



Modelo de Regresión Funcional de la precipitación pentadal, integrando datos de estaciones climáticas y de satélite

Johann Alexis Ospina Galindéz

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2016

Modelo de Regresión Funcional de la precipitación pentadal, integrando datos de estaciones climáticas y de satélite

Johann Alexis Ospina Galindéz

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magister en Estadística

Director(a):

Mercedes Andrade Bejarano, Ph.D.

Codirector(a):

Ramón Giraldo Henao, Ph.D.

Línea de Investigación:

Regresión y Predicción Grupo de Investigación en Estadística Aplicada INFERIR

Universidad del Valle

Facultad de Ingeniería, Escuela de Estadística

Santiago de Cali, Colombia

2016

A ...

Mi madre Ana Rosa

Y

Martha Lucía

Le dedico todos mis objetivos alcanzados.

Agradecimientos

En primer lugar agradezco a mi señora madre por estar siempre a mi lado en los momentos difíciles y a Martha Lucía por darme la oportunidad de estudiar.

Agradezco a la Universidad del Valle, en especial a la Escuela de Estadística por brindarme la oportunidad de hacer el postgrado que siempre había anhelado. A la profesora Mercedes Andrade por sus enseñanzas en el transcurso de la maestría y orientación en el trabajo de grado, y al profesor Ramón Giraldo por toda su disposición para el desarrollo de este trabajo de investigación.

Finalmente agradezco a U.S. Geological Survey EROS Center/FEWS NET y la Universidad de California Santa Barbara, en especial a los investigadores Diego Pedreros y Chris Funk por sus valiosas sugerencias y por haberme suministrado los datos de satélite para el desarrollo del trabajo de investigación.

Resumen

El conocimiento del comportamiento de las variables climáticas y su modelación ayudan a predecir, pronosticar, monitorear el cambio climático y a caracterizar los fenómenos naturales relacionados con el clima que afectan el medio ambiente y los recursos hídricos que determinan la agricultura en una región. Es por ello que la modelación de variables climáticas, es uno de los objetivos del Grupo de Investigación en Estadística Aplicada INFERIR.

El presente proyecto de investigación está enmarcado dentro de la línea de investigación en Regresión y Predicción del grupo INFERIR y tiene como meta modelar la precipitación pentadal en el Valle del Cauca en el período 1993-2011, con el fin de generar información que sirva de insumo para la toma de decisiones de los agricultores, centros de investigación, corporaciones autónomas regionales e instituciones de salud, entre otros. Este proyecto se realizó con la colaboración del grupo de investigación de Estadística Espacial de la Universidad Nacional de Colombia, a través de la vinculación de un profesor del Departamento de Estadística de dicha universidad.

Se utilizaron datos de precipitación de 82 estaciones climáticas suministrados por la Corporación Autónoma Regional (CVC) y las imágenes satelitales de precipitación del período 1993 a 2011, por el Climate Hazard Group de la Universidad de California Santa Bárbara (UCSB). La precipitación pentadal fue modelada a través de un modelo de regresión funcional con covariable funcional. A través de geoestadística funcional se exploró la existencia de correlación espacial de los errores funcionales y se realizó la estimación de los parámetros funcionales del modelo a través de mínimos cuadrados generalizados. Los resultados mostraron la existencia de correlación espacial debido a la cercanía entre las estaciones climáticas. Finalmente, la información del satélite tuvo un aporte significativo en el modelo ajustado en la mayor parte de los años de estudio.

Palabras claves: precipitación, datos funcionales, regresión funcional, correlación espacial.

Abstract

The knowledge of the behavior of the climatic variables and modeling help predict, forecast, monitor climate change and to characterize the natural phenomena related to climate affecting the environment and water resources in agriculture that determine a region. That is why the modeling of climate variables, is one of the objectives of the Research Group in Applied Statistics INFERIR.

This research project is framed within the research on Regression and Prediction INFERIR group and aims to model the pentadal precipitation in the Valle del Cauca in the period 1993-2011, in order to generate information to provide input for decision-making of farmers, research centers, regional autonomous corporations and health institutions, among others. This project was conducted in collaboration with the research group Spatial Statistics from the National University of Colombia, through linking a professor in the Department of Statistics at the university..

They were used rainfall data from 82 weather stations supplied by the Regional Autonomous Corporation (CVC) and satellite images of precipitation for the period 1993-2011, the Climate Hazard Group at the University of California Santa Barbara (UCSB). The pentadal precipitation was modeled through a functional regression model with functional covariate. Through functional geostatistics explored the existence of spatial correlation of functional errors and estimation of functional parameters of the model was performed using generalized least squares. The results showed the existence of spatial correlation due to the proximity between stations. Finally, the satellite information had a significant contribution in the adjusted model in most of the years of study.

Keywords: precipitation, functional data, functional regression, spatial correlation.

Contenido

Agradecimientos	7
Resumen	9
Abstract	10
1. Introducción	2
2. Antecedentes	7
3. Objetivos	13
4. Caso de estudio	14
4.1. Precipitación	15
4.2. Zona de estudio	15
4.3. Climatología del Valle del Cauca	16
4.4. Datos de estaciones climáticas	16
4.5. Datos de satélite	17
5. Predicción de datos faltantes y análisis exploratorio de datos	19
5.1. Geoestadística univariada	23
5.1.1. Variable regionalizada	24
5.1.2. Momentos de una variable regionalizada	24
5.1.3. Semivariograma experimental	25
5.1.4. Semivariograma teórico	25
5.1.5. Predicción espacial (Kriging)	27
5.2. Resultados de la predicción de datos faltantes	30
5.3. Resultados del análisis exploratorio de datos completos	31
5.4. Construcción de las curvas pentadales de estaciones climáticas	36
5.5. Conclusiones	45

6. Análisis descriptivo de datos funcionales	46
6.1. Conceptos preliminares	46
6.2. Base de funciones	48
6.3. Construcción de datos funcionales	51
6.4. Estimación de la la varianza	53
6.5. Número óptimo de funciones de la base (K)	54
6.6. Estadística descriptiva funcional	55
6.7. Resultados de la construcción y análisis exploratorio de datos funcionales . .	56
6.8. Conclusiones	63
7. Modelo de regresión lineal para datos funcionales	64
7.1. Modelo de regresión funcional concurrente	65
7.2. Intervalos de confianza en el modelo de regresión funcional	66
7.3. Bondad de ajuste del modelo	67
7.4. Modelo de regresión funcional entre precipitación y altura	68
7.5. Conclusiones	71
8. Estimación de la estructura de la correlación espacial de los errores	72
8.1. Geoestadística para datos funcionales	72
8.2. Identificación de la estructura de correlación espacial	74
8.3. Resultados de la modelación de la correlación espacial de los residuales funcionales	75
8.4. Conclusiones	78
9. Modelo de regresión lineal funcional con estructura de correlación espacial en los errores	79
9.1. Definición del modelo y estimación de los parámetros funcionales	79
9.2. Ajuste del modelo integrando la estructura de correlación espacial	82
9.3. Conclusiones	89
10. Predicción espacial	90
10.1. Evaluación del nivel predictivo del modelo	90
10.2. Mapas de predicción con el modelo propuesto	96
10.3. Conclusiones	100
11. Conclusiones	101
11.1. Discusión	101
11.2. Conclusiones generales	102
11.3. Trabajo futuro	104
Bibliografía	105

A. Curvas ajustadas y observadas de 82 estaciones, para el año 2011	110
B. Modelo ajustado con correlación y sin correlación de 82 estaciones, para el año 2011	112
C. Código para estimar datos faltantes a través de geoestadística univariada	114
D. Código para la construcción de los datos funcionales de precipitación de estaciones climáticas y de satélite	116
E. Código para estimar un modelo de regresión funcional asumiendo independencia en los errores	118
F. Código para estimar las matrices de correlación espacial a partir de geoestadística funcional	119
G. Código para estimar un modelo de regresión lineal funcional con errores correlacionados	120

Lista de Tablas

4.1. Distribución del número de estaciones con información de precipitación disponible en el Valle del Cauca, en el período 1993-2011.	17
5.1. Distribución del número de estaciones por año.	20
5.2. Resumen de estadísticas descriptivas de la altura de las estaciones climáticas.	20
5.3. Listado de estaciones de análisis ($n = 82$).	21
5.4. Parámetros estimados de semivariograma por año, en el período 1993-2011.	30
5.5. Comportamiento del ECM según modelo ajustado, en el período 2003-2011.	31
5.6. Estadísticas descriptivas de la correlación lineal entre los datos de estaciones climáticas y de satélite, en el período 1993-2011	35
5.7. Porcentaje de la existencia de asociación lineal entre los datos de estaciones climáticas y de satélite, en el período 1993-2011.	36
6.1. Número de funciones para datos de estaciones y de satélite, en el período 1993-2011.	57
6.2. Coeficientes del ajuste de los datos funcionales, para 82 estaciones climáticas en el año 2011.	58
8.1. Estructuras de covarianza espacial.	74
8.2. Suma minimizada de cuadrados según modelo de covarianza espacial, en el período 1993-2011.	75
8.3. Resultados del modelo ajustado de correlación espacial, en el período 1993-2011.	76
9.1. Aporte de la precipitación de satélite, en el período 2003-2011.	84
10.1. Comportamiento del SCE_{VC} (<i>Mediana</i>) con correlación espacial en los errores y sin correlación espacial en los errores , en el período 1993-2011.	92

Lista de Figuras

4.1. Zona de estudio. Fuente: Instituto Geográfico Agustín Codazzi (IGAC, 2012).	16
4.2. Ejemplo de una imagen de precipitación procedente de satélite para el Valle del Cauca.	18
5.1. Localización de las estaciones climáticas.	22
5.2. Diagrama de dispersión de las estaciones climáticas según altura.	23
5.3. Componentes del semivariograma.	26
5.4. Comportamiento de la precipitación pentadal de estaciones climáticas sin acumular, en el período 1993-2011.	33
5.5. Comportamiento de la precipitación pentadal de satélite sin acumular, en el período 1993-2011.	34
5.6. Ilustración del proceso generador de la curva pentadal de precipitación. . . .	37
5.7. Relación entre los datos observados de precipitación de estaciones climáticas y de satélite, en el período 1993-2011.	38
5.8. Precipitación de estaciones climáticas, en el período 1993-2011.	39
5.9. Precipitación de satélite, en el período 1993-2011.	40
5.10. Precipitación de estaciones climáticas por región geográfica, en el período 1993-2011.	41
5.11. Precipitación de satélite por región geográfica, en el período 1993-2011. . . .	42
5.12. Precipitación promedio de estaciones climáticas por región geográfica, en el período 1993-2011.	43
5.13. Precipitación promedio de satélite por región geográfica, en el período 1993-2011.	44
6.1. Comportamiento de los datos observados (puntos), datos suavizados (línea) e intervalo del 95 % de confianza (línea azul) de las primeras cuatro estaciones, para el año 2011.	59
6.2. Datos funcionales de precipitación de estaciones climáticas, promedio y desviación estándar en el período 1993-2011.	61
6.3. Datos funcionales de precipitación de satélite, promedio y desviación estándar en el período 1993-2011.	62
7.1. Coeficiente de regresión asociado a la altura, en el período 1993-2011.	69

