(s_i(t_j), j = 1, \dots, M, i = 1, \dots, n)$ a las cuales se ajustó un modelo paramétrico para reconstruir la curva completa asumiendo tanto un número de observaciones como de parámetros pequeño.

Giraldo (2009) retoma este proceso y añade además, como paso previo la suavización no paramétrica de las curvas observadas, y generaliza además, el caso a un número considerable de parámetros y observaciones superando las restricciones impuestas por Goulard & Voltz (1993). Además, Giraldo (2009), presenta una generalización al caso funcional de variables regionalizadas; cabe destacar, que en general, el FDA se puede adaptar para los tres tipos

de variables espaciales (Datos geoestadísticos, datos de área y patrones puntuales) (Delicado et al., 2010).

Definición 6.3.1. Se define, el proceso espacial funcional de acuerdo a Giraldo (2009) quien generaliza las definiciones de proceso espacial propuesta por Cressie (1993) y de variables aleatorias funcionales propuesta por Ferraty & Vieu (2006) al caso de variables funcionales regionalizadas:

$$\{\chi_s : s \in D \subseteq \mathbb{R}^d\} \quad (6.29)$$

Con s , siendo la localización genérica en el espacio euclidiano d -dimensional (d usualmente es 2); el conjunto $D \subseteq \mathbb{R}^d$ puede ser en este caso, fijo o aleatorio.

χ_s es una variable aleatoria funcional tomando valores en el espacio infinito dimensional o espacio funcional (Giraldo, 2009).

Momentos de un Proceso Espacial funcional

Se asumen para (6.29), las mismas características de un proceso univariado, como lo son estacionariedad en segundo orden e isotropía. Formalmente:

- $E[\chi_s(t)] = \mu_s(t)$ y $\text{Var}[\chi_s(t)] = \sigma_s^2(t)$
- $\text{cov}[\chi_s(t), \chi_{s+h}(u)] = C(h; t, u)$ con h , el vector de rezago espacial y para todo $t, u \in [a, b]$. Esto asegura, que la varianza del proceso asociado $C(0; t, u)$ existe y es finita. Se puede usar la notación $C(h; t)$ para $C(h; t, t)$.

Estos supuestos, implican que existe un mapeo $h \rightarrow (h; t, u)$ tal que:

$$\begin{aligned} \text{Var}[\chi_{s+h}(t) - \chi_s(u)] &= E[(\chi_{s+h}(t) - \chi_s(u))^2] \\ &= \gamma(h; t, u) = C(0; t, u) - C(h; t, u) \quad \forall t \in [a, b] \end{aligned} \quad (6.30)$$

La expresión (6.30) es conocida como *variograma funcional*.

Definición 6.3.2. Variograma Funcional. Para el caso funcional, la estructura de correlación espacial puede adaptarse mediante algunas modificaciones menores. Así, bajo el supuesto de media constante, Giraldo (2009) define la traza-variograma funcional como

$$\gamma(h) = \text{Var}[\chi_{s_i}(t) - \chi_{s_j}(t)] = E[(\chi_{s_i}(t) - \chi_{s_j}(t))^2] \quad (6.31)$$

Mediante el teorema de Fubini y a partir de la definición de valor esperado,

$$\gamma(h) = \frac{1}{2} \mathbb{E} \left[\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt \right] \quad (6.32)$$

Donde $s_i, s_j \in D$ y $h = \|s_i - s_j\|$

El estimador clásico por el método de los momentos (6.5) puede adaptarse fácilmente al caso funcional y se obtiene así, el estimador de (6.32),

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt, \quad (6.33)$$

Siendo $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$ el total de diferentes elementos en $N(h)$.

Nuevamente, para datos que no se encuentran igualmente espaciados, se propone la misma modificación a $N(h)$ tal como para el estimador (6.5).

Con la traza-variograma empírica estimada, se procede a realizar su ajuste a alguno de los modelos teóricos de semivarianza ya conocidos, por ejemplo aquellos en (6.1) por algún método de estimación como OLS, WLS o GLS. Nótese que la traza-variograma ajustada $\hat{\gamma}_\alpha(h)$ es siempre válida dado que hereda las propiedades de un variograma teórico ajustado de un conjunto de datos univariados (Giraldo, 2009).

6.3.1. Kriging Funcional Ordinario

Giraldo (2009) retoma el método propuesto por Goulard & Voltz (1993) superando las restricciones impuestas por dicho método como lo son un pequeño número de observaciones por función y de modelación paramétrica.

El predictor propuesto por los autores, se denomina *Kriging Ordinario para datos funcionales* (u OKFD por sus siglas en inglés), donde previo al procesamiento espacial de las curvas, las funciones observadas se someten a una suavización no paramétrica. Nuevamente, el objetivo en este estudio es la predicción de un dato no observado en un punto no muestreado s_0 , salvo que esta vez, no se considerará una realización de un proceso univariado, sino, una realización de una variable funcional o curva $\chi_{s_0}(t)$.

Goulard & Voltz (1993) definen como el Mejor Predictor Lineal Insesgado (**BLUP**, por sus siglas en inglés) para χ_{s_0} a:

$$\hat{\chi}_{s_0}(t) = \sum_{i=1}^n \lambda_i \chi_{s_i}(t) \quad \forall t \in [a, b], \lambda_1, \lambda_2, \dots, \lambda_n \in \mathbb{R} \quad (6.34)$$

Los $\lambda_i, i = 1, 2, \dots, n$ se estiman de tal manera que se minimice el error cuadrático medio de la predicción:

$$\mathbb{E} \left[\int_a^b (\hat{\chi}_{s_0}(t) - \chi_{s_0}(t))^2 dt \right] = \int_a^b \text{Var} [\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)] dt$$

Lo cual se reduce al siguiente problema de optimización

$$\begin{aligned} \min_{\lambda_1, \lambda_2, \dots, \lambda_n} \quad & \int_a^b \text{Var} [\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)] dt \\ \text{sujeto a} \quad & \sum_{i=1}^n \lambda_i = 1 \end{aligned} \tag{6.35}$$

Donde $\sum_{i=1}^n \lambda = 1$ es la restricción de insesgadez. Observese que a partir del teorema de Fubini y la restricción de insesgadez, se tiene

$$\begin{aligned} \int_T \text{Var} [\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)] dt &= \int_T \mathbb{E} [(\hat{\chi}_{s_0}(t) - \chi_{s_0}(t))^2] dt \\ &= \mathbb{E} \left[\int_T (\hat{\chi}_{s_0}(t) - \chi_{s_0}(t))^2 dt \right]. \end{aligned}$$

Resolviendo (6.35) y teniendo en cuenta que $\gamma(h) = C(0) - C(h)$, los pesos óptimos $\lambda_i, i = 1, 2, \dots, n$ pueden encontrarse como solución del siguiente sistema de ecuaciones

$$\begin{bmatrix} \int_T \gamma_t(|s_1 - s_1|) dt & \cdots & \int_T \gamma_t(|s_1 - s_1|) dt & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \int_T \gamma_t(|s_n - s_1|) dt & \cdots & \int_T \gamma_t(|s_n - s_n|) dt & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -m \end{bmatrix} = \begin{bmatrix} \int_T \gamma_t(|s_0 - s_1|) dt \\ \vdots \\ \int_T \gamma_t(|s_0 - s_n|) dt \\ 1 \end{bmatrix}$$

El predictor basado en la ecuación (6.34) hace uso de la función de traza-variograma por $\gamma(h) = \int_a^b \gamma(h; t, t) dt, h \geq 0$ para obtener los coeficientes $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_n$ por lo que los cálculos serían parecidos a los del kriging para datos univariados.

Se puede demostrar adicionalmente, que la traza-varianza de la predicción por kriging

funcional ordinario basado en la traza-variograma está dada por

$$\sigma_{s_0}^2 = \int_T \text{Var}[\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)] = \sum_{i=1}^n \lambda_i \int_T \gamma_t(||s_i - s_0||) dt - m$$

$$\sum_{i=1}^n \lambda_i \gamma(||s_i - s_0||) - m. \quad (6.36)$$

Adicionalmente, Giraldo (2009) propone otras variaciones a este estimador incluyendo la posibilidad de indexarlo respecto a un parámetro t y se denomina estimador completamene funcional; no obstante dada la complejidad que supone dicha estimación, el estimador a usar en el siguiente trabajo será el OKFD.

Capítulo 7

Metodología

El presente capítulo al estar enmarcado en la teoría geoestadística, puede dividirse en dos subsecciones, correspondientes al análisis estructural y al proceso predictivo como ya se mencionó anteriormente. Adicionalmente, se debe incluir una etapa previa que correspondiente a la construcción de los datos funcionales, mediante la suavización no paramétrica de las curvas a estudiar Giraldo (2009).

7.1. Datos

Los datos a analizar corresponden a mediciones tomadas de la precipitación en 92 estaciones climáticas en el periodo 1983-2011 ubicadas en la región de valle y montaña del departamento del Valle del Cauca (Figura 7.1). Estos datos fueron suministrados por la Corporación Autónoma Regional del Valle del Cauca (CVC) y completados por Ospina (2016) en su tesis de maestría, mediante geoestadística univariada.



Figura 7.1: Zona de Estudio. Fuente: IGAC (2012)

Inicialmente se consideró trabajar solo con aquellas estaciones correspondientes a la región del valle geográfico del río Cauca, pero dado que dicha red de puntos de muestreo no es lo suficientemente densa para llevar a cabo el proceso de predicción espacial se trabajó con la totalidad de estas; es importante mencionar que del total de observaciones anuales suministradas por la CVC y por Ospina (2016) solo se emplearon aquellas del último año de observación (2011).

En primer lugar, se procedió a caracterizar espacialmente el comportamiento de la variable precipitación. Para ello, un primer paso fue analizar la distribución espacial de los puntos de medición.

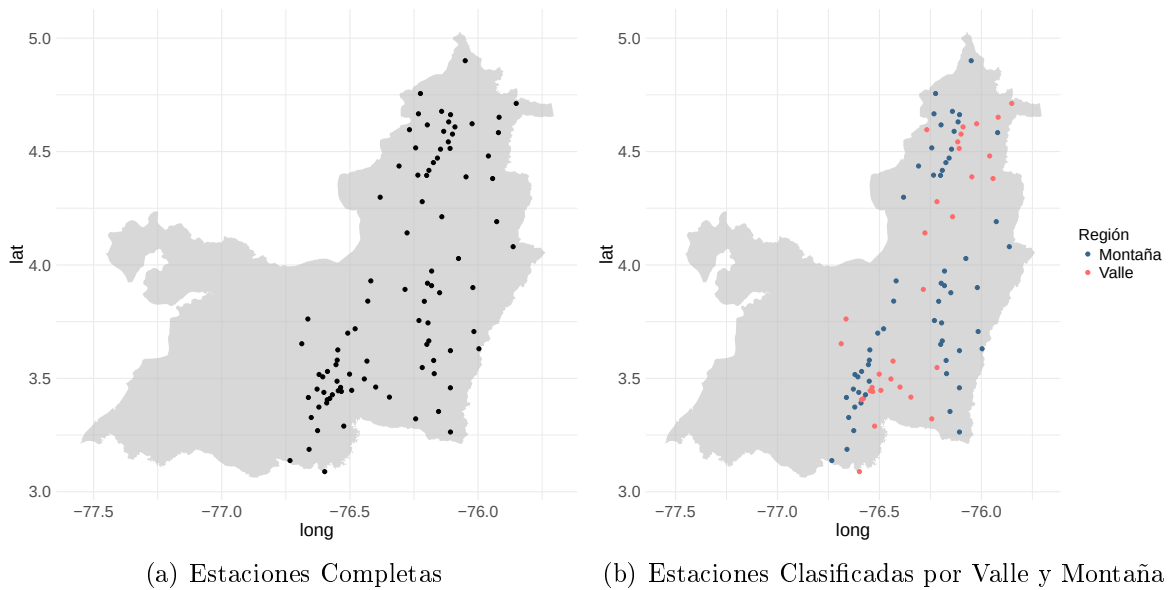


Figura 7.2: Mapa de las estaciones Climáticas

Inicialmente, se graficaron la totalidad de las 92 estaciones proporcionadas por la CVC, como se muestra en el panel izquierdo de la figura (7.2). Ahora bien, si se decide clasificar entre las estaciones del valle o las de montaña, estas se distribuyen espacialmente como se muestra en el panel derecho de la figura (7.2); se nota que existen dos grupos distinguibles para las estaciones ubicadas en el valle geográfico, uno hacia la parte norte de éste y otro hacia al sur, mientras, hacia la zona central del mismo, no existe un conjunto considerable de estaciones, lo cual dificultaría el proceso de predicción al no contarse con una red suficientemente densa de observaciones en este sitio. Por tanto contar con ambos grupos de estaciones disponibles en dicha zona (Valle y Montaña), permitió tener un conjunto considerable de observaciones, lo cual facilita el proceso de predicción espacial.

En concreto, de las 92 estaciones disponibles, sólo 32 se encuentran en la región del valle. La altura promedio de las estaciones fue de 1498 msnm con una mediana de 1494 msnm y una desviación estándar de 532.6857 msnm; el coeficiente de variación de la altura en msnm fue de 35.6 %, por lo cual, la altura a la cual se encuentran ubicadas las estaciones es variable con un coeficiente de variación considerable.

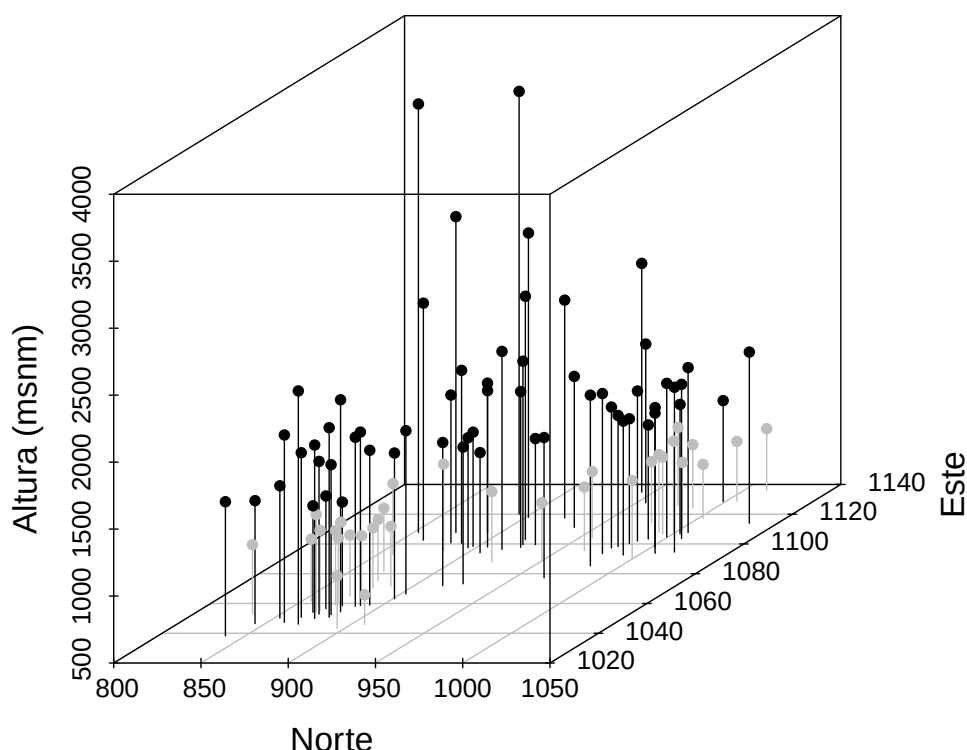


Figura 7.3: Estaciones según altura

A partir de la Figura 7.3 es posible identificar que la gran mayoría de las estaciones se encuentran ubicadas en la parte de montaña, tal como ya se había especificado y existen tres estaciones que se ubican en puntos extremos de altura, más exactamente a 2858, 3656 y 3699 msnm correspondientes a las estaciones de Bolo Blanco ubicada en el municipio de Pradera, Bellavista en el Municipio de Palmira y Los Alpes en el municipio de Florida.

7.2. Construcción de los Datos de Precipitación Pentadal

Posterior a la selección de las estaciones y el periodo a trabajar, se procedió a construir los pentadales; se define como un *pentadal* a la precipitación acumulada cada cinco días. Esto

es, si se define x_i como la precipitación en el día i , $i = 1, 2, \dots, 365$, entonces, los pentadales vendrán dados por $p_1 = \sum_{i=1}^5 x_i$, $p_2 = \sum_{i=6}^{10} x_i, \dots, p_{72} = \sum_{i=360}^{365} x_i$. El objetivo de trabajar con dicha escala es tanto disminuir el ruido de las observaciones, las cuales poseen demasiados valores iguales a cero, como permitir comparar los resultados obtenidos aquí con aquellos de trabajos de igual naturaleza (Ospina, 2016, Verdin et al., 2015, 2016).

Por ejemplo, de acuerdo con la Figura 7.4 que corresponde a la estación de la Balastera con coordenadas geográficas (1053.277, 833.3149) la precipitación diaria presenta el siguiente comportamiento (Figura 7.4),

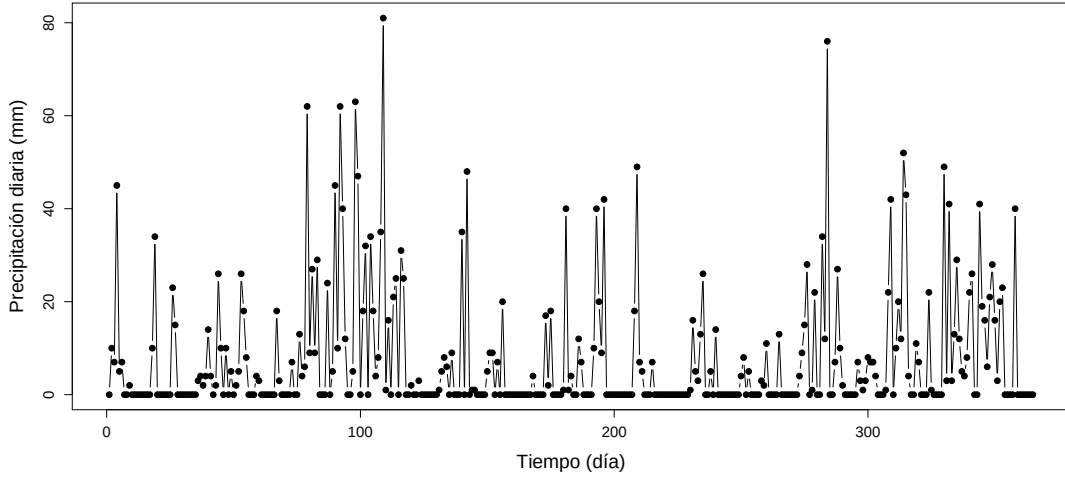


Figura 7.4: Precipitación diaria para la estación la Balastera

Tras acumular cada cinco días, se tienen los correspondientes pentadales los cuales se muestran en la siguiente figura (Figura 7.5)

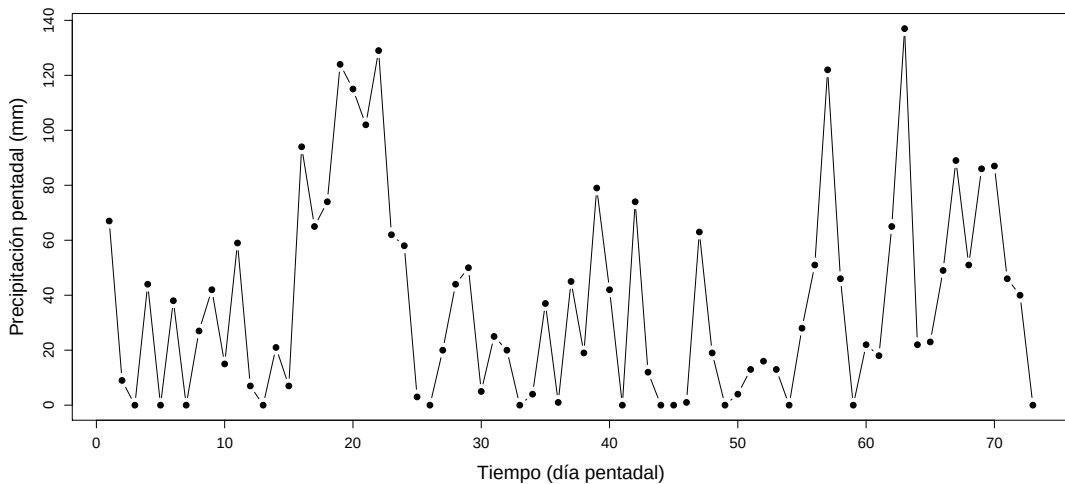


Figura 7.5: Precipitación pentadal para la estación la Balastera

Se aprecia que este proceso disminuye considerablemente la variabilidad de las observaciones y el ruido del proceso subyacente, propiedad necesaria para llevar a cabo la suavización de las curvas (Ramsay & Silverman, 2005). Este proceso se repitió para cada una de las 91 estaciones restantes.

El proceso posterior a la construcción de los pentadales por estación, es la caracterización univariada para cada una de estas.

7.3. Análisis Exploratorio

Con los pentadales construidos, el siguiente paso es la caracterización de los datos pentadales anuales por estación mediante medidas descriptivas univariadas y mediante la visualización de posibles comportamientos que podrían influenciar la conducta de las curvas suavizadas.

Primeramente, se graficaron en mapas las ubicaciones de cada una de las 92 estaciones del estudio a partir de sus respectivas coordenadas. Este tipo de herramienta ayudó a preestablecer si existían comportamientos en el espacio que pudieran influenciar el proceso predictivo espacial.

Posterior a esto, el siguiente paso fue obtener medidas que permiten resumir las características más importantes de la precipitación pentadal para cada una de las estaciones. Esto se logró mediante métodos descriptivos univariados convencionales, tales como medidas de tendencia central y de dispersión.

Como siguiente paso en el análisis exploratorio, se graficó la distribución de cada estación en gráficos de cajas y alambres, estos ayudaron a visualizar de manera global los aspectos más relevantes de la variabilidad por estación de los datos. Además, se realizó el análisis mensual de dicha variable mediante los métodos descriptivos previamente mencionados.

7.4. Análisis de Datos Funcionales

7.4.1. Selección del número de funciones base K y el parámetro de penalización por Rugosidad

El primer paso en el análisis geoestadístico funcional corresponde a la parte funcional. Esta fase consiste en expresar mediante funciones base los datos discretos. Como ya se mencionó en el capítulo (6), existen diversos sistemas mediante los cuales se puede llevar a cabo dicha tarea, el sistema de bases usado en este caso fue el sistema de bases B-splines, más exactamente, B-splines de orden cuatro o cúbicos. Su elección se basó tanto en la ventaja que presentan estos sobre sistemas base como el de Fourier a la hora de describir

comportamientos de los datos con pocas bases, así como mejores resultados en el proceso predictivo de geoestadística funcional, como lo mencionan Giraldo et al. (2011). Adicional al número de bases, debe encontrarse el valor de η , denominado parámetro de penalización por rugosidad, el cual controla el intercambio entre la varianza y el sesgo dada una penalización por la segunda derivada. Giraldo et al. (2011) señalan que si dicho parámetro tiende a infinito, el ajuste de las bases a los datos discretos será una recta de regresión lineal.

Si se define la base de B-splines cúbicos,

$$x(t) = \sum_{k=1}^{L+4} c_k \phi_k(t) = \mathbf{c}^T \phi(t) \quad (7.1)$$

Siendo L el número de puntos interiores de $T = [a, b]$, esto es, $a < \tau_1 < \tau_2 < \dots < \tau_L < b$ (definiendo $\tau_0 = a$ y $\tau_L = b$) en este caso, $a = 1$ y $b = 73$ y $\mathbf{c}$ es el vector de coeficientes c_k y $\phi(t)$ es una función $L + 4$ dimensional. Adicionalmente, y teniendo en cuenta perturbaciones en la función $y_j = f(t_j) + \varepsilon_j$, los coeficientes $\mathbf{c}$ se hallan entonces resolviendo el siguiente problema,

$$\min_{\mathbf{c} \in \mathbb{R}^{L+4}} \sum_{j=1}^M (y_j - \phi(t_j))^2 + \int_T (\phi''(t))^2 dt \quad (7.2)$$

Así, para llevar a cabo el proceso de suavización, se realizó lo siguiente,

- Primero se escogió el número óptimo de funciones base para la suavización, para ello y usando validación cruzada generalizada se halló aquél valor K y el valor η correspondiente a la penalización por rugosidad que proporcionaran la menor suma de cuadrados de los errores
- Hallado el valor de K y de η , y mediante la siguiente relación (Ramsay & Silverman, 2005)

$$\text{número de bases} = \text{orden de los splines} + \text{longitud de los puntos de corte} - 2$$

se determinó el total de nodos a utilizar en la suavización. Con el número de nodos, el parámetro de penalización y el total de funciones base a utilizar definidos, se procedió a ajustar los datos a una base B-splines mediante mínimos cuadrados.

Este proceso se describe en detalle a continuación.

Validación Cruzada Generalizada

El proceso de validación cruzada generalizada hace uso de la función 6.27, la cual, se propone minimizar.

Así, se define un conjunto de parámetros iniciales donde se busca encontrar aquellos que generen el menor valor de (6.27).

Tomando como intervalos para η y K $[0.0001, 0.0002, \dots, 0.1]$ y $[6, 7, \dots, 71, 72]$ respectivamente, se evaluó la disminución en la suma de cuadrados para cada combinación de estos valores, hasta hallar aquella que proporcionara el menor valor posible.

En cuanto al número de funciones base, el número K que proporcionó el menor valor de $ECMP_{CV}$ fue 6. Los resultados correspondientes al proceso, se resumen en la siguiente tabla (Tabla 7.1)

Tabla 7.1: Validación Cruzada Generalizada para el número óptimo de funciones Base

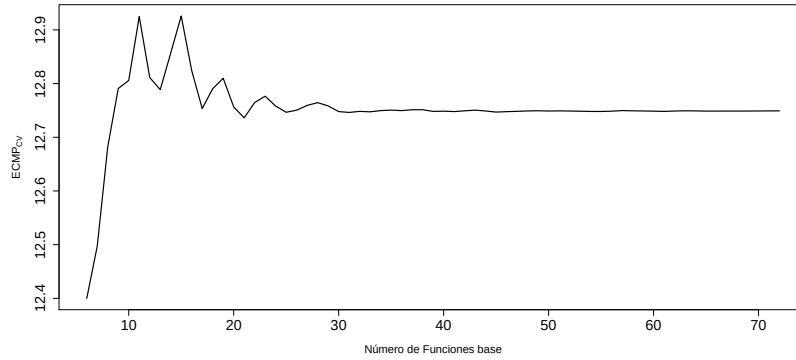
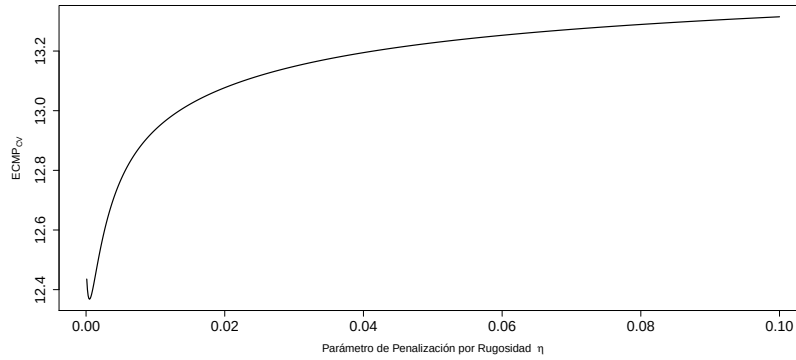
K	$ECMP_{CV}$
6.0000	12.3999
7.0000	12.4973
8.0000	12.6822
9.0000	12.7907
10.0000	12.8058
$\vdots$	$\vdots$
70.0000	12.7490
71.0000	12.7491
72.0000	12.7491

Asimismo, se determinó el valor de η que minimizará el $ECMP_{CV}$; del posible rango de valores asignados a η se halló, que $ECMP_{CV}$ es minimizado cuando $\eta = 0.0005$. Los resultados igualmente se resumen en la Tabla 7.2 que se muestra a continuación.