7.2. Coeficiente de regresión asociado a la altura al cuadrado, en el período 1993-2011.	70
8.1. Traza-Variograma bin de los residuales funcionales, en el período 1993-2011.	77
9.1. Intercepto funcional, en el período 1993-2011.	85
9.2. Coeficiente de regresión asociado a los datos de satélite, en el período 1993-2011.	86
9.3. Residuales estudentizados del modelo ajustado, en el período 1993-2011. . .	87
9.4. Residuales versus valores predichos, en el período 1993-2011.	88
10.1. Curva de $R^2(t)$ funcional, con correlación espacial en los errores y sin correlación espacial en los errores, en el período 1993-2011.	91
10.2. Comportamiento de las curvas observadas y ajustadas, en el período 1993-2011.	94
10.3. Comparación entre las curvas observadas y estimadas para 20 estaciones climáticas, para el año 2011.	95
10.4. Comportamiento del <i>RECM</i> , en el período 1993-1998.	97
10.5. Comportamiento del <i>RECM</i> , en el período 1999-2004.	98
10.6. Comportamiento del <i>RECM</i> , en el período 2005-2011.	99

Capítulo 1

Introducción

El departamento del Valle del Cauca está localizado en el sur occidente de Colombia, limita con 5 departamentos y en su zona occidental tiene costas sobre el Océano Pacífico. Esto hace que esté muy expuesto a los efectos del fenómeno Niño-Niña (BIOPALMIRA, 2008). La precipitación de esta zona tiene un régimen bimodal, resultado de la actividad de la zona de convergencia intertropical, con dos trimestre secos y dos lluviosos. El 70 % de la precipitación total registrada en la región se presenta en los meses lluviosos (Carvajal et al., 1998). La economía de este departamento, fundamentalmente agrícola, es con frecuencia afectada por los cambios bruscos en las condiciones climáticas y en particular por los del patrón de precipitación, puesto que éste modifica la disponibilidad hídrica y a la postre repercute en la producción agrícola (Puertas Orozco et al., 2011). En algunas regiones del departamento se han registrado eventos extremos de precipitación que incluso generan cambios en la composición y estructura de los ecosistemas en estas mismas (Carvajal et al., 1998). Por todo lo anterior resulta muy relevante contar con herramientas que permitan caracterizar y predecir su comportamiento. En este trabajo se presentan alternativas estadísticas para estudiar su distribución espacial.

En Colombia las estaciones climáticas han sido instaladas teniendo en cuenta primordialmente criterios de accesibilidad y seguridad (Garcés & Santana, 2005). Esto hace que la distribución geográfica de la red de monitoreo climático sea muy irregular y conlleva por ende a que el estudio del comportamiento espacial de las variables climáticas (temperatura, precipitación, humedad, etc.) sea con frecuencia insatisfactorio, puesto que hay incertidumbre en muchas zonas (aquellas donde no hay estaciones). Una forma de solventar esta falencia es acudiendo al uso de información satelital. Así se proporciona una mayor cobertura geográfica y se puede hacer una caracterización más amplia del patrón de distribución de las variables en la región de interés. La utilización de satélites y sensores remotos como, entre otros, el Geostationary Operational Environmental Satellite (GOES) y el Tropical Rainfall Measuring Mission (TRMM) permite obtener un amplio número de registros de variables climáticas en espacio y tiempo, facilitando así el proceso

de caracterización de las mismas (Rojas, 2003).

No obstante su importancia en términos de cobertura geográfica y temporalidad, la información satelital presenta problemas de sesgo, es decir que no coincide exactamente con la obtenida en campo (cuando es posible hacer la comparación) (Xie et al., 2003, Funk & Verdin, 2003, Woldemeskel et al., 2013, Moazami et al., 2014). Se requiere entonces desarrollar modelos que permitan ajustarla de tal manera que se pueda aprovechar todo su potencial espacial y temporal. En particular para la realización de este trabajo se cuenta con datos de precipitación obtenidos por dos fuentes (campo y satélite) en estaciones climáticas del Valle del Cauca. Se dispone de información diaria y pentadal (acumulada de 5 días) de 19 años (1993-2011) en cada una de las estaciones de esta región. El propósito es estimar un modelo que permita aprovechar las bondades de los sensores remotos en términos de su posibilidad de registrar gran cantidad de datos en espacio y tiempo, pero corrigiendo a partir de su comparación con la información de campo. Utilizando este mismo, se podrá estimar la información histórica de precipitación en sitios de la región donde no hay estaciones climáticas, aumentando así la cobertura de la red.

Desde el punto de vista estadístico es complicado implementar un modelo de regresión lineal clásico, para modelar de forma simultanea un amplio volumen de información que surge cuando el número de años y de sitios de medición se incrementa. En estos casos, en lugar de utilizar los datos discretos por estación e imagen de satélite para cada año, se propone utilizar toda una entidad funcional que permita caracterizar el comportamiento de forma anual. Una alternativa es convertir los datos de cada sitio en curvas (por ejemplo usando técnicas no paramétricas) y hacer su estudio desde la perspectiva del análisis de datos funcionales (ADF) (Ramsay & Silverman, 2005). Para la realización de este trabajo se adopta esta metodología. Específicamente, se obtienen las curvas de precipitación de campo y de satélite empleando bases de funciones y se propone un modelo de regresión funcional concurrente que permita establecer la relación entre las curvas de campo (variable funcional respuesta) y de satélite (variable funcional predictora). El modelo estimado involucra la dependencia espacial debida a las estaciones.

Según el Departamento Nacional de Estadística (DANE), el Valle del Cauca basa su economía en actividades agrícolas e industriales. Los productos agrícolas más relevantes de la región son la caña de azúcar y el café (DANE, 2013). Dentro del sector industrial se destacan la fabricación de productos químicos y la producción de papel y cartón (DANE, 2013). No obstante la buena producción agrícola e industrial de los últimos años, es también reconocido que ésta podría haber sido mucho mejor, si no se hubiesen presentado eventos extremos en las condiciones climáticas. Este departamento ha sido afectado considerablemente por los efectos del fenómeno Niño-Niña, dada su cercanía al Océano Pacífico (Carvajal et al., 1998) y esto ha repercutido en su actividad económica. Para el

óptimo manejo de la actividad agroindustrial del departamento resulta de vital importancia el pronóstico y la alerta temprana de los cambios de corto y largo plazo en las condiciones ambientales. Una variable esencial en este aspecto es la precipitación. Su seguimiento, monitoreo y modelación es esencial para tomar medidas preventivas en muchos contextos y para muchas entidades públicas y privadas. Agricultores, centros de investigación, corporaciones autónomas e instituciones de salud, entre otros, serían beneficiados con información oportuna de la variación climática y en particular de la precipitación.

El registro de la precipitación puede hacerse a través de mediciones en campo (estaciones climáticas) o por medio de imágenes de satélite. Los datos de campo son en general insuficientes desde el punto de vista de cobertura espacial, puesto que las estaciones climáticas se distribuyen de manera irregular dentro de la región. Por otro lado con el satélite meteorológico se puede medir la precipitación con una resolución temporal constante y abarcando prácticamente toda el área de estudio. En este caso el problema es que la información registrada puede contener errores asociados a las limitaciones del equipo y a la variabilidad meteorológica (Funk & Verdin, 2003). Dadas las bondades y limitaciones de los dos tipos de información, resulta muy relevante proponer metodologías que permitan integrarlas de forma que se obtenga información con la precisión de la estación climática y la cobertura del satélite.

Desde el punto de vista estadístico ha habido diferentes enfoques para la modelación de la precipitación. Modelos lineales generalizados (Coe & Stern, 1982), mixtos (Zakaria & Moslim, 2015), series temporales (Cowpertwait et al., 1996), cadenas de Markov (Hughes et al., 1999), regresión no paramétrica (Lall et al., 1996) y geoestadística (Zquiza, 1998), entre otros, son enfoques para el análisis de este tipo de datos. El uso de una u otra herramienta depende de los objetivos específicos del estudio. Descripción espacio temporal, estimación de la variación media, predicción en el tiempo, predicción espacial, son solo algunos de los objetivos que se pueden abordar con el uso de herramientas estadísticas con datos de precipitación tomados en estaciones climáticas de una región.

El análisis estadístico de un gran volumen de datos espacio-temporales como los considerados en este trabajo se convierte en un reto. El uso de técnicas clásicas como las arriba mencionadas puede resultar limitado desde el punto de vista teórico o por su implementación computacional. Por ello es necesario proponer herramientas estadísticas que sean flexibles desde el punto de vista de sus supuestos y que sean fáciles de implementar computacionalmente.

Una alternativa relativamente nueva que es apropiada para describir y modelar fenómenos con variación espacio-temporal es el Análisis de Datos Funcionales (ADF). Esta aproximación parte de la construcción de curvas o superficies que después son utilizadas como insumos

(inputs) para el uso de técnicas clásicas como regresión, ANOVA, series de tiempo o geostatística, entre otras. En este trabajo se aborda esta aproximación para dar solución al problema de buscar relaciones entre series históricas de precipitación registradas en la estaciones y a través del satélite. Se parte de la construcción de curvas de precipitación a partir de técnicas no paramétricas de suavizado que después son usadas como entradas de un modelo de regresión funcional (Ramsay & Silverman, 2005).

Las técnicas de regresión funcional propuestas en la literatura (Ramsay & Silverman, 2005) no consideran la dependencia espacial entre los datos funcionales. Por ello en el presente trabajo se propone una modificación de los modelos de regresión funcional de tal forma que consideren la variabilidad espacial debida a las estaciones en campo.

En resumen se busca proponer un modelo de regresión que a partir de series históricas de precipitación registradas a través del satélite meteorológico permita hacer estimación de series de precipitación en estaciones climáticas con información deficiente o en sitios de la región no observados. Las estimaciones deben estar libres del sesgo que la información de satélite conlleva y tener en cuenta la correlación espacial inherente a la información registrada en estaciones climáticas.

En el departamento del Valle del Cauca la producción agrícola, uno de los principales actores de su economía y el de mayor crecimiento (DANE, 2011), es seriamente afectada por la variabilidad climática (Carvajal et al., 1998). La precipitación en particular influye notablemente sobre el rendimiento agrícola porque afecta el ciclo hidrológico (BIOPALMIRA, 2008). El conocimiento de su comportamiento es indispensable, entre otros, en los modelos de escorrentía (Cerón et al., 2011), la estimación de áreas diferenciadas de precipitación (Galindo et al., 2005) y modelos de cuencas (Puertas Orozco et al., 2011). Su modelación espacial y temporal es de gran relevancia ambiental y económica (Méndez Venegas, 2008).

Surge entonces la necesidad de proponer modelos que permitan caracterizar el comportamiento espacio temporal de la precipitación en el departamento. Los resultados de esta modelación son de gran utilidad en muchos contextos aplicados. Desde el punto de vista estadístico el análisis de esta información es un reto. Hay diferentes estructuras de correlación involucradas (espacial, temporal y espacio-temporal), información incompleta, se requiere hacer predicción espacial (predecir la precipitación en sitios sin información) y hay que analizar y modelar un gran volumen de datos.

Con la realización de este trabajo se busca de un lado aportar en la solución de un problema práctico que es de interés para la región y desde la perspectiva estadística, contribuir con una metodología que permita analizar grandes conjuntos de datos espacio-temporales y que

combinando distintas fuentes de información de precipitación permita hacer predicción de esta misma en sitios y zonas sin información.

El trabajo está organizado de la siguiente forma: En el capítulo 1 se presenta la introducción del trabajo de investigación. En el capítulo 2 se realiza una revisión de los antecedentes de la modelación de precipitación usando análisis de datos funcionales y otras metodologías estadísticas. En el capítulo 3 se presenta el objetivo general y los objetivos específicos del trabajo de investigación. En el capítulo 4 se describe la información analizada y el área de estudio. El capítulo 5 ilustra la metodología usada para depurar la información y el análisis exploratorio de la misma. En el capítulo 6 se muestra la aproximación usada para convertir los datos de satélite y de campo en curvas de precipitación y su respectivo análisis descriptivo. En los capítulos 7 y 8 se presenta la teoría de la regresión funcional concurrente con errores independientes y se hace una revisión de la teoría del análisis geoestadístico funcional como herramientas para medir la dependencia espacial entre curvas. En los capítulos 9 y 10 se propone una modificación del modelo de regresión funcional de forma que involucre la estructura de correlación espacial entre la estaciones y se aplica al caso de las curvas de precipitación bajo estudio. Finalmente en el capítulo 11 se hace una discusión de los resultados, se dan las conclusiones y algunas perspectivas futuras de análisis.