Tabla 7.2: Validación cruzada para el valor óptimo del parámetro de penalización

η	$ECMP_{CV}$
0.0001	12.4355
0.0002	12.3999
0.0003	12.3799
0.0004	12.3704
0.0005	12.3679
$\vdots$	$\vdots$
0.0996	13.3144
0.0997	13.3145
0.0998	13.3146
0.0999	13.3147
0.1000	13.3149

Gráficamente, el proceso puede resumirse mediante las siguientes figuras (Figura 7.6),

(a) Validación Cruzada Generalizada para K (b) Validación Cruzada Generalizada para η **Figura 7.6:** Resultados de la Validación Cruzada Generalizada

En la figura 7.6 se aprecia como la función $ECMP_{CV}$ alcanza sus correspondientes mínimos

para los valores previamente mencionados de K y η y además, los valores a partir de los cuales dicha función converge en sus respectivos casos, alrededor de 30 para K y aproximadamente 0.90 para el η .

Tras escoger el número de bases y el parámetro de rugosidad que minimice la Suma de Cuadrados de los Errores, el siguiente paso es el ajuste por Mínimos cuadrados Ordinarios al sistema de bases para cada una de las estaciones.

7.4.2. Ajuste de las funciones Base mediante Mínimos Cuadrados

La estimación mediante Mínimos Cuadrados Ordinarios es análoga al caso univariado y propende por la minimización del cuadrado de la diferencia entre los valores observados y los ajustados. Para este proceso se recurrió a los elementos presentados en la sección (6.2.13) del capítulo (6).

Cabe anotar que dado el elevado número de valores de precipitación iguales a cero, el proceso tiende a estimarlos como negativos (Ramsay & Silverman, 2005), por ello se hizo necesario realizar una transformación que garantizara valores positivos sobre el dominio de las curvas suavizadas.

Para garantizar lo anterior, Giraldo et al. (2011), Reyes et al. (2012) y Marín (2004) trabajaron con la raíz cuadrada de los datos y posteriormente, tras obtener la suavización de estos valores, se elevaron al cuadrado para obtener las curvas de la variable en su escala original. En el presente trabajo, se utilizó esta metodología; así, los valores suavizados son no negativos y consecuentemente, lo son las estimaciones de las medidas descriptivas funcionales como la media, la varianza y la desviación estándar funcional.

Así, el proceso para el ajuste mediante mínimos cuadrados ordinarios de los valores discretos a funciones base puede resumirse como sigue,

- Se definió el número de funciones base y el parámetro de suavización hallados en (7.4.1).
- A partir de estos, se define el número de funciones que conforman las bases B-splines las cuales se ajustaron a los datos discretos.
- Posteriormente, se ajustaron las bases de tipo spline con la que se va a trabajar a los datos mediante Mínimos Cuadrados Ordinarios.

Entonces, con el número de funciones que conforman una base y el parámetro η definido, las funciones base para la suavización fueron las siguientes,

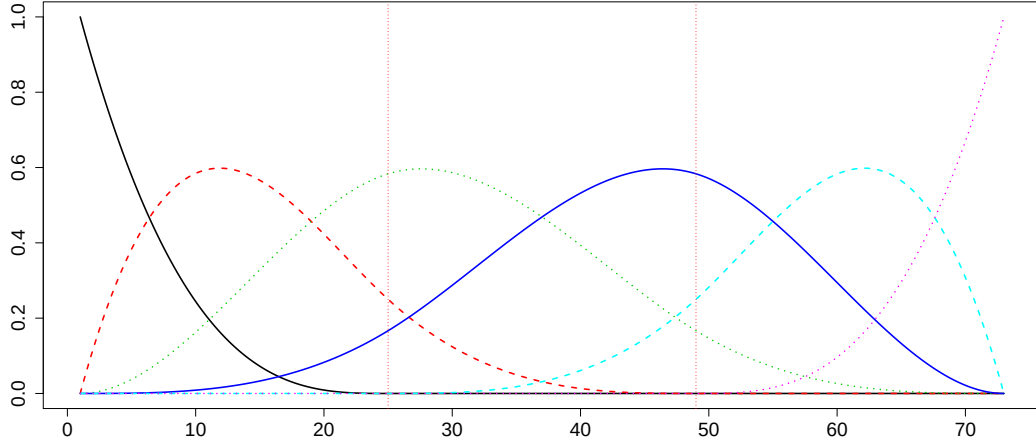


Figura 7.7: Bases B-splines

Donde las líneas rojas punteadas verticales en la Figura 7.7 corresponden a los nodos interiores utilizados; los valores correspondientes a dichos nodos son, $\tau_1 = 25, \tau_2 = 49$.

Previo al ajuste, debe tenerse en cuenta las posibles complicaciones producto de los valores iguales a cero en la base de datos (aproximadamente el 18 % de las observaciones son cero), ya que como lo menciona Ramsay & Silverman (2005), esto conllevaría a que el ajuste produzca valores negativos, lo cual en el estudio carece de sentido dada la naturaleza de la variable precipitación.

La solución aquí planteada consiste en realizar el ajuste de los valores de la raíz de la precipitación y posterior a esto, elevar los resultados al cuadrado garantizando así valores de precipitación positivos para la suavización (Giraldo et al., 2011).

Entonces, sea la matriz de datos pentadales,

$$P = [P_1 \ P_2 \ \dots \ P_{92}] \quad (7.3)$$

Donde (7.3) es una matriz 73×92 , cada columna de esta representa una de las 92 estaciones y cada fila está asociada a un pentadal; esto es, $[p_{ij}]$ es el pentadal i -ésimo de la j -ésima estación, con $i = 1, \dots, 73, j = 1, 2, \dots, 92$.

Luego, cada $p_{ij} \in P$ se elevó a la $1/2$ y se construyó la matriz P' . Para cada columna de P' se realizó el procedimiento de suavización mediante MCO. Reemplazando cada columna de

P' en la expresión (6.16) de la sección (6.2.13) del capítulo (6), entonces,

$$\text{SMSSE}_j(P'_j|c_j) = \sum_{i=1}^{73} [P'_{ij} - \sum_k^6 c_{jk} \phi_k(t_i)]^2 \quad \forall j = 1, 2, \dots, 92 \quad (7.4)$$

Hallando los valores $\hat{c}_j$ que para cada columna de P' disminuyan (7.4), se tienen posteriormente los valores ajustados,

$$\hat{P}'_j = \Phi \hat{c}_j = \Phi (\Phi' \Phi)^{-1} \Phi' P'_j \quad (7.5)$$

Luego, para la matriz de valores ajustados $\hat{P}'$, se elevó cada elemento $\hat{p}'_{ij}$ al cuadrado y se obtuvieron así, los valores de precipitación suavizados positivos. Estos valores se almacenaron en la matriz P^* .

$$P^* = [P_1'^2 \ P_2'^2 \ \dots \ P_{92}'^2] \quad (7.6)$$

Con estos valores, se llevó a cabo el proceso de análisis estructural para determinar la estructura de correlación espacial y posteriormente se realizó el proceso de predicción espacial para curvas.

7.4.3. Análisis Exploratorio Funcional

Con las curvas de precipitación suavizadas para cada una de las estaciones, se procedió a caracterizarlas descriptivamente. Análogo al caso univariado, algunas medidas que permiten resumir aspectos relevantes de los datos funcionales son la media, la varianza y la desviación estándar funcional. Asimismo, una medida útil de la estructura de correlación entre los datos respecto al parámetro que los indexa es la covarianza cruzada y consecuentemente la correlación cruzada, la cual asocia para distintos tiempos como se relacionan todas y cada una de las curvas de par en par (Ramsay & Silverman, 2005).

Estas medidas se toman punto a punto, por lo cual una manera apropiada de observar el comportamiento de las mismas es visualmente.

Para este caso, la atención se centró en la media y la desviación. En el caso de la primera medida; de acuerdo a su definición, para cada instante de tiempo t se procedió a promediar los valores de precipitación pentadal para las 92 curvas; así, la media funcional muestral en este caso está dada como,

$$\bar{p}(t) = (92^{-1}) \sum_{j=1}^{92} P_j^*(t), t = 1, \dots, 73$$

En el caso de la desviación estándar funcional, el procedimiento es análogo,

$$\sqrt{Var_p(t)} = \{(91)^{-1} \sum_{j=1}^{92} [P_j^*(t) - \bar{p}(t)]^2\}^{1/2}, t = 1, \dots, 73$$

7.5. Geoestadística Funcional

7.5.1. Determinación de la estructura de Correlación Espacial

Para establecer la estructura de correlación de las curvas pentadales suavizadas se recurrió a los métodos de geoestadística funcional propuestos por Giraldo (2009) en su trabajo de tesis doctoral.

En primer lugar y con las curvas suavizadas, el primer paso fue estimar la trazo-variograma, mediante la traza-variograma empírica expuesta en la sección (6.3). El proceso es análogo al análisis estructural para datos univariados. La principal diferencia entre ambos métodos radicó en que se consideraron curvas en vez de realizaciones puntuales de una variable aleatoria indexada en el espacio. Este estimador se construyó a partir de 6.33; se tomó una distancia máxima entre estaciones de 50 *km* dado que después de esta la variabilidad espacial es despreciable y además, se consideró un total de 14 lags o rezagos espaciales los cuales garantizaron un equilibrio entre el ruido y el sesgo del estimador.

Así, la traza-variograma empírica para este caso viene dada por,

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt,$$

Con $i, j = 1, 2, \dots, 92 \wedge i \neq j$, $T = [1, 73]$ y $N(h) = 14$.

Posterior a la estimación de la traza-variograma, se procedió al ajuste por mínimos cuadrados ponderados del modelo de correlación espacial para las curvas de manera similar al caso univariado. Se consideraron para este caso los modelos de semivarianza acotados, Gaussiano, Esférico, Exponencial, y Matérn presentados en el capítulo (6) en la sección (6.2.3); luego, el mejor modelo se seleccionó con base a la reducción más significativa de la Suma de Cuadrados de los Errores.

Una vez encontrada la estructura de correlación espacial, se procedió a realizar el proceso de predicción en sitios no muestreados.

7.5.2. Kriging funcional ordinario y predicción espacial

Con el fin de determinar la efectividad del modelo de predicción espacial, se procedió a eliminar las estaciones observadas, una a la vez y posteriormente se predijeron mediante Kriging Funcional Ordinario (OKFD por sus siglas en inglés) haciendo uso del predictor (6.34) presentado en la sección (6.3.1) del capítulo (6).

Así, el procedimiento para evaluar el poder predictivo del modelo fue el siguiente

1. En primer lugar se selecciona de manera aleatoria una estación del total de 92 sin reemplazo.
2. Se elimina la estación seleccionada en el paso previo y se estiman los valores de la precipitación a partir de los observados en las 91 estaciones restantes; para ello, se usa el estimador de kriging funcional ordinario (expresión (6.34) la sección 6.3 del capítulo 6), cuya forma está dada en este caso por,

$$\hat{\chi}_{s_j}(t) = \sum_{i=1}^{91} \lambda_i \chi_{s_i}(t) \quad \forall t \in [1, 73], \lambda_1, \lambda_2, \dots, \lambda_{91} \in \mathbb{R}$$

Donde $\chi_{s_i}(t)$ es la i -ésima estación de las restantes, $i = 1, 2, \dots, 91$ y $j = 1, 2, \dots, 30$.

3. Una vez se predice el valor para la estación, se calculan los residuales de cada valor como la diferencia entre cada valor observado y el predicho; entonces,

$$e(t) = \chi_{s_j}(t) - \hat{\chi}_{s_j}(t) \tag{7.7}$$

4. De los residuales, interesa conocer su esperanza y desviación. Por ello, se calcula como se ha venido realizando la función de media estimada para estos y su función de desviación estándar (punto a punto) así como se calcularon en (7.4.3),

$$\bar{e}(t) = (92^{-1}) \sum_{j=1}^{92} e_j(t), t = 1, \dots, 73$$

$$\sqrt{Var_e(t)} = \{(91)^{-1} \sum_{j=1}^{92} [e_j(t) - \bar{e}(t)]^2\}^{1/2}, t = 1, \dots, 73$$

Para el caso de la media, si la predicción fue óptima, su valor será cero, indicando insesgadez en las predicciones y por ende el estimador captura la variabilidad espacial subyacente entre las curvas observadas.

Por otro lado si no existió una gran variabilidad en las observaciones y el predictor (6.34) logró capturarla, el valor de la desviación estándar funcional se mantendrá constante a lo largo del intervalo de variación de t .

5. Una vez obtenidos estos valores, se grafican y se analizan con el fin de detectar si existen, comportamientos asociados a las curvas; adicionalmente, se compara el comportamiento de todos los valores observados con los predichos mediante un gráfico de dispersión. Es de esperarse bajo condiciones ideales que la predicción sea lo más parecida al valor real, por tanto estos valores tenderán a aproximarse a la recta $x = y$ si el estimador fue óptimo.

En el siguiente capítulo se presentarán los resultados de la aplicación de la metodología aquí planteada.

Capítulo 8

Resultados y Discusión

Tras la aplicación de la metodología planteada en el capítulo anterior (Capítulo 7), se plasmarán los resultados obtenidos con base en ésta. El orden de esta sección está definido de la siguiente forma: primero se expondrá el análisis descriptivo de las estaciones estudiadas, posteriormente se mostrarán los resultados del proceso de suavización para las curvas pentadales observadas, consecuentemente se analizarán las curvas obtenidas mediante métodos descriptivos para datos funcionales; los resultados que le siguen a los anteriormente descritos corresponden al análisis estructural, donde se incluye la estructura de correlación espacial estimada para las curvas pentadales. En conjunto con esto se expondrán los diferentes modelos teóricos que se ajustaron para la traza-variograma experimental escogiendo aquel que generó la menor Suma de Cuadrados del Error.

Finalmente, a partir del modelo teórico ajustado, se utilizaron los parámetros de este en el predictor de Kriging Ordinario funcional; para validar dichos resultados y mediante validación cruzada, se procedió a eliminar de manera paulatina una a una las 92 estaciones observadas y a predecir las curvas correspondientes para estudiar la efectividad del predictor. Consecuentemente se estudió si el ajuste fue adecuado mediante el análisis de los residuales y finalmente se dictaminó la efectividad del modelo para el proceso predictivo.

8.1. Análisis Exploratorio

8.1.1. Construcción de los Pentadales

En primer lugar, se procede a construir las curvas pentadales para cada una de las 92 estaciones como se especificó en el capítulo correspondiente a la metodología (Capítulo 7); esto se hace con el fin de reducir la variabilidad y/o ruido subyacente en las observaciones.

De acuerdo con la Figura 8.1 la precipitación diaria para cada una de las estaciones presenta el siguiente comportamiento

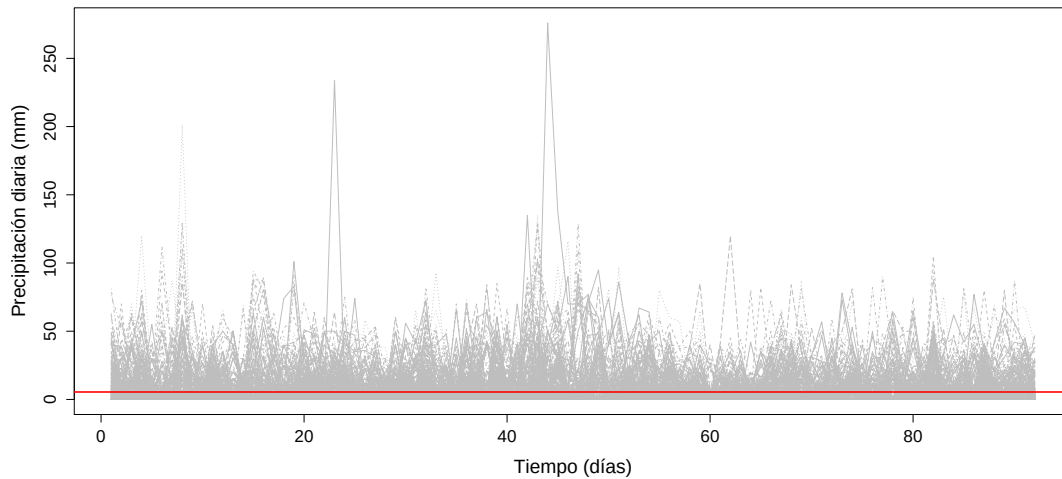


Figura 8.1: Precipitación diaria 2011

Se aprecia un comportamiento aproximadamente cíclico con picos extremos que alcanzan alrededor de los 300 *mm* de lluvia como máximo en lo transcurrido del año. En promedio para este año, la precipitación diaria fue de aproximadamente 5.46 *mm* con una desviación estándar de 11.06 *mm* y un coeficiente de variación de 202.26 %, lo cual es un claro indicio de la alta variabilidad asociada a esta variable, donde se puede pasar de periodos secos a periodos donde la precipitación es alta (Euscátegui & Hurtado, 2011). Posterior a esto, se procedió a construir los pentadales como se especificó en el capítulo anterior (7).

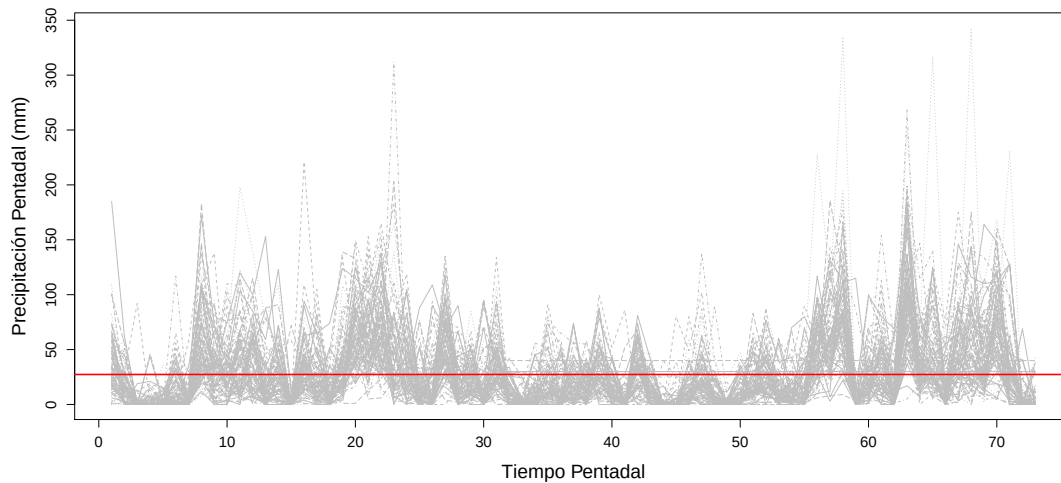


Figura 8.2: Lluvia pentadal 2011.

Analizando la Figura (8.2), para la precipitación pentadal se tiene que durante la gran mayoría del periodo de estudio, ésta tiende a estar por encima de la precipitación promedio para dicho año.

Otro resultado que podría rescatarse posteriormente, es el comportamiento periódico que se presenta en la intensidad de la precipitación pentadal (Figura 8.2); aproximadamente cada tres meses se presenta un patrón estacional que si bien podría ser modelado mediante otro tipo de bases como bases tipo Fourier y no B-Splines, dado el elevado número de bases que esto requiere, se justifica el uso de las últimas (Ramsay & Silverman, 2005).

Respecto a las estaciones, la lluvia pentadal se comporta de la siguiente forma (Figura 8.3)

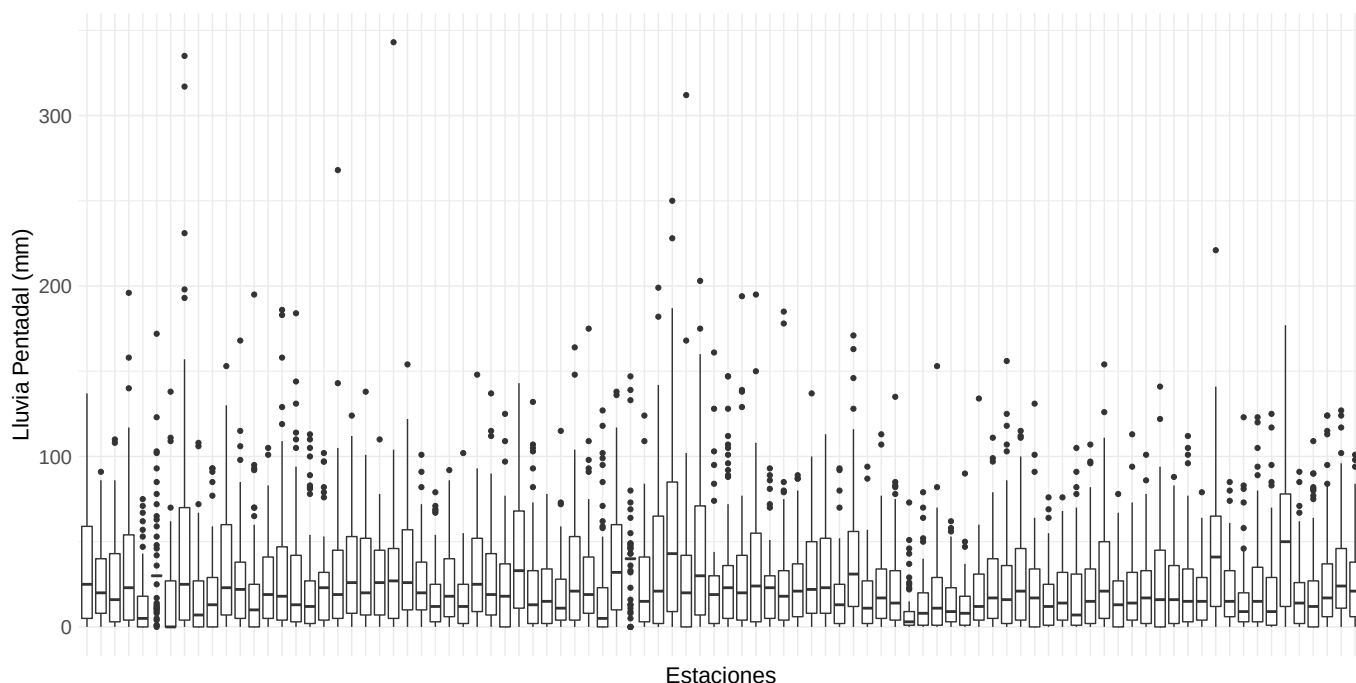


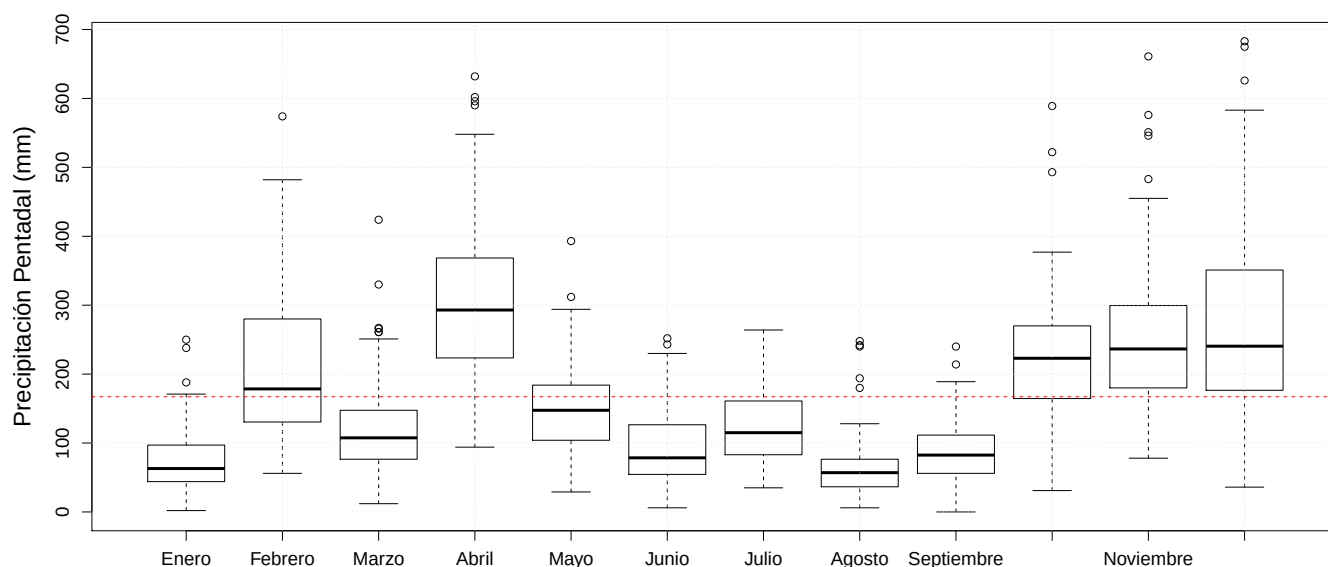
Figura 8.3: Lluvia pentadal por Estación

Se tiene que para el año de estudio, se presentó una máxima precipitación pentadal de 343.00 mm , mientras, gran parte de las estaciones (75 %) no registraron más de 40.00 mm de precipitación pentadal. Se tuvo además, una precipitación pentadal mínima de 0 mm y que la precipitación media pentadal del año 2011 fue de 27.31 mm con una mediana de 17.00 mm . El valor máximo de precipitación pentadal se dio en la estación *Dos quebradas* ubicada en la municipalidad Guadalajara de Buga en el Valle del Cauca, con coordenadas ($3^{\circ}54'31.28'' \text{ N}$, $76^{\circ}10'52.64'' \text{ E}$) a una altura de 1870 msnm y el valor mínimo se dio en la estación *la Balastera* ubicada en la municipalidad de Buenos Aires, Cauca, en los límites con los municipios de Jamundí y Buenaventura en el Valle del Cauca (Alcaldía de Buenos Aires Cauca, 2016) y sus coordenadas son ($3^{\circ}05'19.36'' \text{ N}$, $76^{\circ}35'53.63'' \text{ E}$) con una altura de 1014 msnm . La estación que presentó el mayor coeficiente de variación fue *Ing. la Quinta* con un valor de 164% , la cual posee coordenadas ($3^{\circ}25'02.12'' \text{ N}$, $76^{\circ}20'45.02'' \text{ E}$) y se encuentra ubicada en la municipalidad de Candelaria a una altura de 975 msnm ; de manera análoga la estación cuya precipitación presentó el menor coeficiente de variación corresponde a la estación de El Silencioso ubicada en la municipalidad de Jamundí con coordenadas ($3^{\circ}08'15.12'' \text{ N}$, $76^{\circ}44'01.39'' \text{ E}$) y con una altura de 1502 msnm .

Finalmente, se resalta la bimodalidad de la variable, ya que se observan fuertes crecimientos aproximadamente a principio y a final del año, además de la inestabilidad de la variable, pues esta presentan varias caídas en toda la ventana de observación.

En cuanto al comportamiento mensual para los pentadales se refiere, se tiene entonces (Figura

8.4)

**Figura 8.4:** Precipitación Pentadal Mensual.**Tabla 8.1:** Resumen, Precipitación Pentadal Mensual

	Enero	Febrero	Marzo	Abril	Mayo	Junio
1	Min. : 2.00	Min. : 56.0	Min. : 12.00	Min. : 94.0	Min. : 29.0	Min. : 6.00
2	1st Qu.: 44.50	1st Qu.:131.2	1st Qu.: 76.75	1st Qu.:223.8	1st Qu.:104.5	1st Qu.: 54.75
3	Median : 63.00	Median :178.5	Median :107.50	Median :293.0	Median :147.5	Median : 78.50
4	Mean : 73.77	Mean :209.7	Mean :122.86	Mean :312.6	Mean :148.9	Mean : 95.83
5	3rd Qu.: 96.50	3rd Qu.:279.5	3rd Qu.:147.25	3rd Qu.:368.2	3rd Qu.:184.0	3rd Qu.:126.25
6	Max. :250.00	Max. :574.0	Max. :424.00	Max. :632.0	Max. :393.0	Max. :252.00

	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
1	Min. : 35.0	Min. : 6.00	Min. : 0.00	Min. : 31.0	Min. : 78.0	Min. : 36.0
2	1st Qu.: 84.0	1st Qu.: 36.75	1st Qu.: 56.00	1st Qu.:164.8	1st Qu.:180.0	1st Qu.:177.2
3	Median :115.0	Median : 57.00	Median : 82.50	Median :223.0	Median :236.5	Median :240.5
4	Mean :122.7	Mean : 64.26	Mean : 89.09	Mean :229.3	Mean :259.1	Mean :277.8
5	3rd Qu.:160.0	3rd Qu.: 75.25	3rd Qu.:111.25	3rd Qu.:269.5	3rd Qu.:298.8	3rd Qu.:350.5
6	Max. :264.0	Max. :248.00	Max. :240.00	Max. :589.0	Max. :661.0	Max. :683.0

En la figura (8.4) , es más fácil apreciar la bimodalidad correspondiente a la precipitación pentadal. Esta se da alrededor de los meses de Marzo, Abril y Mayo y en los últimos tres meses del mismo (Octubre, Noviembre, Diciembre). Se tiene de igual forma que el valor máximo de la precipitación se da en Diciembre mientras el mínimo se da en septiembre (Tabla 8.2).