Capítulo 2

Antecedentes

En este capítulo, se presentan los antecedentes relacionados con la modelación de la precipitación, la integración de información de estaciones climáticas y de satélite y el análisis de datos funcionales.

Modelación con otras metodologías

Con respecto a la modelación de la precipitación en el Valle del Cauca, se destaca el proyecto de investigación *Modelación Estadística de Variables Climáticas en el Suroccidente Colombiano*, desarrollado por el Grupo de Investigación en Estadística Aplicada INFERIR (Andrade et al., 2012). En el marco de este proyecto se encuentran los subproyectos de acuerdo a dos metodologías:

Modelo Lineal Mixto

Cruz & Males (2012) modelaron la precipitación mensual en el Valle del Cauca, a través de un Modelo Lineal Mixto Generalizado. Los autores trabajaron con datos de precipitación de 155 estaciones climáticas, en el período 1971 a 2009. En una primera fase, identificaron dos componentes: uno debido al tiempo (períodos secos y húmedos), los fenómenos del Niño y la Niña y otro a la localización de las estaciones climáticas relacionadas con la altitud, las cuales intervienen en el comportamiento de la precipitación mensual del Valle del Cauca. Finalmente, los autores encontraron que la precipitación mensual sigue una distribución Gamma. Así mismo, modelaron la presencia de correlación espacial utilizando un modelo Matern y hallaron evidencia de un aporte importante en la estimación de los parámetros de la variable altura y su componente cuadrática en el modelo propuesto.

Portillo & Medina (2012) modelaron la precipitación diaria en el Valle del Cauca, utilizando un modelo lineal mixto generalizados, en el período 1971 a 2009. Hallaron relación entre la precipitación y la altitud y la influencia de los fenómenos del Niño y

la Niña, además de los períodos secos y húmedos en los sitios de estudio. Así mismo, encontraron que la cantidad de precipitación sigue una distribución log-normal o Gamma. En el modelo, incorporaron factores fijos como el Índice de Oscilación Sur, la altitud y factores aleatorios como el Tiempo (años) y la localización de las estaciones climáticas.

Geoestadística

Zape & Castañeda (2012) caracterizaron la precipitación mensual durante los años 1971 a 2009 en la región del Valle del Cauca, Colombia, examinando la correlación temporal y espacial de las series a lo largo de los años, para 154 estaciones climáticas. Los autores, encontraron ausencia de correlación temporal a lo largo de los años, sin embargo, encontraron patrones estacionales de la variable durante cada año. Posteriormente, evaluaron la componente espacial, a partir de Modelos Mixtos Lineales Generalizados examinando la correlación espacial por medio de los residuales y ajustaron un modelo incorporando la estructura de correlación espacial, esta estructura fue identificada por medio del variograma teórico basado en métodos Geoestadísticos. Finalmente obtuvieron como resultado, que el modelo óptimo para modelar la correlación espacial fue el modelo Matern.

Modelación integrando imágenes satelitales

Un acercamiento en la modelación incorporando imágenes de satélite lo realizaron Grimes et al. (1999), quienes utilizaron el Kriging con derivada externa para integrar datos de estaciones climática y de satélite. Los autores concluyeron que el procedimiento de integración de ambas fuentes de información, mejoró las estimaciones de precipitación. Posteriormente Rojas (2003) comparó la precipitación mensual obtenida del satélite Tropical Research Measurement Mission (TRMM) y la precipitación observada en seis estaciones climáticas, ubicadas en la costa norte del Perú. Este proceso fue realizado mediante el análisis de la correlación de las series temporales de ambas fuentes de datos, concluyendo que las estimaciones de satélite siguen el mismo comportamiento de los datos de superficie.

Teo & Grimes (2007) integraron datos de precipitación diaria de satélite y estación climática en Gambia (África), evaluando las variaciones de los valores de precipitación de ambas fuentes. Este proceso lo realizaron, utilizando la metodología doble Kriging y un indicador de rendimiento a partir de Kriging por bloques, donde cada bloque es el valor del pixel de precipitación proveniente de satélite. A la integración la denominaron “gauge-pixel” que fue comparada con los datos de precipitación de satélite, a partir de la función de distribución de los datos, para evaluar el efecto de las variaciones de los datos calibrados con el valor observado de precipitación de cada pixel. Finalmente, los autores concluyen que la integración de ambas fuentes de datos se puede utilizar como una método para mejorar la sensibilidad en modelos hidrológicos.

Méndez Venegas (2008) utilizó métodos geoestadísticos como el Kriging, Cokriging Ordinario y Cokriging Colocado para modelar de manera conjunta la precipitación mensual de datos de radar y estaciones climáticas en el Distrito Federal de México, con el objetivo de dar seguimiento a los ciclones tropicales. Para esto, el autor trabajó con 61 estaciones climáticas ubicadas dentro de la zona de cobertura del radar. Para el análisis geoestadístico conjunto, utilizó los valores de ambas mediciones en ubicaciones comunes. A las coordenadas correspondientes a la ubicación de una estación dada, le asociaron un valor de precipitación de radar de la siguiente manera: se localizó el píxel asociado a la coordenada de la estación y se tomó su valor de precipitación acumulada de radar junto con los valores de los ocho píxeles circundantes, finalmente se seleccionó el que tuviera la menor diferencia absoluta con respecto al valor de precipitación medida en la estación, obteniendo estimaciones de precipitación acumulada de estaciones climáticas y radar ubicadas en las localizaciones de las estaciones climáticas.

Alvarez-Villa et al. (2011) modelaron la precipitación media anual en Colombia, a partir de una herramienta geoestadística denominada Kriging con deriva externa. Utilizaron información de precipitación media anual de 1,180 estaciones climáticas en el período 1997-2002. Posteriormente estimaron un campo de precipitación media anual con una resolución espacial de 4x4 km basándose en la información del sensor PS del satélite Tropical Rainfall Measurement Mission (TRMM). Luego realizaron un análisis estructural por regiones (6 en total), ajustando semivariogramas a cada región y su respectivo modelo teórico. Finalmente promediaron los valores de las celdas comunes adyacentes para estimar un nuevo campo de precipitación para toda la zona de estudio.

Woldemeskel et al. (2013) unieron datos de precipitación mensual proveniente de estaciones climáticas y del satélite TRMM 3B42 en Australia, con el objetivo de evaluar la precisión asociada a esta integración y las ventajas del proceso de integrar ambas fuentes de datos para las condiciones climáticas de Australia y de esta manera generar una base de datos retrospectiva de precipitación. Para cumplir los objetivos del trabajo, los autores compararon las mediciones de estaciones climáticas con las mediciones del satélite (TRMM) 3B42 tomadas en las localizaciones de las estaciones. Luego ajustaron los valores de precipitación utilizando Splines y el Inverso de la Distancia Ponderada. Posteriormente, integraron los valores de precipitación de estaciones climáticas con el acumulado mensual TRMM 3B42 utilizando un procedimiento de linealización ponderada. Finalmente, evaluaron los resultados de los errores de las localizaciones observadas, a partir de validación cruzada (CV), concluyendo que la fusión de los dos conjuntos de datos mejoró la estimación de precipitación, cuando la red de estaciones era escasa.

Moazami et al. (2014) desarrollaron un método de simulación conjunto basado en

cóputas para el análisis de la incertidumbre y el ajuste de dos productos de precipitación diaria de satélite de alta resolución (PERSIANN y TMPA-3B42) en la parte Suroeste de Irán, en el período 2003 a 2006. Para esto, utilizaron un conjunto de sesenta eventos de precipitación diaria, luego evaluaron las diferencias entre las mediciones de estaciones climáticas como dato de precipitación de referencia y estimaciones de precipitación por satélite, a estas diferencias la denotaron como el sesgo observado. Posteriormente, utilizaron una cópula gaussiana multivariante, construida a partir de la distribución normal multivariante, con el objetivo de generar múltiples campos aleatorios basados en los sesgos observados y asignando una estructura de dependencia. Por último, realizaron simulaciones del sesgo y las incorporaron sobre las estimaciones originales de precipitaciones del satélite, con el fin de obtener un conjunto de realizaciones de precipitación ajustada, integrándolas a las estimaciones de satélite. Finalmente, los autores concluyeron que la fiabilidad y el rendimiento del modelo desarrollado en lo que respecta a la corrección de sesgo, corresponden al 10 % de eventos de precipitación diaria y que mejora en 35.42 % el Error Cuadrático Medio, cuando se ajustaron los datos con las estimaciones de satélite.

Modelación de la precipitación pentadal

Xie et al. (2003) estimaron la precipitación pentadal, integrando datos de estaciones climáticas y estimaciones de satélite en todo el mundo, en el período 20 a 24 de Julio de 1988. Para este propósito, estimaron la precipitación utilizando el algoritmo de integración basado en la reducción del error aleatorio y combinando las estimaciones de satélite, a través del método de Máxima Verosimilitud, que consiste en combinar linealmente el inverso al cuadrado de los coeficientes obtenidos a través del modelo de regresión lineal. Los resultados mostraron que la integración de ambas fuentes de información reprodujo con una alta calidad los patrones de distribución espacial de la precipitación total y las variaciones temporales, no obstante, mostraron la existencia de un alto sesgo en algunas áreas de la tierra, específicamente en zonas oceánicas debido a la falta de observaciones de estaciones climáticas.

Verdin et al. (2015) modelaron la precipitación pentadal en Centro y Sur América, específicamente en Guatemala, Belice, El Salvador, Honduras, Nicaragua, Costa Rica, Panamá, Colombia y Venezuela, en el período de Enero a Julio de 2009. Para esto, utilizaron las estaciones climáticas de las zonas de estudio y los datos de satélite estimados con el producto CHIRP (Climate Hazard Infrared Precipitation). Así mismo, aplicaron métodos geoestadísticos como el Kriging Bayesiano para predecir la precipitación de las estaciones incorporando como covariable la precipitación de satélite. De acuerdo a lo propuesto por los autores, concluyeron que el modelo presentó un buen ajuste, evaluándolo con validación cruzada, excluyendo el 25 % y 50 % de las observaciones. También, mostraron que la integración espacial de ambas fuentes de datos fueron mejores y más robustas en

comparación con los datos de satélite.

Modelación de la precipitación a partir de ADF

Con respecto a la modelación utilizando análisis de datos funcionales, se destacan los siguientes trabajos:

Suhaila et al. (2011) modelaron la precipitación diaria en la península de la Malasia con observaciones diarias de 50 estaciones climáticas, en el período 1975-2004. Los datos de lluvia se recogieron en una base diaria, en forma de observaciones discretas. El objetivo de su estudio fue convertir los valores de precipitación en una curva suave o función para representar el proceso de precipitación continua en cada región utilizando del análisis de datos funcionales. Puesto que los datos de precipitación muestra un patrón periódico en cada región, los autores utilizaron una base de Fourier para los datos funcionales. Los autores utilizaron, once bases de funciones para describir el patrón de precipitaciones para las estaciones en el Este y cinco funciones de base para describir el patrón de lluvias en el oeste. Adicionalmente, tuvieron en cuenta los períodos húmedos y secos, así como los valores máximo y mínimo de precipitaciones. Para determinar diferentes patrones de lluvia, compararon las curvas entre las regiones estudiadas, utilizando el análisis funcional de la varianza. Finalmente, concluyeron que existen diferencias significativas entre las curvas medias funcionales de cada región. Las diferencias más grandes las encontraron entre las regiones del Este y del Noroeste y estas diferencias pueden ser, probablemente, debido al efecto de la topografía de la zona de estudio.