Dicho comportamiento puede deberse al fenómeno de la Niña para el periodo 2010-2011, ya que tal y como lo indica el IDEAM, este tuvo repercusiones en el clima colombiano generando además, catástrofes humanas de diversa índole como inundaciones lentas, crecientes súbitas y deslizamientos de tierras acarreando asimismo pérdidas humanas y materiales (Euscátegui & Hurtado, 2011).

Aún más, al nivel del Valle del Cauca, dicho fenómeno repercutió considerablemente en el comportamiento del río Cauca, el cual “mantuvo niveles críticos desde la cuenca media en adelante, superando cotas de desbordamiento para numerosas poblaciones ubicadas a lo largo de la zona ribereña, que afectaron infraestructuras y ocasionaron muchos damnificados por destrucción de viviendas y vías de acceso” (Euscátegui & Hurtado, 2011).

8.2. Suavización de las Curvas

Como ya se expuso en el capítulo anterior (7), se tuvieron en cuenta 6 funciones base (K) y un parámetro de suavización $\eta = 0.0005$. Inicialmente como se mencionó, el proceso se llevó a cabo tomando en cuenta los datos sin realizar transformación alguna. Así, el resultado de la suavización para los datos pentadales se aprecia en la Figura 8.5

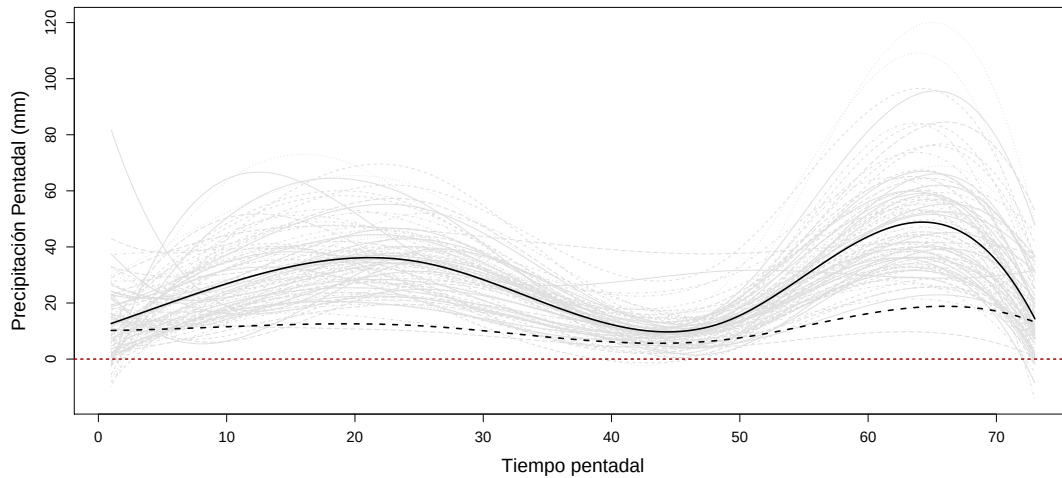


Figura 8.5: Curvas Suavizadas para el año 2011

Donde la curva negra sólida en la Figura 8.5 representa las media funcional, mientras, la punteada representa la desviación estándar funcional.

Lo primero que cabe notar es que las curvas toman valores negativos, los cuales no tienen sentido en este trabajo dado que los datos corresponden a precipitación. Ramsay & Silverman (2005) menciona que este tipo de problemas puede originarse en un número considerable de valores observados iguales a cero.

Para subsanar este problema, como se indicó en el capítulo (7) en la sección (7.4) se propone el método implementado por Giraldo et al. (2011), Reyes et al. (2012) y Marín (2004), que consiste en trabajar con la raíz cuadrada de la precipitación pentadal y elevar los resultados obtenidos al cuadrado; esto garantiza valores positivos para la curva; además, esta transformación no afecta considerablemente la estructura de los datos al ser esta monótona, creciente e inyectiva (Marín, 2004).

Tras aplicar la transformación pertinente, las curvas suavizadas presentan el comportamiento mostrado en la Figura(8.6).

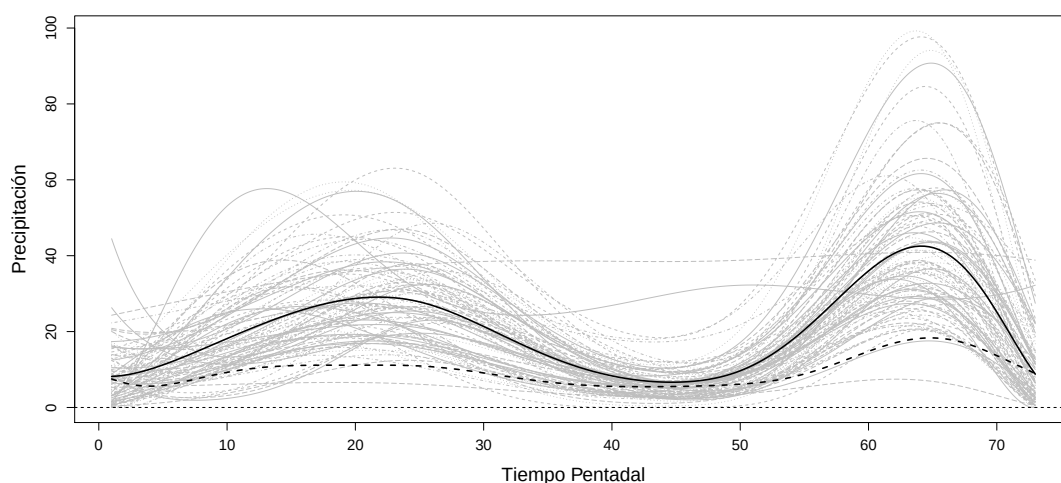


Figura 8.6: Curvas Suavizadas 2011, tras aplicar la transformación

8.3. Análisis Estructural

Posteriormente, para estudiar la estructura de correlación espacial es necesario hallar la traza-variograma empírica en donde se evalúa la distancia promedio entre parejas de sitios dentro de intervalos de distancia; análogo al caso univariado, esta se interpreta de acuerdo al valor que tome; esto es, entre menor sea la distancia entre los sitios mayor similitud o correlación espacial habrá entre las observaciones (Giraldo, 2009).

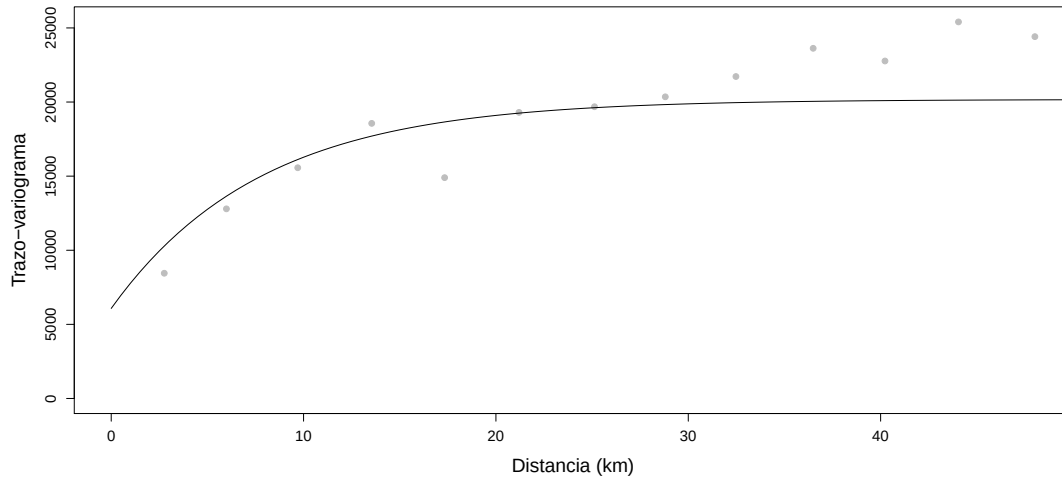


Figura 8.7: Trazo-variograma teórica ajustada

Tras haber graficado la traza-variograma empírica (Figura 8.7) y ajustar sobre esta los modelos teóricos, se halla el modelo que mejor se ajuste a ésta; este paso se muestra en la siguiente figura (Figura 8.8), donde se llega a la conclusión que el modelo que mejor se ajusta es el exponencial. Éste modelo se confunde con el Matérn, dado que conforme al valor que tome κ este modelo representa distintos campos aleatorios con distintos grados de suavidad (Diggle & Ribeiro, 2007).

Por ejemplo, para el caso del modelo exponencial, dicho grado de suavidad se da cuando $\kappa = 0.5$; no obstante y dado que de antemano no se conoce la estructura de la variable de la cual provienen las observaciones es aconsejable usar el modelo exponencial (Diggle & Ribeiro, 2007). Aún más, dado que este modelo se utiliza cuando la dependencia espacial tiene un crecimiento exponencial respecto a la distancia entre las observaciones, su uso se justifica. Cabe resaltar además, que este modelo es el más usado entre las funciones de semivariograma (Giraldo, 2002).

Tabla 8.2: Modelos teóricos

Modelos	τ	σ_q	ϕ	Rango	SCE
Esférico	6076.00	14100.00	7.78	7.78	0.0002
Exponencial	6076.00	14100.00	7.78	23.30	0.0001
Gaussiano	6076.00	14853.34	7.78	13.46	0.0001
Mattern	6076.00	14100.00	7.78	23.30	0.0001

Para un valor $\kappa = 0.5$ el modelo Mattern se aproxima a un modelo exponencial. De acuerdo a un análisis de los parámetros obtenidos anteriormente se seleccionó el modelo exponencial como el modelo que mejor se ajusta a la traza-variograma empírica por las razones antes expuestas. Estos resultados se pueden observar gráficamente (Figura 8.8):

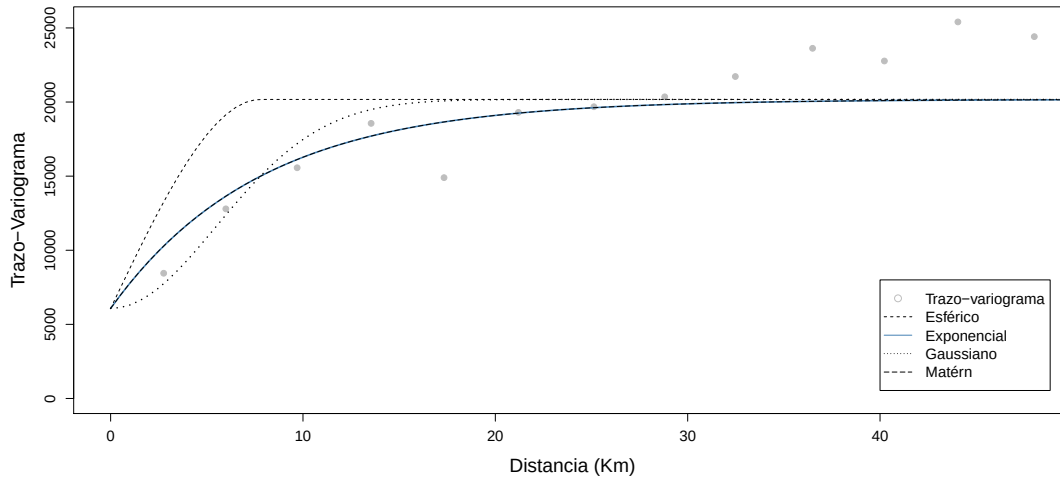


Figura 8.8: Modelos de Traza-variograma teóricos ajustados

Finalmente, el modelo ajustado para la estructura de correlación espacial viene dado por uno exponencial:

$$\hat{\gamma}(h) = 6076.00 + 20176.00 \left(1 - \exp\left(\frac{-3h}{23.30}\right) \right) \quad (8.1)$$

8.4. Kriging Funcional Ordinario Predicción Espacial

Ahora, definido el modelo teórico de la estructura de correlación espacial, se procedió a realizar el proceso de predicción para la precipitación pentadal. Dado que el objetivo del trabajo de grado es realizar predicciones de la precipitación pentadal en un lugar donde no se tenga información, para verificar si las predicciones que se están realizando son acertadas y se aproximan al valor verdadero, se eliminó una a una las 92 estaciones de forma gradual y se predijo su valor correspondiente con la información de las 91 estaciones restantes; se utilizó el modelo de Kriging Funcional Ordinario para predecir los valores correspondientes a las curvas de precipitación en dichas estaciones. Con el fin de ilustrar el procedimiento, se analizarán 30 de las 92 estaciones estudiadas.