Scheipl et al. (2015) propusieron un modelo funcional aditivo mixto donde la variable respuesta tiene una estructura de correlación espacial o temporal. Los autores, desarrollaron el modelo con variable explicativa funcional y escalar. Además, se basaron en el modelo aditivo mixto clásico para desarrollar la parte de estimabilidad e inferencia en el campo funcional. Para ilustrar el método aplicaron el modelo propuesto para predecir la precipitación promedio mensual a partir de la temperatura, para esto utilizaron 35 estaciones climáticas en Canadá, en el período 1960 a 1994.

Norliyana & Suhaila (2015) los autores utilizaron un análisis de datos funcionales aplicado a la precipitación diaria en la Península de Malasya en el período 1985 a 2009, con el objetivo de caracterizar las variaciones de precipitación y el viento debida a la influencia de los monzones. Para esto, los autores transformaron los datos de precipitación y del viento en una curva suave utilizando bases de Fourier, puesto que los datos de viento y lluvia mostraron un patrón periódico. Así mismo, agregaron una penalización sobre de rugosidad en la construcción del objeto funcional y la estimación la realizaron por mínimos cuadrados. Finalmente, concluyeron que utilizando un número pequeño de funciones de

bases, la diferencia es pequeña, independiente del parámetro de penalización. No obstante, cuando se emplea una función de base grande, la penalización debe añadirse, con el objetivo de optimizar el ajuste de los datos funcionales. Adicional a esto, los autores encontraron una relación entre los datos funcionales del viento y la lluvia, a través, de la correlación cruzada funcional entre ambas variables.

Capítulo 3

Objetivos

Objetivo general

- Ajustar un modelo de regresión funcional concurrente integrando datos de estaciones climáticas y de satélite, considerando la estructura de correlación espacial debido a las estaciones climáticas.

Objetivos específicos

- Ajustar un Modelo de Regresión Funcional con variable de respuesta y variable explicativa funcional.
- Modelar la estructura de correlación espacial de la precipitación pentadal.
- Predecir curvas de precipitación pentadal en aquellas zonas no observadas.

Capítulo 4

Caso de estudio

La precisión en la estimación de la precipitación es crucial para muchos propósitos, tales como la modelación climática, para usos ambientales, en la agricultura, en los grupos de investigación, en las secretarías de salud, con el propósito de generar información como insumo para la toma de decisiones en procura de la sostenibilidad en los recursos hídricos, mitigar el riesgo de desastres naturales y en la generación de conocimiento y aplicación de nuevas metodologías para la caracterización y predicción de la precipitación en una región (Puertas Orozco et al., 2011, Méndez Venegas, 2008, Ravelo & Santa, 2000).

Actualmente, se encuentran dos métodos comunes para medir la precipitación: el primer método es la medición directa a través de estaciones climáticas y el segundo método es el uso de técnicas de teledetección como, radares o satélites meteorológicos (Funk & Verdin, 2003). Ahora bien, la forma más común de predicción de precipitación en una región es a través de la interpolación o el uso del promedio como estimación del dato de precipitación generado por un conjunto de estaciones. Sin embargo, esta predicción es confiable siempre y cuando haya una densa red de estaciones climáticas en la zona (Ravelo & Santa, 2000, Rojas, 2003, Teo & Grimes, 2007, Méndez Venegas, 2008).

En las zonas donde hay solamente una red dispersa de estaciones, el enfoque para obtener la precipitación a partir de predicción espacial está sujeta a un alto grado de incertidumbre. Además, la medición de la precipitación utilizando estaciones climáticas puede ser incierta debido a los efectos del viento y las fallas en la calibración en los sistemas de medición (Yozgatligil et al., 2013). No obstante, algunos de los problemas anteriores pueden ser remediados con el uso de técnicas de medición basadas en satélites (Funk & Verdin, 2003, Verdin et al., 2015). Estas técnicas proporcionan una medición continua a una resolución más fina tanto temporal como espacialmente, con respecto a las mediciones de estaciones climáticas. Sin embargo, presentan errores de medición debido a la pérdida de información en la imagen de precipitación emitida por el satélite debido a las fallas en la transferencia en tiempo real y por la nubosidad que genera datos faltantes en la

imagen (Woldemeskel et al., 2013). El presente trabajo de investigación está enfocado a la modelación de la precipitación integrando los datos de estaciones climáticas y de satélite en el Valle del Cauca, en el período 1993-2011 a partir de un modelo de regresión funcional.

A continuación se describen algunos conceptos relacionados con la variable de estudio, zona de estudio y climatología de la zona de estudio.

4.1. Precipitación

Según la definición oficial de la Organización Meteorológica Mundial (OMM) (OMM, 2001), la precipitación es la caída de partículas líquidas de agua de diámetro mayor de 0.5 mm o de gotas menores, pero muy dispersas. Es un fenómeno atmosférico que se inicia con la condensación del vapor de agua que se encuentra en las nubes cuando éstas atraviesan capas de aire frío contenido en las nubes. La formación de gotas de precipitación, a partir de las diminutas gotas que forman la nube, necesita la existencia de cristales de hielo sobre los que se depositan las gotas de agua. De este modo van creciendo los cristales hasta que son lo suficientemente grandes para caer por su propio peso, dando lugar a la precipitación en el momento en que se funden antes de llegar al suelo (OMM, 2001).

4.2. Zona de estudio

En el sector central del Sur Occidente colombiano se encuentra el Departamento del Valle del Cauca, entre los 3° 05'y 5° 0'de latitud norte y los 75° 41'y 77° 33'de longitud oeste; comprende 42 municipios y presenta dentro de sus límites políticos tres regiones bien definidas; la costa Pacífica, el valle interandino del Río Cauca y la región andina a la que corresponden las cordilleras Occidental y Central (ver Figura 4.1). La máxima altura (4,080 metros) se encuentra en los farallones de Cali. El departamento del Valle del Cauca cuenta con una extensión territorial de 21,195 km^2 que representa el 1.5 % del área nacional (Escobar Carmona, 2006, BIOPALMIRA, 2008).



Figura 4.1: Zona de estudio. Fuente: Instituto Geográfico Agustín Codazzi (IGAC, 2012).

4.3. Climatología del Valle del Cauca

El régimen de precipitación en el Valle del Cauca se define en tres regiones: costa pacífica, valle interandino del río cauca y la región andina. En la costa pacífica, las precipitaciones máximas se presentan en un solo período del año, durante los meses de Septiembre, Octubre y Noviembre (más de 700 mm./mes), los meses restantes se registran precipitaciones de menor intensidad. El régimen unimodal es producido por los vientos provenientes del Océano Pacífico, originando así una zona de alta pluviosidad durante todos los meses del año (Escobar Carmona, 2006).

El Valle geográfico del Río Cauca y las cordilleras Occidental y Central, presentan comportamientos similares, aunque los promedios son diferentes para ambas zonas. El régimen de distribución de la precipitación es bimodal, con dos trimestres secos en los meses de Diciembre - Enero - Febrero y Junio - Julio - Agosto y dos períodos lluviosos en Marzo - Abril - Mayo y Septiembre - Octubre - Noviembre. Este comportamiento es el resultado de la actividad de la Zona de Confluencia Intertropical (ZCIT); esta se localiza en el sur durante los meses de Enero y Febrero, comenzando su desplazamiento hacia el norte a mediados de Marzo, concluyendo su recorrido a mediados del año. Durante el mes de Septiembre inicia el desplazamiento hacia el sur, el cual termina en el mes de Enero (Carvajal et al., 1998).

4.4. Datos de estaciones climáticas

Los datos de estaciones climáticas utilizados en esta investigación, fueron proporcionados por la Corporación Autónoma Regional del Valle del Cauca (CVC). En el proceso de recolección de la información, se obtuvo la información correspondiente a las series históricas

de precipitación total diaria proveniente de las estaciones, posteriormente se verificó la información del número de estaciones disponibles por año en el período 1993-2011, con el objetivo de establecer uniformidad en el número de estaciones con información disponible para todo el período de análisis. Para tener una idea de la información procedente de las estaciones climáticas, en la Tabla 4.1 se muestra la distribución del número de estaciones con información disponible para el período de estudio (1993-2011).

Tabla 4.1: Distribución del número de estaciones con información de precipitación disponible en el Valle del Cauca, en el período 1993-2011.

Año	Número de estaciones	Año	Número de estaciones
1993	162	2003	141
1994	167	2004	141
1995	168	2005	142
1996	163	2006	134
1997	149	2007	134
1998	150	2008	134
1999	155	2009	134
2000	152	2010	129
2001	143	2011	127
2002	143		

4.5. Datos de satélite

Los datos de satélite fueron proporcionados por el Climate Hazard Group de la Universidad de California Santa Barbara (UCSB). La información de satélite se denomina Climate Hazards IR Precipitation (CHIRP) y fue construida a partir de la combinación de dos estimaciones de precipitación por infrarrojo, una es la B1 que comprende el período de 1981 hasta el 2008 y la otra es la CCP-IR que comprende desde el año 2000 hasta el presente. Posteriormente el Climate Hazard Group ajustaron un modelo de precipitación mensual para producir el porcentaje de anomalía del IRP (Infrared precipitation), que finalmente fue integrado con el FCLEAM (FEWS Climatology a climatological field calculated using multivariable regression) produciendo el CHIRP (Pedreros, 2012).

El período mas desagregado de la información corresponde a valores de precipitación total pentadales (cada 5 días). La información de satélite fué parte del TRMM (Tropical Rainfall Measurement Mission), con 25x25 km de resolución inicial y el otro dato es la precipitación estimada por infrarrojo a partir del satélite NOAA a una resolución de 4x4 km. Se transformaron los valores originales del TRMM a una resolución de 5 km y se dividió por su climatología misma del TRMM, esa anomalía porcentual se promedió con

la anomalía porcentual del infrarrojo y se multiplicó por el FCLEAM (FEWS Climatology a climatological field calculated using multivariable regression), todo esto es el producto denominado FTIP (FEWS TRMM IR Precipitation).

Los datos de precipitación de satélite utilizados para el estudio tienen la resolución de 5x5 km (Figura 4.2) al igual que la distribución espacial de los dos satélites y la estructura del FCLEAM (Pedreros, 2012). Con esta información, se extrajo el dato en las localizaciones de las estaciones climáticas para cada año y se acumularon, obteniendo finalmente la precipitación acumulada pentadal para datos de satélite en el período 1993 a 2011.

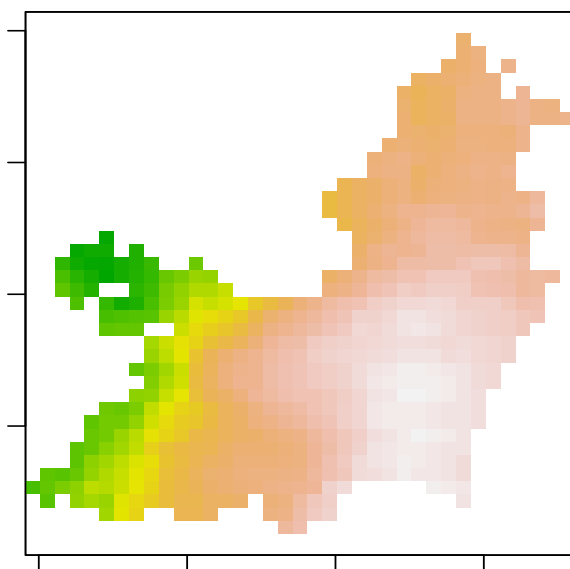


Figura 4.2: Ejemplo de una imagen de precipitación procedente de satélite para el Valle del Cauca.

Capítulo 5

Predicción de datos faltantes y análisis exploratorio de datos

A continuación, se muestra la descripción del procesamiento de la información que va a ser analizada en el trabajo de investigación. Inicialmente se hace una descripción del método de estimación de datos faltantes a partir de Geoestadística clásica aplicado a los datos de estaciones climáticas, y finalmente, se muestra el análisis exploratorio de las curvas observadas de precipitación provenientes de estaciones climáticas y de satélite. Los elementos teóricos para el desarrollo de este capítulo son extraídos de Cressie (1993) y Giraldo (2002).

Una de las dificultades que se presenta en el análisis de variables climatológicas al momento de procesar información con el objetivo de realizar análisis estadísticos, es la falta de datos en algunas estaciones climatológicas que se encuentran en actividad de forma irregular, generando una serie de observaciones faltantes en dichos registros. Como solución metodológica a este problema, generado por factores humanos o informáticos, es necesario estimar la información faltante de la variable climática de interés (Urrutia et al., 2010, Chen & Liu, 2012). A continuación se describe el proceso de estimación de los datos faltantes de este trabajo de investigación.

El proceso de estimación se realizó para la información proveniente de estaciones climáticas, ya que los datos de satélite tienen información completa dentro de la zona de cobertura. Inicialmente, se excluyeron las estaciones que tenían información faltante por encima del 20% y las estaciones localizadas por debajo de 700 msnm y por encima de 2,000 msnm. Dentro del grupo de estaciones establecidas para analizar, se propuso estimar los datos faltantes de las estaciones climáticas interpolando los datos por día. Luego se pasó de datos diarios a totales cada cinco días (datos pentadales). Finalmente, la información proveniente de estaciones climáticas, con la cual se trabajó en este proyecto fue la precipitación acumulada pentadal para datos de estaciones climáticas en el período 1993

a 2011.

En la Tabla (5.1) se muestra el número de estaciones por año, para el período de estudio 1993-2011. De acuerdo a esta tabla, se puede observar que hubo entre 110 y 144 estaciones con información disponible en el período de análisis. Inicialmente se contaba con 238 estaciones climáticas, de las cuales fueron excluidas 28 estaciones porque estaban ubicadas en corregimientos del departamento del Cauca.

De las 210 estaciones climáticas se excluyeron 118 que tuvieron datos faltantes por encima del 20 % y 10 estaciones que estaban localizadas por debajo de los 900 metros por encima de 2000 metros, finalmente se mantuvieron ($n = 82$) estaciones climáticas con información en los 19 años de estudio. En la Tabla 5.2 se muestran las estadísticas descriptivas de las estaciones climáticas que van a ser analizadas en el estudio, se destaca una altura promedio de 1,194 metros con una desviación de 332 metros y un rango de altura entre 721 metros y 1,915 metros. En la Tabla 5.3 se muestra el listado de las 82 estaciones que son incluidas en el análisis.

Tabla 5.1: Distribución del número de estaciones por año.

Año	Número	Año	Número
1993	138	2003	122
1994	143	2004	122
1995	144	2005	122
1996	140	2006	115
1997	130	2007	115
1998	131	2008	115
1999	134	2009	116
2000	132	2010	113
2001	124	2011	110
2002	124		

Tabla 5.2: Resumen de estadísticas descriptivas de la altura de las estaciones climáticas.

Indicador	Valor
Media	1194.0
Mediana	1003.2
Moda	1795.0
Desviación estándar	332.0
Rango	1194.0
Mínimo	721.0
Máximo	1915.0

Tabla 5.3: Listado de estaciones de análisis ($n = 82$).

Código de la estación	Nombre de la estación	ESTE (metros)	NORTE (metros)	ALTURA	Código de la estación	Nombre de la estación	ESTE (metros)	NORTE (metros)	ALTURA
2610000102	LA BALASTRERA	1053.277	833.315	1014	2622400101	COLEGIO SAN LUIS	1060.145	874.386	1106
2610000103	CRUCES	1128.670	1006.201	947	2622400103	PLANTA RIO CALI	1059.116	872.750	1054
2610000105	PUERTO MOLINA	1116.983	1003.001	906	2622400104	BRASILIA	1049.942	873.548	1795
2612700105	MATAPALO	1071.550	887.219	947	2622410102	YANACONAS	1052.902	871.956	1640
2612710101	FLORIDA	1092.694	859.077	1031	2622420101	MONTEBELLO	1058.590	877.406	1273
2612710103	ING LA QUINTA	1081.307	869.664	975	2622420102	SAN PABLO	1050.799	880.677	1915
2612710106	GUACHAZAMBOLO	1070.420	878.516	949	2622420103	AGUACATAL	1052.410	879.525	1621
2612720101	YUNDE CARRIZAL	1073.363	874.565	957	2622500101	DAPA	1058.141	885.463	1762
2612800104	AUII	1107.741	892.337	1615	2622500102	LLOREDA GRASAS	1064.021	880.860	961
2612810108	TENJO	1100.725	881.119	1604	2622600101	SANTA INES	1058.923	892.661	1656
2612810112	AUSTRIA	1100.481	887.551	1791	2622600103	YUMBILLO	1058.658	887.682	1795
2612810122	PLANTA NIMA	1095.516	884.101	1146	2622900101	VILLAMARIA	1066.418	902.968	1719
2613000102	EL PARAISO	1097.514	895.392	1321	2622900102	OCACHE	1063.222	900.852	1588
2613000103	EL CASTILLO	1098.371	897.088	1353	2623300101	BUENOS AIRES	1071.894	916.465	1570
2613100103	LA SELVA	1098.009	905.879	1667	2623500101	EL CANEY	1073.210	926.315	1521
2613100107	COSTA RICA	1094.087	907.025	1248	2624000102	RIOFRIO	1088.879	949.742	932
2613200105	LA PATAGONIA	1096.397	916.387	1978	2624600102	EL RETIRO	1093.540	977.913	1695
2613600101	DOSQUEBRADAS	1099.540	924.028	1870	2624600103	LA HERRADURA	1095.419	964.977	976
2613600103	LA PRIMAVERA	1097.750	925.211	1663	2624700103	EL VESUBIO	1109.573	1001.447	1002
2613600105	ACUEDUCTO BUGA	1088.027	922.199	1025	2624700104	EL PORVENIR	1106.692	994.155	1074
2613800101	ANGOSTURAS	1099.557	931.142	1292	2624700108	LAS PENAS	1107.650	1007.418	1732
2614200101	VENUS	1111.189	937.278	1627	2624700109	MONTECRISTO	1098.319	980.253	1478
2614400102	BUGALAGRANDE	1103.939	957.633	997	2624700110	EL AGUACATE	1097.347	977.786	1552
2614400104	IRLANDA	1127.716	955.283	1685	2624700113	EL OSO	1108.498	997.946	1173
2614900101	EL SOCORRO	1124.070	987.294	974	2624700114	BUENAVISTA	1101.994	986.244	1620
2614900102	LA ELIVIRA	1114.397	977.096	955	2624700115	EL OREGANO	1100.251	984.030	1432
2615000102	LA ARBOLEDA	1128.396	998.691	1255	2624700117	EL LUCERO	1107.468	990.959	1084
2615400107	QUEBRADANUEVA	1125.918	976.296	1082	2624700118	PATIO BONITO	1104.700	999.284	1647
2615400111	PIEDRAS DE MOLER	1136.071	1012.945	960	2624700120	LA DESPESA	1103.323	990.595	1352
2620000103	PLANTA RIO CAUCA	1064.958	872.965	956	2624700121	LA QUIEBRA	1106.854	1003.890	1466
2621500101	VILLACOLOMBIA	1046.488	844.184	1418	2625500102	EL AGUILA EL GRANARIO	1113.958	1033.741	1778
2621500103	EL SILENCIOSO	1038.214	838.708	1502	5301500102	LOBOGUERRERO	1045.940	907.727	721
2622100102	EL PALACIO	1061.537	855.515	964	5301500107	DAGUA	1043.305	895.668	892
2622100103	SAN VICENTE	1050.177	853.301	1488	5420410107	BETANIA	1085.370	982.351	1774
2622110101	LA ARGENTINA	1047.419	859.711	1898	5420410112	LA MARIA	1092.565	991.208	1500
2622320101	LA FONDA	1054.120	866.787	1295	5420410113	EL BOSQUE	1103.782	1009.039	1650
2622320103	ALTO DE LAS IGLESIAS	1050.727	864.791	1729	5420410114	ARABIA	1097.599	1002.416	1504
2622330101	COLEGIO SAN JUAN BOSCO	1060.471	872.417	983	5420410115	PUERTO NUEVO	1089.864	1000.060	1086
2622330102	CANAVERALEJO PG	1055.461	868.987	1097	5420410116	EL BALSAL	1093.749	1007.816	1588
2622330105	LOS CRISTALES	1056.568	870.803	1341	5420410117	EL CAIRO	1094.601	1017.713	1730
2622330106	LAS BRISAS	1054.355	868.346	1227	5420410118	MORAVITO	1077.222	967.135	1548

En las Figuras 5.1 y 5.2 se muestra las coordenadas de las estaciones climáticas y el diagrama de dispersión entre la localización y la altitud. De las 82 estaciones, 32 se encuentran en zona de valle (menor a 1,240 metros de altura) y 50 en zona de montaña (1,240 a 2,000 metros de altura).

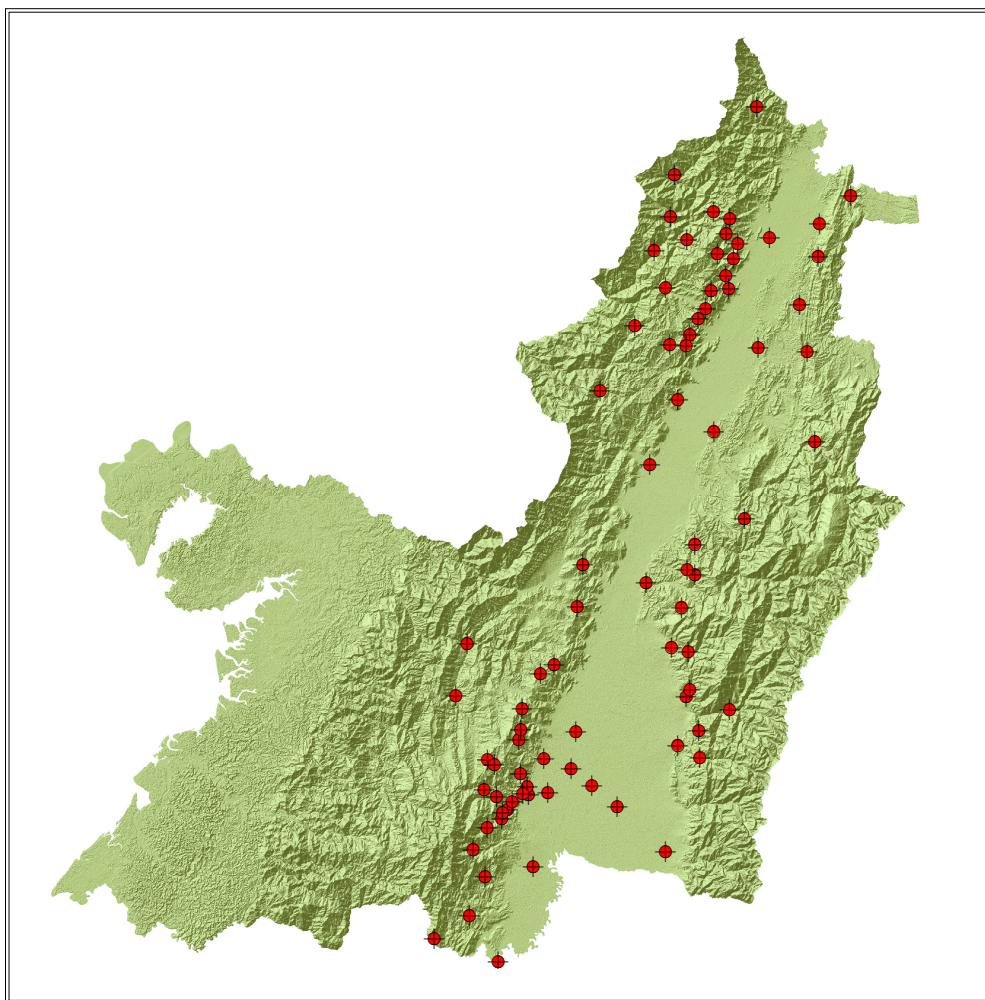


Figura 5.1: Localización de las estaciones climáticas.