Tabla 8.3: Estaciones escogidas para ilustrar

SAN PABLO	EL PARAISO
EL BOSQUE	VILLA ARACELLY
EL SOCORRO	EL BALSAL
EL PORVENIR	LA PATAGONIA
PUERTO NUEVO	EL VESUBIO
GUACHAZAMBOLO	YANACONAS
EL PALACIO	MORAVITO
EL RETIRO	BUENOS AIRES
LA ARBOLEDA	MONTECRISTO
BOLO BLANCO	AGUACATAL
DAGUA	EL OSO
ACUEDUCTO BUGA	MATAPALO
VILLACOLOMBIA	ALTO DE LAS IGLESIAS
LLOREDA GRASAS	BUGLAGRANDE

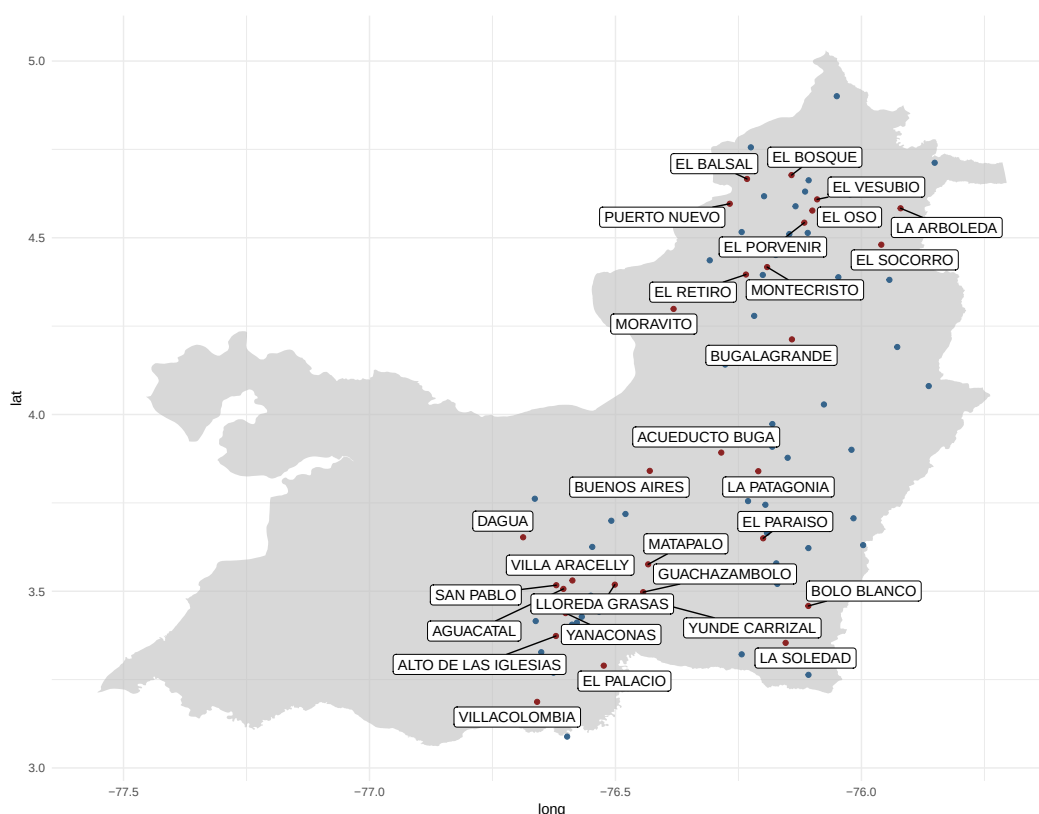
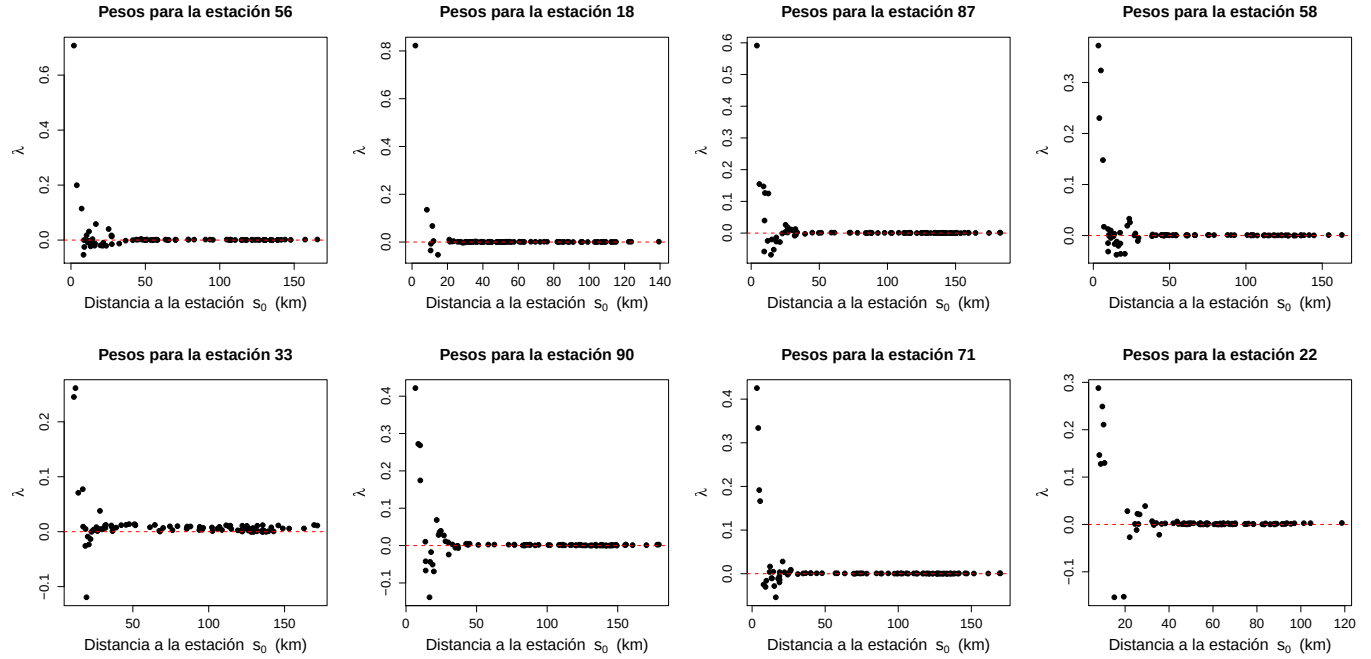
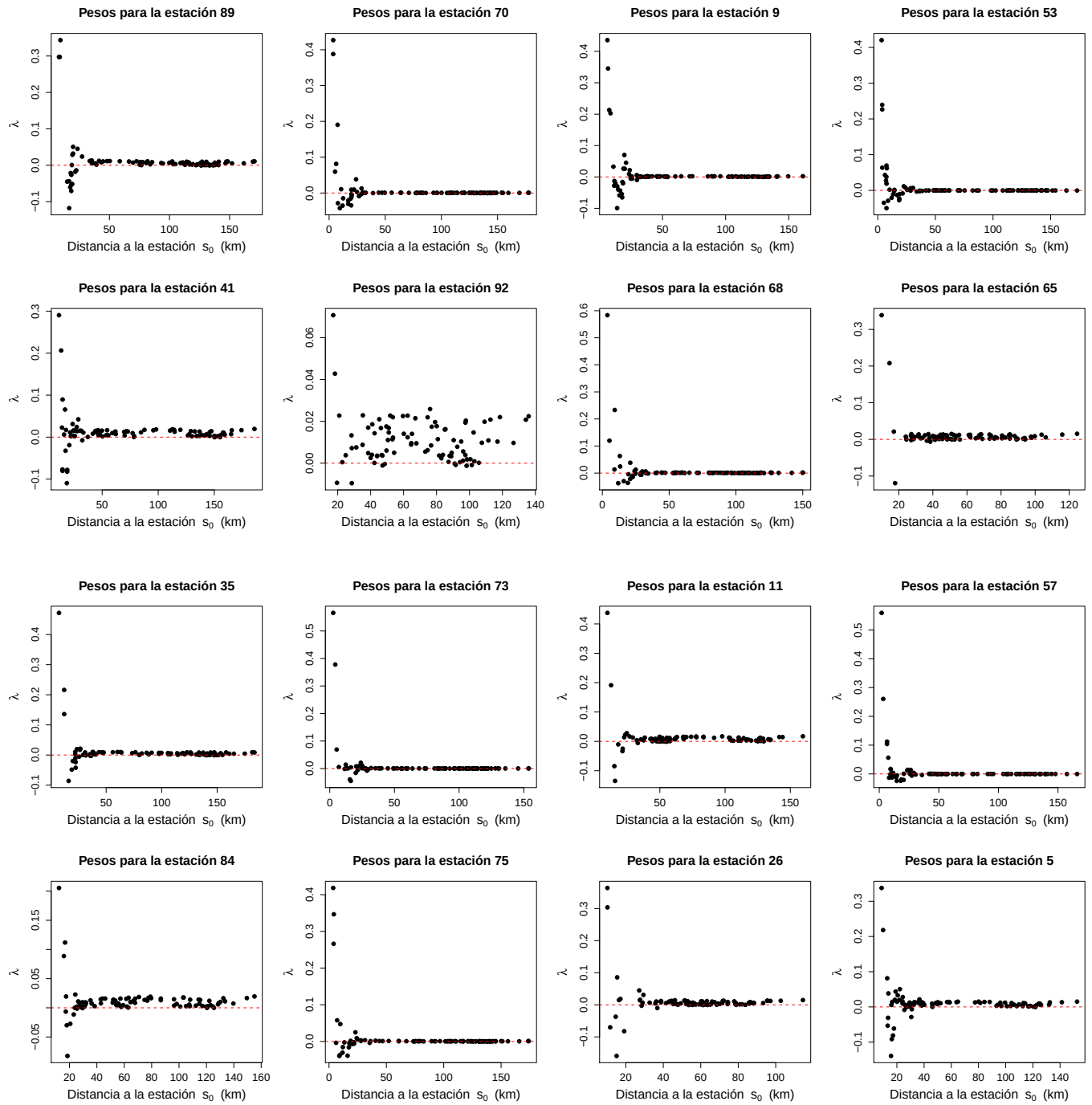


Figura 8.9: Estaciones escogidas

Es notable aún en la figura (8.9), la ausencia de un número de estaciones considerables hacia la región central del valle, lo cual influenció las predicciones realizadas para las estaciones en esta zona. Por ejemplo, en el caso de las estaciones, “*Buenos Aires*”, “*Moravito*”, y “*Acueducto Buga*” las cuales no cuentan con puntos de muestreo lo suficientemente cercanos como se muestra en (8.9).

Para cada una de las 30 estaciones que ilustran los resultados, los pesos $\lambda_i, i = 1, 2, \dots, 91$ asociados a las estaciones restantes obtenidos mediante Kriging Funcional Ordinario para su predicción, presentan el siguiente comportamiento.





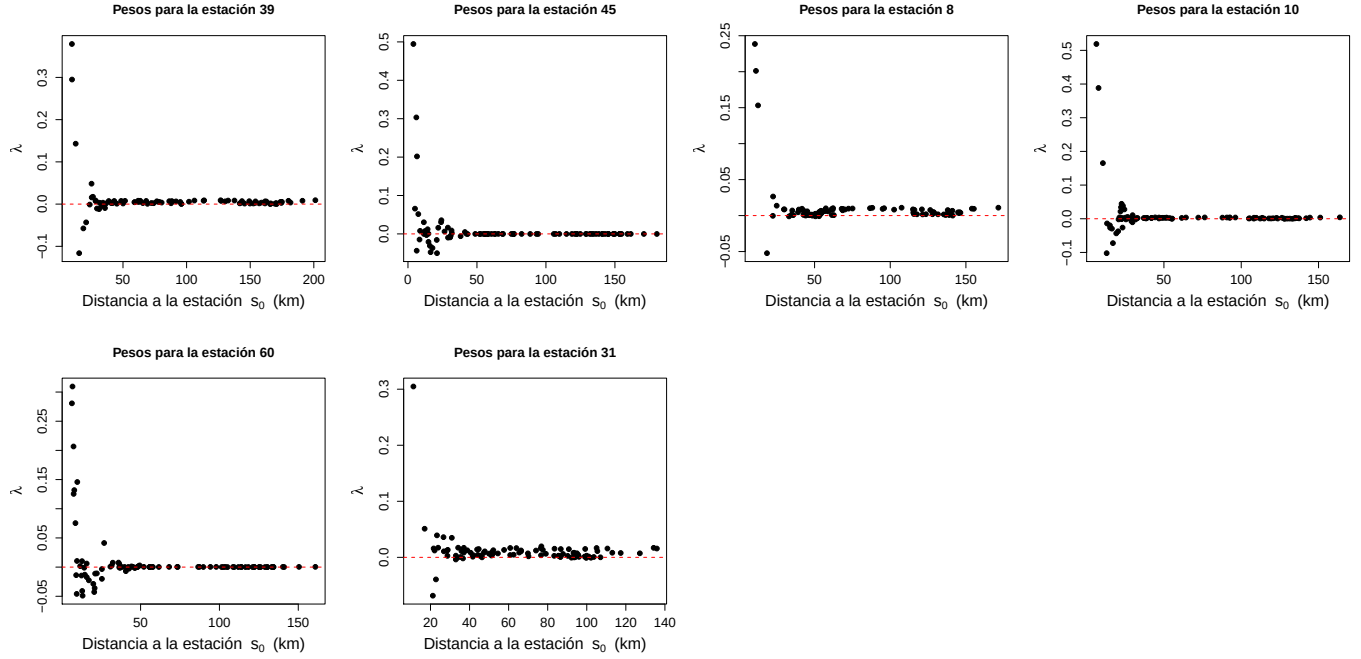


Figura 8.10: Pesos asociados al predictor OKFD

Para los pesos asignados a cada una de las estaciones escogidas para ilustrar el procedimiento, se tiene que la gran mayoría tuvo valores cercanos a cero. Por ejemplo, el primer cuartil de la distribución de estos toma el valor de cero y el 50 % toma valores menores o iguales a 0.001.

En la figura (8.10), se aprecia para la estación de *Buenos Aires* (estación número 65) que la mayoría de los pesos asociados tiene valores cercanos a cero, lo cual corrobora el supuesto de mayor ponderación a puntos cercanos; esto mismo se corrobora para la estación *Acueducto de Buga* (estación número 26) y la estación *Moravito* (estación número 92).

Ahora, para el caso de las 92 estaciones, superponiendo las curvas predichas a las observadas, el comportamiento es el siguiente (Figura 8.11).