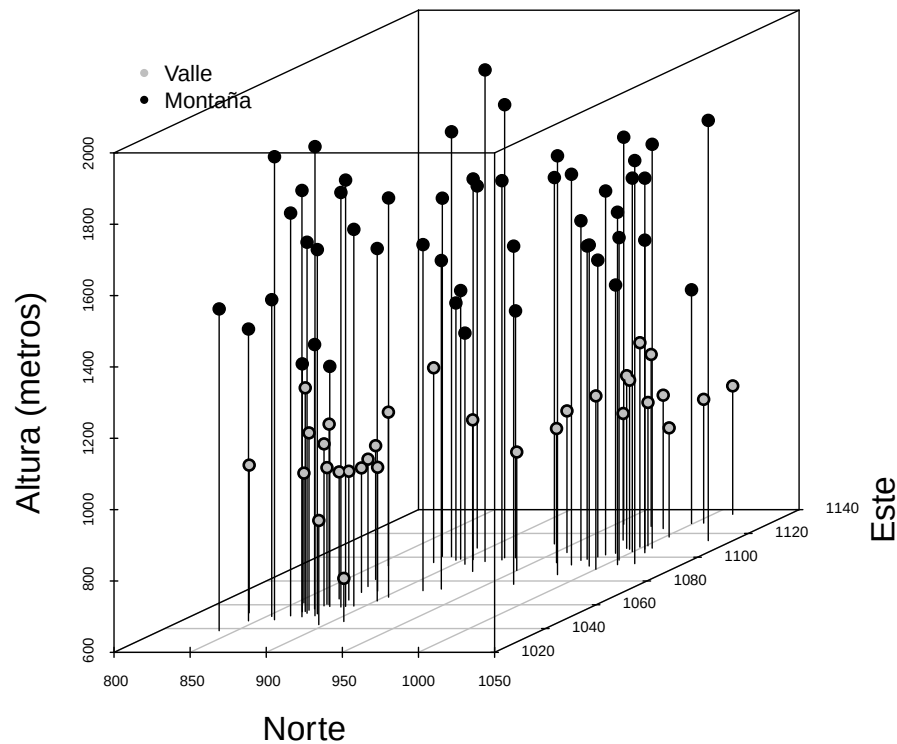


Figura 5.2: Diagrama de dispersión de las estaciones climáticas según altura.

Dado que se tuvieron problemas de datos faltantes y considerando que los datos fueron registrados en distintas localizaciones alrededor de la zona de estudio. Entonces se usó la Geoestadística univariada, en particular la utilización del Kriging Ordinario como método de predicción, para estimar los datos faltantes requeridos para completar la base de datos. A continuación, se presentan los aspectos teóricos de la Geoestadística univariada expuestos en Cressie (1993) y Giraldo (2002).

5.1. Geoestadística univariada

La metodología de estimación de la estructura de correlación espacial, conjuga una serie de metodologías para el análisis de datos de variables observadas en diferentes puntos del espacio de una región. Los datos espaciales se pueden clasificar en tres grupos fundamentales según el contexto de observación del que provienen: patrones espaciales, lattices y geoestadística (Cressie, 1993).

En la presente investigación, se trabajó con Geostatística. A continuación, se definen algunos términos importantes dentro del contexto de la Geoestadística.

5.1.1. Variable regionalizada

Una variable regionalizada es una variable medida en el espacio. Esta se puede definir como un proceso estocástico con un dominio en el espacio euclidiano d -dimensional R^d , $\{Z(x) : x \in D \subset R^2\}$. Si el dominio $d = 2$ es un plano, entonces $Z(x)$ es una variable medida en un punto x del plano.

5.1.2. Momentos de una variable regionalizada

A continuación se definen los momentos de primer y segundo orden de una variable regionalizada (Cressie, 1993, Giraldo, 2002)

- **Momentos de primer orden:** Si la función de distribución de $Z(x_i)$ tiene una media definida, será una función de localización x_i .

$$E[Z(x_i)] = m(x_i)$$

- **Momentos de segundo orden:** Si la varianza de $Z(x_i)$ existe, entonces se define como el momento de segundo orden y estará en función de la localización x_i .

$$V[Z(x_i)] = E[Z(x_i) - m(x_i)]^2 = \sigma_i^2$$

$$Cov[Z(x_i), Z(x_j)] = E[(Z(x_i) - m(x_i))(Z(x_j) - m(x_j))]$$

La función de variograma es definida como la varianza de la diferencia de $Z(x_i) - Z(x_j)$ (Cressie, 1993):

$$V[Z(x_i) - Z(x_j)] = 2\gamma(x_i, x_j)$$

donde la expresión,

$$\gamma(x_i, x_j) = \frac{1}{2}V[Z(x_i) - Z(x_j)] \quad (5.1)$$

se denomina función de semivarianza.

Una variable regionalizada presenta estacionariedad si su función de distribución conjunta no varía con respecto a un cambio en el vector h , es decir, $\vec{Z}(x) = [Z(x_1), Z(x_2), \dots, Z(x_n)]$ es igual a $\vec{Z}(x) = [Z(x_1 + h), Z(x_2 + h), \dots, Z(x_n + h)]$

5.1.3. Semivariograma experimental

Bajo el supuesto de que la varianza de los incrementos de una variable regionalizada es finita, la función denotada por $2\gamma(h)$ se denomina variograma. Entonces el variograma en términos de valores esperados es (Cressie, 1993):

$$\begin{aligned} 2\gamma(h) &= V [Z(x+h) - Z(x)] \\ &= E [(Z(x+h) - Z(x))^2] - [E (Z(x+h) - Z(x))]^2 \\ &= E [(Z(x+h) - Z(x))^2] \end{aligned}$$

Por lo tanto el variograma se define como la media aritmética de todos los cuadrados de las diferencias entre pares de valores experimentales separados por una distancia h (Giraldo, 2002).

La mitad del variograma $\gamma(h)$, se conoce como la función de semivarianza. La función de semivarianza estimada por el método de los momentos está definida por:

$$\gamma(h) = \frac{1}{2Np(h)} \sum_{i=1}^{Np(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (5.2)$$

donde:

$Np(h)$: número de pares de puntos a la distancia h .

h : es el incremento en la distancia.

$Z(x_i)$: son los valores de la variable en el punto x_i .

x_i : punto de medición de $Z(x_i)$

5.1.4. Semivariograma teórico

Después de construir el semivariograma experimental, es necesario ajustar a este un modelo teórico que permita determinar los parámetros del semivariograma para ser usados en la predicción (ver Figura 5.3).

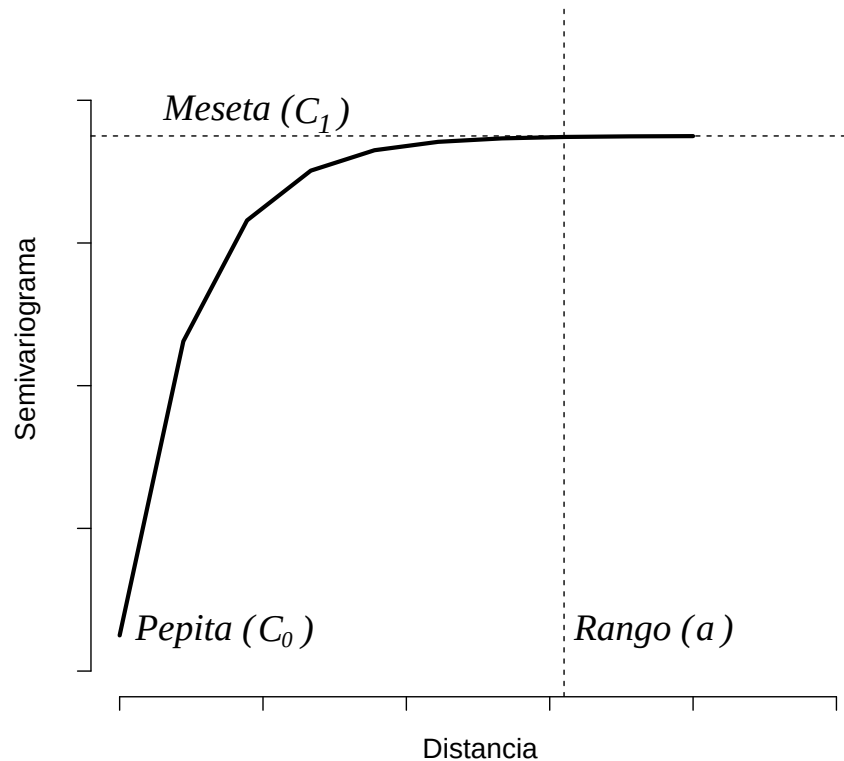


Figura 5.3: Componentes del semivariograma.

Los componentes del semivariograma son:

- **Pepita (Nugget):** representa el intercepto del semivariograma y es expresado como C_0 .
- **Meseta (Sill):** representa el valor de $\gamma(h)$, para el cual un aumento de h permanece constante y se expresa como C_1 .
- **Alcance (Range):** es la distancia h para la cual no hay correlación espacial entre $Z(x)$ y $Z(z + h)$ y se expresa por a .

Existen modelos teóricos de semivariogramas que pueden ajustarse al semivariograma experimental. A continuación se muestran algunos modelos, su forma y definición expuestas en Cressie (1993) y Giraldo (2002):

- **Modelo exponencial:** es utilizado cuando la dependencia espacial tiene un crecimiento exponencial con respecto a las distancias entre observaciones. Se expresa

teóricamente como:

$$\gamma(h) = C_0 + C_1 \left[1 - \exp \left\{ \frac{-h}{a} \right\} \right] \quad (5.3)$$

- **Modelo Gaussiano:** este modelo se caracteriza por alcanzar la meseta asintótica y el rango es definido como la distancia a la cual el variograma alcanza el 95 % de la meseta. Este modelo tiene la siguiente forma teórica:

$$\gamma(h) = C_0 + C_1 \left[1 - \exp \left\{ \frac{-h^2}{a^2} \right\} \right] \quad (5.4)$$

- **Modelo Esférico:** se caracteriza por tener un crecimiento rápido cerca al origen, sin embargo, los incrementos marginales van decreciendo para distancias grandes. El modelo esta expresado de la siguiente forma:

$$\gamma(h) = \begin{cases} C_0 + C_1 \left[1 - \frac{3}{2} \left(\frac{h}{a} \right) + \frac{1}{2} \left(\frac{h}{a} \right)^3 \right], & h \leq a \\ C_0 + C_1, & h > a \end{cases} \quad (5.5)$$

- **Modelo Matern:** Este modelo es flexible, varía de acuerdo al parámetro κ . Cuando $\kappa = 0.5$ coincide con el modelo exponencial.

$$\gamma(h) = \frac{1}{2^{\kappa-1} \Gamma(\kappa)} \left(\frac{h}{a} \right)^{\kappa} K_{\kappa} \left(\frac{h}{a} \right) \quad (5.6)$$

donde $K_{\kappa} \left(\frac{h}{a} \right)$ es una función de Bessel modificada de la segunda clase de orden κ .