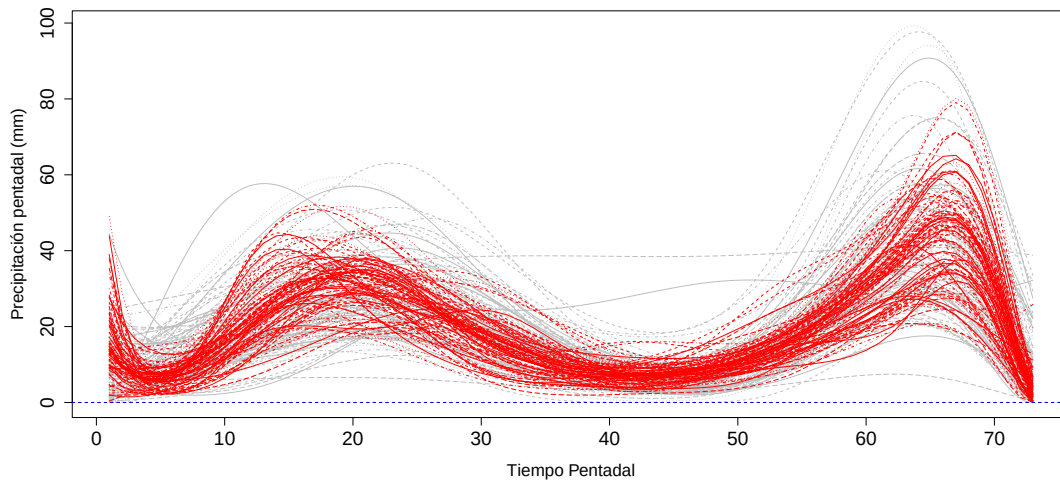


Figura 8.11: Predicción de curvas pentadales para las 92 estaciones

Es notable que las curvas obtenidas de este proceso siguen el comportamiento global de la precipitación pentadal en el Valle del Cauca, aunque las curvas predichas tienden a subestimar las observadas; aún más, dado que no es visible si el comportamiento pronosticado es acertado y/o preciso del todo y se busca estudiar los residuales de dichas predicciones para tener mayor información sobre más propiedades del modelo de predicción funcional se graficaron los valores observados vs. los predichos y los predichos vs. los residuales. El comportamiento esperado en el gráfico de datos observados y predichos es lineal, y para el gráfico de residuales y predicciones se espera no encontrar algún tipo de patrón o comportamiento, lo que describiría un proceso homocedástico.

A continuación, se realiza un gráfico de residuales, en el cual se superpondrán la media y variación de esta.

8.4.1. Análisis de Residuales

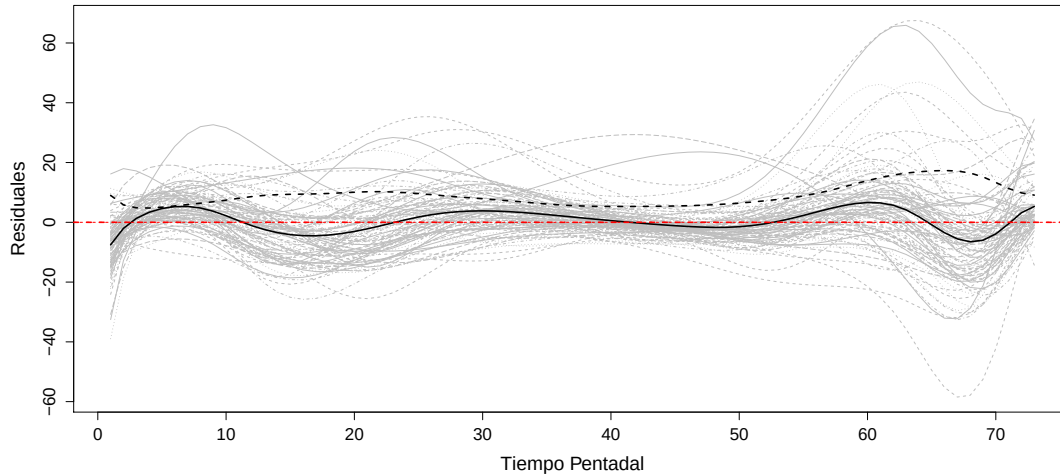


Figura 8.12: Residuales de curvas pentadales para las 92 estaciones

En la Figura (8.12), la línea discontinua hace referencia a la desviación estándar de los residuales y la línea negra sólida representa la media de los mismos.

Para este caso, se valida el supuesto de media cero, conforme a lo expuesto por Giraldo et al. (2011) garantizando insesgadez en las predicciones. Respecto al comportamiento de la desviación estándar y algunas curvas residuales, se aprecia un comportamiento creciente respecto a los valores finales de t . Este comportamiento puede atribuirse a los altos valores que la variable tomó durante los meses Marzo, Abril y Mayo y Octubre, Noviembre y Diciembre durante el fenómeno de La Niña acaecido entre los años 2010 y 2011 (Euscátegui & Hurtado, 2011). De igual manera, se debe a la naturaleza bimodal (Reyes et al., 2012) y el proceso acumulativo de la variable original (precipitación diaria en precipitación pentadal), tal y

como lo expresa Ospina (2016).

Para estudiar más en detalle el ajuste de los valores predichos a los observados, se procedió a graficar para cada estación los valores obtenidos mediante el modelo, en función de los valores observados. Es de esperarse que el ajuste sea aproximadamente lineal acercándose su forma a una recta $x = y$.

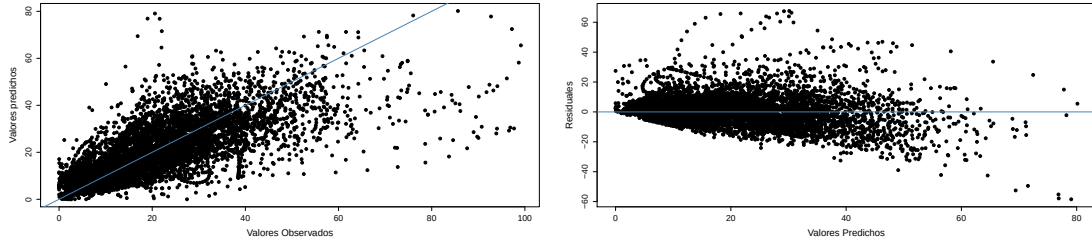


Figura 8.13: Análisis de los residuales

Se procedió entonces a juzgar la homocedasticidad de los errores mediante un gráfico de los residuales vs. los valores ajustados (Panel Derecho de la Figura 8.13). Se evidencia un comportamiento creciente en la última parte de la figura, producto del proceso acumulativo llevado a cabo para la construcción de los datos pentadales como también lo mostró Ospina (2016). También puede deberse a la naturaleza bimodal de la variable precipitación (Reyes et al., 2012) y a los altos valores observados durante el periodo de análisis (Euscátegui & Hurtado, 2011).

Posteriormente, de acuerdo al panel izquierdo de la Figura (8.13) para los valores predichos y observados, es posible evidenciar que la relación de estos es aproximadamente lineal, además, teniendo en cuenta el análisis de las curvas residuales donde la media fluctuó alrededor de cero, se concluye que el modelo logró predecir la precipitación pentadal.

Así, de los resultados mostrados en las Figuras (8.11), (8.12) y (8.13) se concluye que el modelo logró predecir la variable precipitación, lo cual constituye una alternativa metodológica a aquellas expuestas en el capítulo correspondiente a los antecedentes (Capítulo 4).

Capítulo 9

Conclusiones y Recomendaciones

9.1. Conclusiones

El estudio de datos espaciales se ha hecho necesario en múltiples áreas del conocimiento como la geografía, geología, ciencias ambientales, entre otras. Estos datos además de contener información de la variable de interés en el estudio, se complementan con información geográfica de la ubicación en donde estos se dan. Las técnicas clásicas suponen que las variables están sujetas a condiciones idénticas además de que dichas observaciones son independientes entre sí sin considerar el impacto del tiempo y el espacio, es por esto que, al estudiar datos de carácter espacial se espera que las observaciones cercanas sean correlacionadas entre ellas por lo cual es importante estudiar dicha estructura (Ramsay & Silverman, 2002, 2005, Giraldo et al., 2011).

De acuerdo con las estadísticas descriptivas, se encuentra que existe bimodalidad para la variable precipitación, donde se puede observar que los meses donde se dieron las mayores cantidades de precipitación fueron el periodo Febrero a Abril y de Octubre a Diciembre. También se destaca la forma que toma la precipitación pentadal cuya densidad es aproximadamente la de una distribución Gamma; esta conclusión ha sido ampliamente divulgada en diferentes investigaciones (Cruz & Males, 2012, Portillo & Medina, 2012) siendo importante resaltar esta característica, dado que la transformación raíz cuadrada que se utilizó en este estudio es recomendada para variables cuya distribución es asimétrica positiva (Marín, 2004).

La suavización realizada por el método no paramétrico B-splines con los datos originales generaban algunos problemas teóricos de gran relevancia, pues se calcularon curvas de precipitación funcional menores a cero, lo cual dado el contexto del problema no tiene sentido; así, se justifica el uso de una transformación para los datos previo al proceso de generación de las curvas funcionales. Por consiguiente, el utilizar la raíz cuadrada de la precipitación para realizar el suavizado de las curvas y luego elevar las curvas obtenidas al cuadrado, para

obtener la escala original de la variable (Giraldo et al., 2011), (Reyes et al., 2012) & (Marín, 2004), resolvió dicho inconveniente en este trabajo.

Respecto a la estructura de correlación espacial, la traza-variograma empírica fue ajustada al modelo teórico exponencial, es decir que la dependencia espacial tiene un crecimiento exponencial respecto a la distancia de las observaciones, adicionalmente, se conoce que este modelo es utilizado ampliamente según Giraldo (2002).

Se obtuvieron adicionalmente resultados similares a los de Ospina (2016) donde los residuales producto del proceso de predicción, presentaron cierta variabilidad. Como lo indica Ospina (2016) producto del proceso acumulativo de la variable original; también puede deberse a la naturaleza bimodal de la variable precipitación (Reyes et al., 2012) y a los altos valores de precipitación que se presentaron durante el periodo 2010-2011 debido al fenómeno de la niña (Euscátegui & Hurtado, 2011); otra alternativa para modelar esta variable se encuentra dada por (Reyes et al., 2012). Finalmente con base a lo expuesto en el Capítulo 8 se concluye, que el modelo logró predecir la variable precipitación pentadal.

9.2. Recomendaciones

Se pueden considerar otro tipo de bases para llevar a cabo el proceso de suavización no paramétrica como lo son las bases Wavelets si se está en presencia de datos faltantes, discontinuidades en las mediciones y de perturbaciones aleatorias; este tipo de bases representarían una alternativa que se ha comprobado, genera resultados similares a aquellos hallados mediante B-splines para el proceso predictivo (Giraldo et al., 2011).

Otra alternativa para tratar con el problema de suavizaciones negativas, es recurrir a técnicas de suavizaciones restringidas como lo son la suavización positiva y la monótona (Ramsay & Silverman, 2005, Ospina, 2016).

Por último, cabe mencionar que para la predicción de la precipitación pentadal se pueden probar otros tipos de predictores funcionales como el Universal, el Residual o el Kriging con Deriva Externa, teniendo en cuenta las condiciones de la variable, en caso de que esta cuente con una tendencia en media o variabilidades producto del proceso acumulativo.

Bibliografía

- Agud, R. (2011). *Nuevos modelos de probabilidad de la precipitación: Aplicación a España, y en particular, al País Vasco*. PhD thesis, Universitat de València, España.
- Alcaldía de Buenos Aires Cauca (2016). Alcaldía de Buenos Aires Cauca: Nuestro Municipio. Recuperado de http://buenosaires-cauca.gov.co/informacion_general.shtml#geografia.
- Alcaldía de Santiago de Cali (2016). Datos de Cali y el Valle del Cauca. Recuperado de http://www.cali.gov.co/publicaciones/227/datos_de_cali_y_el_valle_del_cauca/.
- Bedoya, M., Contreras, C., & Ruiz, F. (2010). Estudio del agua 2010. capítulo 7: Alteraciones del régimen hidrológico y la oferta hídrica por variabilidad y cambio climático. Recuperado de <http://documentacion.ideam.gov.co/openbiblio/bvirtual/021888/CAP7.pdf>.
- Bel, L., Bar-Hen, A., Petit, R., & Cheddadi, R. (2011). Spatio-temporal functional regression on paleoecological data. *Journal of Applied Statistics*, 38(4), 695–704.
- Belton, A. B. R. (2006). *Functional Analysis*. Lancaster: Lancaster University, 2 edition.
- Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied spatial data analysis with R*. New York: Springer, 1 edition.
- Castro, D. & Carvajal, Y. (2013). Análisis de tendencia en la precipitación pluvial anual y mensual en el departamento del valle del cauca. *Memorias*, 11(20).
- Chilès, J.-P. & Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. New Jersey, USA: John Wiley & Sons, 2 edition.
- Collado, J. (1988). Estimación óptima de la precipitación media con el método kriging. *Ingeniería hidráulica en México*, 3, 34–45.
- Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley-Interscience, 1 edition.
- Cruz, N. & Males, V. (2012). Modelos mixtos lineales generalizados para la modelación de la precipitación mensual en el Valle del Cauca - Tesis de pregrado en Estadística. Universidad del Valle.

- Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147, 1 – 23.
- DANE (2013). Cuenta satélite piloto de la agroindustria (cspa): Maíz, sorgo y soya y su primer nivel de transformación 2005-2009. https://www.dane.gov.co/files/investigaciones/pib/agroindustria/Doc_Metodologico_%20Maiz_Sorgo_Soya_def_23_05_13.pdf.
- Dauxois, J., Pousse, A., & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1), 136 – 154.
- Delgado, E. (2012). Análisis estadístico mediante modelos de efectos mixtos funcionales. Master's thesis, Universidad de Granada.
- Delicado, P., , R., Comas, C., & Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21, 224–239.
- Delicado, P. (2008). Curso de modelos no paramétricos. *Departamento de Estadística Investigación Operativa, Universidad Politécnica de Cataluña*. Recuperado de http://www-eio.upc.es/~delicado/docencia/Apuntes_Models_No_Parametrics.pdf.
- Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'inséé*, (15), 3–101.
- Di Giuseppe, E. (2013). *Statistical methods for the analysis of climatic phenomena*. PhD thesis, University of Rome La Sapienza.
- Diggle, P. & Ribeiro, P. (2007). *Model-based Geostatistics*. Springer Series in Statistics. Springer New York, 1 edition.
- Dunn, P. K. (2004). Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, 24(10), 1231–1239.
- Escobar, S. (2006). Aplicación de técnicas estadísticas en las series climatológicas mensuales totales de precipitación, evaporación y brillo solar. Corporación Autónoma Regional del Valle del Cauca (CVC). http://190.97.204.39/cvc/RecursoHidrico/aplicativos/Climatologia/Documentos/documento_final_estadistico.pdf.
- Euscátegui, C. & Hurtado, G. (2011). Análisis del impacto del fenómeno de la niña 2010-2011 en la hidroclimatología del país. Recuperado de <http://www.ideam.gov.co/documents/21021/418818/Análisis+Impacto+La+Niña.pdf/640a4a18-4a2a-4a25-b7d5-b3768e0a768a>.

- Ferraty, F. & Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. New York: Springer, 1 edition.
- Giraldo, R. (2002). *Introducción a la geoestadística: Teoría y aplicación*. 1 edition.
- Giraldo, R. (2009). *Geostatistical analysis of functional data*. PhD thesis, Universitat Politècnica de Catalunya.
- Giraldo, R., Delicado, P., & Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, 18(3), 411–426.
- Giraldo, R., Mateu, J., & Delicado, P. (2012). geofd: an R package for function-valued geostatistical prediction. *Revista Colombiana de Estadística*, 35(3), 385–407.
- Goovaerts, P. (1999). Using elevation to aid the geostatistical mapping of rainfall erosivity. *CATENA*, 34(3), 227 – 242.
- Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(1), 113 – 129.
- Goulard, M. & Voltz, M. (1993). *Geostatistical Interpolation of Curves: A Case Study in Soil Science*, (pp. 805–816). Springer Netherlands: Dordrecht.
- Guenni, L., Degryze, E., & Alvarado, K. (2008). Análisis de la tendencia y la estacionalidad de la precipitación mensual en venezuela. *Revista Colombiana de Estadística*, 31(1), 41–65.
- Hasan, M. M. & Dunn, P. K. (2010). A simple poisson–gamma model for modelling rainfall occurrence and amount simultaneously. *Agricultural and Forest Meteorology*, 150(10), 1319 – 1330.
- Horváth, L. & Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer New York, 1 edition.
- Hsing, T. & Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 1 edition.
- IGAC (2012). Instituto Geográfico Agustín Codazzi (IGAC). http://geoportal.igac.gov.co/mapas_de_colombia/igac/mps_fisicos_deptales/2012/Valle_del_Cauca.pdf.
- Isaaks, E. H. & Srivastava, R. M. (1989). *An introduction to applied geostatistics*. New York: Oxford University Press, 2 edition.

- Jaramillo, A. & Chaves, B. (2000). Distribución de la precipitación en Colombia analizada mediante conglomeración estadística. *Centro Nacional de Investigaciones de Café*, (2), 102–113.
- Jewell, S. A. & Gaussiat, N. (2015). An assessment of kriging-based rain-gauge–radar merging techniques. *Quarterly Journal of the Royal Meteorological Society*, 141(691), 2300–2313.
- Kokoszka, P. & Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman & Hall / CRC numerical analysis and scientific computing. CRC Press, 1 edition.
- Levitin, D., Nuzzo, R. L., Wines, B. W., & Ramsay, J. O. (2007). Introduction to functional data analysis. *Canadian Psychology*, 48(3), 135–155.
- Ly, S., Charles, C., & Degré, A. (2011). Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium. *Hydrology and Earth System Sciences*, 15(7), 2259–2274.
- Marín, J. (2004). *Transformaciones de variables: Notas de clase*. Universidad Carlos III de Madrid.
- Matheron, G. (1969). *Curso de Geoestadística*. París, France: Escuela de Minas de París.
- Melo, C. (2012). *Análisis Geoestadístico Espacio Tiempo basado en Distancias y Splines con Aplicaciones*. PhD thesis, Universitat de Barcelona.
- Montero, J., Fernández-Avilés, G., & Mateu, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Wiley Series in Probability and Statistics. Wiley, 1 edition.
- Navarro, V. (2004). Análisis de datos funcionales: Implementación y aplicaciones. Trabajo de Fin de Carrera. Recuperado de <https://upcommons.upc.edu/handle/2099.1/13315?locale-attribute=es>.
- Olaya, J. (2012). *Métodos de regresión no paramétrica*. Cali, Colombia: Editorial: Universidad del Valle, 1 edition.
- Ospina, J. (2016). Modelo de regresión funcional de la precipitación pentadal, integrando datos de estaciones climáticas y de satélite. Master's thesis, Universidad del Valle.
- Parzen, E. (1999). *Stochastic Processes*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1 edition.
- Piqueras, M. (2007). Sector agrario y cambio climático. *Ecologistas en Acción*, 54, 49–58.

- Portillo, R. A. & Medina, C. (2012). Modelo Lineal Mixto Generalizado para la Precipitación Mensual en el Departamento del Valle del Cauca en los Periodos de 1971 a 2009 - Tesis de pregrado. Universidad del Valle.
- Ramsay, J. & Silverman, B. (2005). *Functional data analysis*. New York: Springer, 1 edition.
- Ramsay, J. O. & Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. New York, USA: Springer. 191p.
- Reyes, A., Giraldo, R., & Mateu, J. (2012). Residual kriging for functional data. application to the spatial prediction of salinity curves. Reporte Interno de Investigación no. 20. Departamento de Estadística. Facultad de Ciencias. Universidad Nacional de Colombia. http://www.estadistica.unal.edu.co/unciencias/data-file/user_45/RI20.pdf.
- Rodriguez, R., Benito, A., & Portela, A. (2004). *Meteorología y climatología*. España: Fundación española para la ciencia y la tecnología, 1 edition.
- Schlegel, F., de las Naciones Unidas para la Agricultura y la Alimentación, O., Chiappini, A., & FAO (1996). *Ecología y enseñanza rural: nociones ambientales básicas para profesores rurales y extensionistas*. Estudio FAO montes. FAO.
- Secretaría de la OMM (1993). *Atlas Internacional de Nubes: Manual de Observación de Nubes y Otros Meteoros*. Organización Meteorológica Mundial, Ginebra, Suiza.
- Secretaría Departamental de Salud del Valle del Cauca (2016). Análisis de situación de salud con el modelo de los determinantes sociales. Recuperado de <http://www.valledelcauca.gov.co/salud/descargar.php?id=13189>.
- Serban, N. (2008). *Spatial Statistics: Lecture Notes*. Georgia Tech.
- SIPSA (2014). *Principales afectaciones, enfermedades y plagas de los cítricos durante la temporada seca*. DANE, Bogotá, 27 edition. https://www.dane.gov.co/files/investigaciones/agropecuario/sipsa/insumos_factores_de_produccion_sep_2014.pdf.
- Sánchez, F. J. (2017). *Hidrología Superficial y Subterránea*. Salamanca: Createspace Independent Publishing Platform, 1 edition.
- Solera, A. (2000). *Criterios para la Selección de Modelo Estadísticos*. Banco Central de Costa Rica, Costa Rica.
- STAPH (2017). Groupe de travail en statistique fonctionnelle et opératoirelle. www.math.univ-toulouse.fr/staph/.
- Stern, R. D. & Coe, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, 147(1), 1–34.

- Suhaila, J., Jemain, A. A., Hamdan, M. F., & Wan Zin, W. Z. (2011). Comparing rainfall patterns between regions in Peninsular Malaysia via a functional data analysis technique. *Journal of Hydrology*, 411(3), 197–206.
- Suhaila, J. & Yusop, Z. (2016). Spatial and temporal variabilities of rainfall data using functional data analysis. *Theoretical and Applied Climatology*, 129(1), 229–242.
- Tomaso Poggio, L. R. & Frogner, C. (2010). Statistical learning theory and applications, spring 2010. Module: Functional analysis. Recuperado de <http://www.mit.edu/~9.520/spring10/Classes/mathcamp01.pdf>.
- Valle del Cauca: Colombia, guía Turística (2017). Análisis de situación de salud con el modelo de los determinantes sociales. Recuperado de <https://cdn.colombia.com/docs/turismo/sitios-turisticos/cali/valle-del-cauca.pdf>.
- Verdin, A., Rajagopalan, B., Kleiber, W., & Funk, C. (2015). A bayesian kriging approach for blending satellite and ground precipitation observations. *Water Resources Research*, 51(2), 908–921.
- Verdin, A., Rajagopalan, B., Kleiber, W., & Funk, C. (2016). Kriging and local polynomial methods for blending satellite-derived and gauge precipitation estimates to support hydrologic early warning systems. *IEEE Transactions on Geoscience and Remote Sensing*, 54(5), 2552–2562.
- Weather & Climate (2017). Atmospheric layers. <http://www.weather-climate.org.uk/02.php>.
- Webster, R. & Oliver, M. (2007). *Geostatistics for Environmental Scientists*. Statistics in Practice. Wiley.
- Yu, J.-Y. (2011). *Weather Analysis: Course Notes*. University of California, Irvine.
- Zhang, J. (2013). *Analysis of Variance for Functional Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 1 edition.

Capítulo 10

Anexos

10.1. Código en R para la suavización de los Datos

```
# ----- #
# ----- Cargando librerias ----- #
# ----- #

libs <- c("sp", "raster", "rasterVis", "maptools", "rgeos", "XML", "rgdal",
         "dismo", "rworldmap", "mapview", "geofd", "gtools",
         "RgoogleMaps", "ggplot2", "readr", "readxl",
         "fda", "fda.usc", "xtable", "moments", "colorspace",
         "scatterplot3d", "reshape", "plotly", "calibrate")

lapply(libs, library, character.only = TRUE)

coord <- read_delim("DataFrameLonLat92EstacionesPlanimetricas.csv", ";",
  escape_double = FALSE, trim_ws = TRUE)
attach(coord)

d <- read.csv("Datos/DatosEstacionesDiariasCompletas_2011.csv")
View(d)

# ----- #
# ----- Dato pentadal ----- #
# ----- #

datos <- merge(d, coord)
di <- as.matrix(datos[, -c(1:12)])
dias <- di[, -c(366:368)]
```

```

nom <- datos$STATION_NAME;nom
rownames(dias) <- nom

n <- 5
pent <- matrix(0, nrow = nrow(dias), ncol=73)

for(i in 1:nrow(dias)){
  pent[i,] <- unname(tapply(dias[i,], (seq_along(dias[i,])-1) %/% n, sum))
}
View(pent)

pent <- as.data.frame(pent)
rownames(pent) <- nom

penta <- t(pent)
colnames(penta) <- nom

# ----- #
# ----- Suavización ----- #
# ----- #

time <- seq(1, 73, length.out = 73)
dayrng <- c(1,73)
fd_nombres_penta = list("Datos Funcionales Precipitación",
  "Pentadal", "Precipitación")

penta.f <- fdata(penta,argvals = time, rangeval = range(c(0,1)),
  names = fd_nombres_penta )
lambda <- 5*(1e-4)
nodos <- seq(1,73,length.out = 4)
norder <- 4
nbasis <- 6
bbasis <- create.bspline.basis(rangeval = dayrng, nbasis = nbasis,
norder, breaks =nodos)

x11()
plot(bbasis, lwd = 2, cex.axis = 1.5)
curv.Lfd <- int2Lfd(2)
harmLfd <- vec2Lfd(c(0,(2*pi/73)^2,0), rangeval = dayrng)
curv.fdPar <- fdPar(bbasis,curv.Lfd, lambda)

```

```
pentSmooth1 <- smooth.basis(time, penta, curv.fdPar)
pentfd <- pentSmooth1$fd
```

```
penta.r <- penta^0.5
pentSmooth2 <- smooth.basis(time,penta.r,curv.fdPar)
```

```
pentfd.r <- pentSmooth2$fd^2
```

10.2. Código en R para el Análisis Estructural

```
# ----- #
# ----- Análisis Estructural ----- #
# ----- #

n <- dim(penta)[1]
argvals<-seq(1,n, by=1)

coord.11 <- as.matrix(coor[,c(5,6)])
dista <- dist(coor, method="euclidean", diag=TRUE, upper=TRUE)
d <- as.matrix(dista)
dist <- max(d)

s <- 92
rangeval <- range(argvals)
M <- bsplinepen(bbasis,0)
L2norm <- l2.norm(s,pentSmooth1$fd, M)

dista <- max(dist(coord.11))

tracev <- trace.variog(coord.11, L2norm, bin=FALSE, max.dist=50,uvec="default",breaks="def
tracevbin <- trace.variog(coord.11, L2norm, bin=TRUE, max.dist=50)

models <- fit.tracevariog(tracevbin, models = c("spherical","exponential",
  "gaussian","matern"),
                                sigma2.0 = 14100, phi.0 = 7.78,  nugget = 6076,
                                fix.kappa = TRUE,
                                kappa = 0.5, max.dist.variogram = 50)

plot(tracev)
x11()
plot(tracevbin, xlab="Distancia (Km)", ylab="Trazo-Variograma",
```

```

pch = 19, col = "grey", cex.lab = 1.5, cex.axis = 1.2)
lines(models$fitted[[1]], lty = 2, col = "black", lwd = 1)
lines(models$fitted[[2]], lwd=2, col="steelblue", lty = 1)
lines(models$fitted[[3]], lwd=2, col="black", lty = 3)
lines(models$fitted[[4]], lwd=1, col="black", lty = 5)
legend(40, 8000, c("Trazo-variograma", "Esférico",
  "Exponencial", "Gaussiano", "Matérn"),
  lty=c(-1,2,1,3,5),col=c("gray","black","steelblue","black","black"),
  pch=c(1,-1,-1,-1,-1), bty = n, cex = 1.2)

# Mejor Modelo
modelot <- models$best

x11()
plot(tracevbin, pch = 19, col = "grey",
  xlab = "Distancia (km)", ylab = "Trazo-variograma", cex.lab = 1.5, cex.axis = 1.2)
lines(modelot)

```

10.3. Código en R para la Validación del Modelo

```

okfd.pred <- matrix(0, nrow = 73, ncol = 92)

for (i in 1:92) {
  coord.cero <- matrix(c(coord.11[i,], coord.11[i, ]), ncol = 2, nrow = 1)
  coord.comp <- coord.11[-i,]
  penta.comp <- penta.r[, -i]
  okfd.res <- okfd(new.coords = coord.cero,
    coords=coord.comp, data=penta.comp,
    smooth.type = "bsplines", nbasis = 10,
    argvals = time, fix.nugget = TRUE, lambda = 0.0005)
  nd <- okfd.res$krig.new.data
  okfd.pred[,i] <- nd
}

observados <- eval.fd(pentfd.r,time)
residuales <- observados - okfd.pred^2
medres.fd <- apply(residuales, 1, mean)
sdres.fd <- apply(residuales, 1, sd)

```