5.1.5. Predicción espacial (Kriging)

La metodología de la predicción espacial es conocida como Kriging, esta consiste en hacer inferencias sobre datos no observados de un proceso aleatorio en el espacio. Este método fundamenta las estimaciones a partir de la minimización del error cuadrático medio de predicción. Kriging es un método de estimación que da la mejor estimación lineal insesgada de los valores de los puntos, esto es, elegir el promedio ponderado de los valores de las muestras la cual tenga la mínima varianza (Armstrong, 1998), (Giraldo, 2002).

Existen diferentes variaciones del método kriging, entre ellas están:

- Kriging Simple (SK).
- Kriging Ordinario (OK).
- Kriging Universal (UK).

A continuación se presenta el marco teórico del método de predicción utilizado para estimar los datos faltantes de precipitación diaria de estaciones climáticas:

- **Kriging Ordinario (OK):** se desea estimar un valor $z(x_0)$ en x_0 , usando los valores de los datos $Z(x_\alpha)$, a partir de n puntos muestrales vecinos x_α y combinándolos linealmente con pesos ω_α , i.e .

$$Z^*(x_0) = \sum_{\alpha=1}^n \omega_\alpha Z(x_\alpha) \quad (5.7)$$

Los pesos son elegidos de tal manera que el estimador sea:

- Insesgado: $E[Z^*(x_0) - Z(x_0)] = 0$ y de
- Varianza mínima: $V[Z^*(x_0) - Z(x_0)]$ es mínima.

El insesgamiento del estimador es garantizada bajo la restricción que la suma de los pesos sea 1, esto es, $\sum_{\alpha=1}^n \omega_\alpha = 1$

$$\begin{aligned} E[Z^*(x_0) - Z(x_0)] &= E\left[\sum_{\alpha=1}^n \omega_\alpha Z(x_\alpha) - Z(x_0) \cdot \sum_{\alpha=1}^n \omega_\alpha\right] = 0 \\ &= \sum_{\alpha=1}^n \omega_\alpha E[Z(x_\alpha) - Z(x_0)] = 0 \end{aligned}$$

La varianza de la estimación $\sigma_E^2 = V[Z^*(x_0) - Z(x_0)]$ es la varianza de la combinación lineal

$$Z^*(x_0) - Z(x_0) = \sum_{\alpha=1}^n \omega_\alpha Z(x_\alpha) - 1 \cdot Z(x_0) = \sum_{\alpha=0}^n \omega_\alpha Z(x_\alpha)$$

con un peso ω_0 igual a -1 y $\sum_{\alpha=0}^n \omega_\alpha = 0$.

La varianza de la estimación,

$$\begin{aligned}\sigma_E^2 &= E[(Z^*(x_0) - Z(x_0))^2] \\ &= 2 \sum_{\alpha=1}^n \omega_{\alpha} \gamma(x_{\alpha} - x_0) - \sum_{\alpha=1}^n \sum_{\beta=1}^n \omega_{\alpha} \omega_{\beta} \gamma(x_{\alpha} - x_{\beta}) - \gamma(x_0 - x_0)\end{aligned}$$

Minimizando la estimación de la varianza con la restricción sobre los pesos $\left[V[Z^*(x_0) - Z(x_0)] - 2\mu_{OK} \left(\sum \omega_{\alpha}^{OK} - 1\right)\right]$, se obtiene el sistema de kriging ordinario (OK):

$$\begin{pmatrix} \gamma(x_1 - x_1) & \cdots & \gamma(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n - x_1) & \cdots & \gamma(x_n - x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \omega_1^{OK} \\ \vdots \\ \omega_n^{OK} \\ \mu_{OK} \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \end{pmatrix}$$

La varianza de la estimación del kriging ordinario es,

$$\sigma_{OK}^2 = \mu_{OK} - \gamma(x_0 - x_0) + \sum_{\alpha=1}^n \omega_{\alpha}^{OK} \gamma(x_{\alpha} - x_0) \quad (5.8)$$

Finalmente el estimador kriging ordinario es un interpolador exacto, es decir, que si x_0 es idéntico con una localización de los datos entonces el valor estimado es el mismo con el valor del dato en ese punto (Giraldo, 2002)

$$Z^*(x_0) = Z(x_{\alpha}), \text{ si } x_0 = x_{\alpha}$$

5.2. Resultados de la predicción de datos faltantes

Con las estaciones que tenían datos faltantes se realizó el proceso de estimación, interpolando a partir de geoestadística clásica, para esto se hizo un análisis estructural con la precipitación promedio anual de las 82 estaciones y posteriormente se ajustaron los modelos teóricos (Exponencial, Esférico, Gausiano y Matern) expuestos anteriormente.

En la Tabla 5.4 se muestran los resultados de la estimación de los parámetros de los semivariogramas muestrales por año, en el período 1993-2011. Con estos parámetros fue ajustado un modelo de semivariograma teórico para cada año.

Tabla 5.4: Parámetros estimados de semivariograma por año, en el período 1993-2011.

Año	C_1	a	C_0
1993	21.74	0.98	0.67
1994	24.5	0.25	0.24
1995	27.63	1.26	0.5
1996	19.19	0.71	0.42
1997	14.33	1.65	0.59
1998	21.91	1.51	1.34
1999	24.07	0.21	0.36
2000	35.49	0.84	0.05
2001	41.47	1.49	1.26
2002	31.92	0.77	0.08
2003	31.88	0.58	0.51
2004	28.37	0.73	0.99
2005	33.81	0.04	0.63
2006	32.52	0.23	0.39
2007	30.24	0.05	0.66
2008	33.52	0.38	0.02
2009	56.94	1.54	0.85
2010	50.29	0.02	1.22
2011	25.55	0.12	0.23

Para seleccionar el modelo óptimo, se utilizó el criterio del mínimo Error Cuadrático Medio (ECM) obtenido de la estimación de los residuales a partir de validación cruzada. Finalmente, con el modelo seleccionado se realizó el proceso de estimación de las observaciones faltantes utilizando Kriging Ordinario para cada año. En la Tabla 5.5 se muestra los resultados obtenidos y se resalta el modelo de semivariograma con menor ECM en el período 1993-2011.

Tabla 5.5: Comportamiento del ECM según modelo ajustado, en el período 2003-2011.

Año	Exponencial	Esférico	Gaussiano	Mattern
1993	1.82	0.77	0.48	0.44
1994	3.53	2.69	4.46	3.42
1995	1.22	2.05	2.16	3.8
1996	3	2.13	5.67	3.98
1997	0.57	2.97	4.32	0.45
1998	1.51	1.67	4.51	2.15
1999	2.23	4.63	3.56	2.46
2000	6.53	5.86	6.78	6.08
2001	7.23	6.69	5.41	7.48
2002	7.4	5.87	2.34	6.28
2003	4.22	5.6	2.63	4.43
2004	5.91	4.17	4.43	6.16
2005	7.1	5.89	5.4	7.04
2006	10.26	8.37	9.98	10.04
2007	5.87	6.52	5.78	4.08
2008	9.05	6.03	13.07	8.2
2009	10.16	11.31	11.89	9.77
2010	13.47	12.19	12.54	11.94
2011	13.16	12.76	10.5	11.15

5.3. Resultados del análisis exploratorio de datos completos

Esta sección comprende el análisis exploratorio de los datos observados procedentes de estaciones climáticas y de satélite, después de realizar el proceso de estimación de los datos faltantes en la información proveniente de las estaciones climáticas.

Se observó que los datos reflejaron en forma clara los períodos secos y húmedos que indican la estacionalidad de la precipitación en el transcurso del año, de los datos procedentes de estaciones climáticas y de satélite, sin embargo, los datos provenientes de

estaciones climáticas presentaron un rango de variación mayor a los datos procedentes de satélite, en este último se observa claramente la estacionalidad de precipitación, este comportamiento puede ser atribuido a que los datos de estaciones presentaron mayor variabilidad, debido a que en muchas estaciones climáticas no hubo registro de precipitación o por el contrario hubo registros que sobrepasaron el promedio diario (Figura 5.4). Por otro lado, los datos de satélite mostraron mayor homogeneidad en el comportamiento de la precipitación debido a que es una estimación uniforme en toda la región de estudio (Figura 5.5).

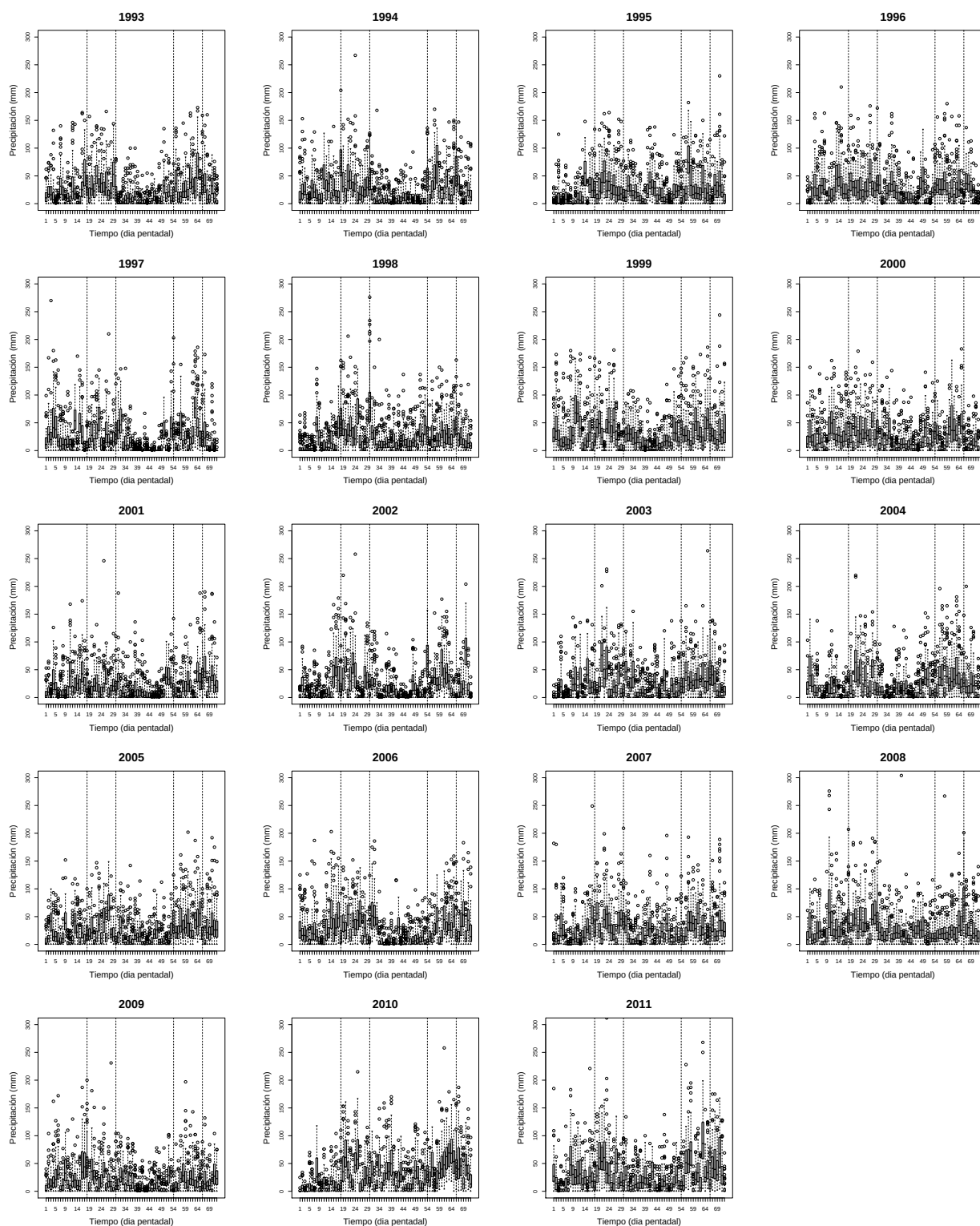


Figura 5.4: Comportamiento de la precipitación pentadal de estaciones climáticas sin acumular, en el período 1993-2011.

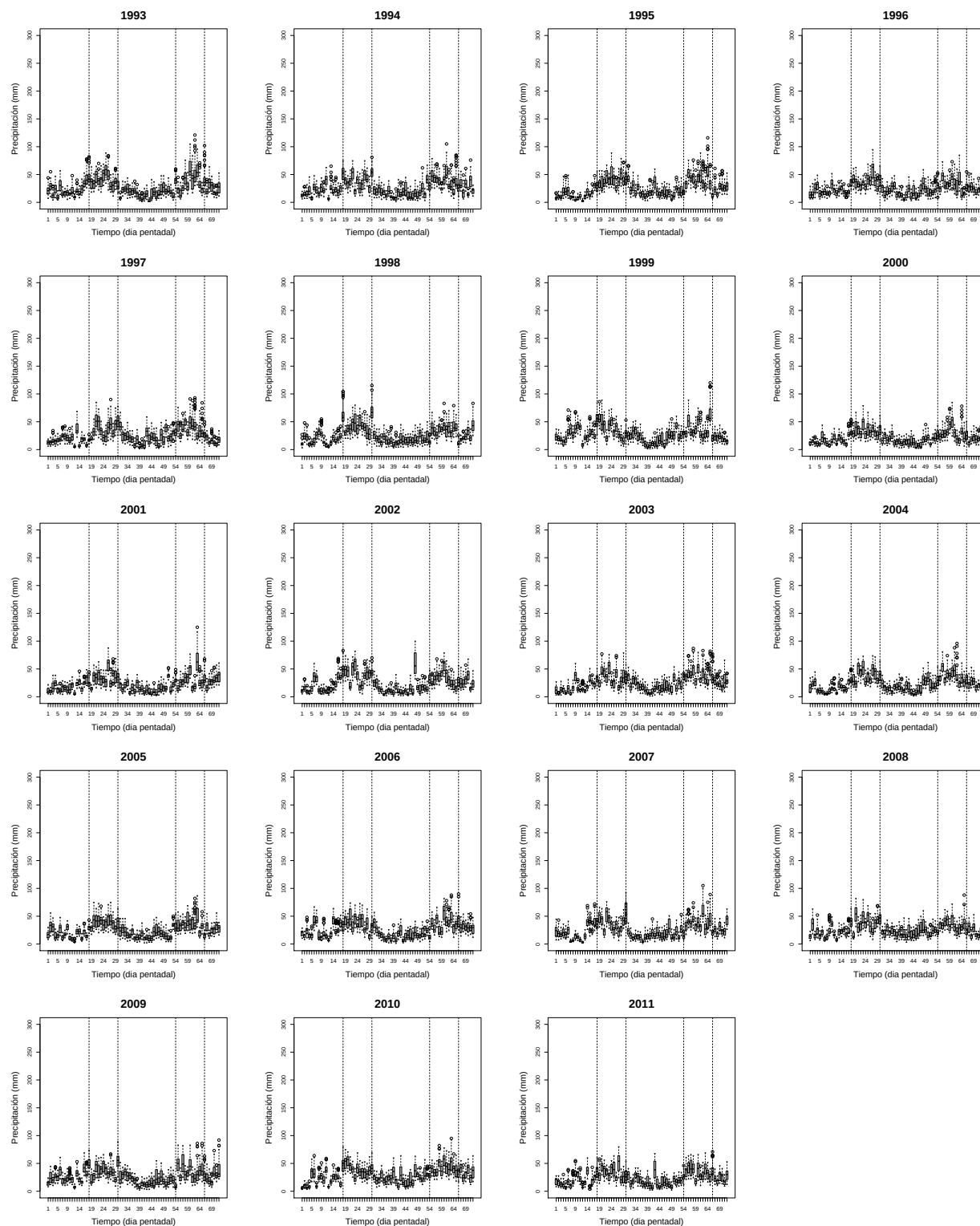


Figura 5.5: Comportamiento de la precipitación pentadal de satélite sin acumular, en el período 1993-2011.

Con el objetivo de analizar de forma exploratoria el nivel de asociación entre los datos estaciones climáticas y de satélite, en la Tabla **5.6** se muestra el comportamiento de las estadísticas descriptivas de la correlación lineal entre los datos observados de precipitación de estaciones climáticas y de satélite de las 82 estaciones climáticas por año. Se denota la existencia de correlaciones entre 0.1 a 0.39 y correlaciones máximas alrededor de 0.79 a 0.89. También se observa que la correlación promedio estuvo entre 0.33 a 0.67, evidenciando una relación intermedia y baja entre los conjuntos de datos mencionados.

Tabla 5.6: Estadísticas descriptivas de la correlación lineal entre los datos de estaciones climáticas y de satélite, en el período 1993-2011

Año	Mínimo	Máximo	Promedio	Mediana	Desviación estándar
1993	0.28	0.69	0.43	0.42	0.19
1994	0.29	0.68	0.42	0.45	0.2
1995	0.17	0.55	0.42	0.41	0.29
1996	0.14	0.65	0.35	0.49	0.14
1997	0.06	0.5	0.35	0.35	0.2
1998	0.11	0.63	0.57	0.6	0.26
1999	0.27	0.64	0.51	0.56	0.14
2000	0.19	0.67	0.38	0.41	0.13
2001	0.17	0.64	0.29	0.31	0.23
2002	0.38	0.89	0.67	0.6	0.17
2003	0.28	0.72	0.6	0.61	0.27
2004	0.37	0.69	0.61	0.67	0.22
2005	0.24	0.65	0.46	0.58	0.27
2006	0.21	0.72	0.49	0.61	0.25
2007	0.18	0.79	0.52	0.56	0.13
2008	0.14	0.67	0.33	0.42	0.3
2009	0.15	0.64	0.34	0.37	0.24
2010	0.39	0.6	0.57	0.47	0.23
2011	0.11	0.74	0.59	0.46	0.17

Adicionalmente, se contrastó la correlación de las 82 observaciones de estaciones y de satélite por año ($H_0 : \rho = 0$ Vs $H_0 : \rho \neq 0$). En la Tabla **5.7**, se observa que los porcentajes de rechazo siempre estuvieron por encima del 80 %, es decir, que existe una asociación lineal entre los datos de estaciones climáticas y de satélite en todo el período de estudio.

Tabla 5.7: Porcentaje de la existencia de asociación lineal entre los datos de estaciones climáticas y de satélite, en el período 1993-2011.

Año	n	%	Año	n	%
1993	67	81.71	2003	78	95.12
1994	79	96.34	2004	81	98.78
1995	75	91.46	2005	77	93.9
1996	69	84.15	2006	79	96.34
1997	50	60.98	2007	79	96.34
1998	72	87.8	2008	61	74.39
1999	77	93.9	2009	62	75.61
2000	64	78.05	2010	81	98.78
2001	56	68.29	2011	80	97.56
2002	82	100			

Un primer acercamiento para el ajuste de un modelo de regresión lineal es realizar una exploración de la relación entre la variable respuesta y la variable explicativa. Con respecto a esto, los datos de estaciones climáticas y de satélite acumulados, mostraron una relación directa pero débil al principio del año, sin embargo, en los pentadales finales la relación es más fuerte (Figura 5.7).

5.4. Construcción de las curvas pentadales de estaciones climáticas

Para trabajar con la misma escala temporal entre ambas fuentes de información, fue necesario pasar los datos de precipitación diaria a precipitación acumulada cada cinco días (dato pentadal).

Sea x_i el dato de precipitación en el día i , entonces: $p_1 = \sum_{i=1}^5 x_i, p_2 = \sum_{i=6}^{10} x_i, \dots, p_{72} = \sum_{i=360}^{365} x_i$ representa el pentadal j , donde $j = 1, \dots, 72$. De esta manera se generaron 72 datos pentadales para cada año.

Sin embargo, la información del sensor remoto además de presentar una escala pentadal, contiene datos acumulados de imágenes de precipitación pentadal. Por lo tanto, el proceso generador inicial tiene la siguiente forma funcional (línea azul clara), al tener dicha estructura funcional creciente en cada pentadal, se tomó la decisión de trabajar con las curvas acumuladas de precipitación (ver Figura 5.6).

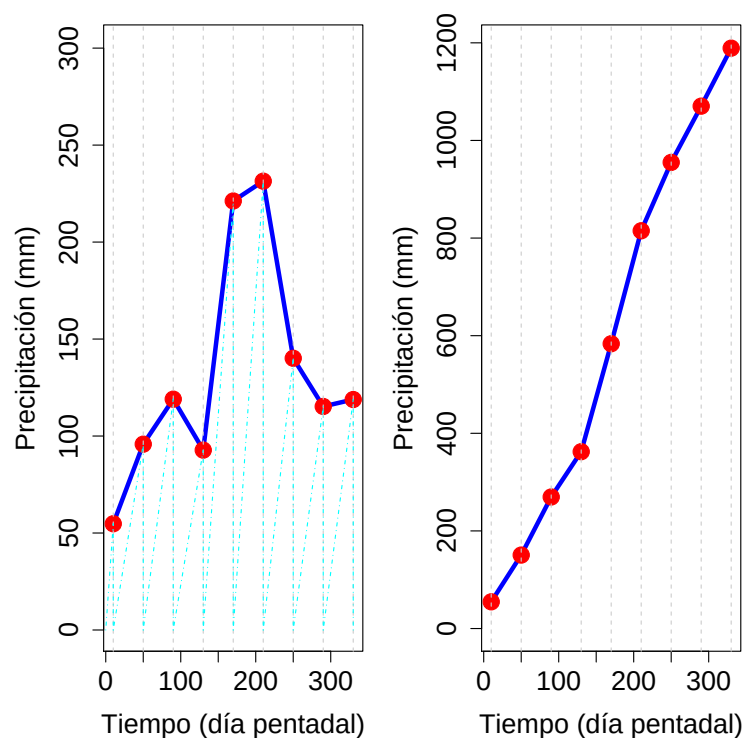


Figura 5.6: Ilustración del proceso generador de la curva pentadal de precipitación.

Las curvas observadas de las estaciones climáticas presentaron mayor variabilidad que las curvas observadas de satélite en todo los años (Figura 5.8). Esto puede deberse a que los datos de estaciones climáticas tienen de manera implícita la variabilidad espacial debida a la cercanía o lejanía entre las estaciones, mientras que los datos observados de satélite no presentaron tanta variabilidad (Figura 5.9), sin embargo, son mas uniformes entre los puntos ya que proviene de un proceso de estimación homogéneo en todo el Valle del Cauca.

Con el objetivo de caracterizar la influencia de la altura de las estaciones con respecto al comportamiento de los valores de precipitación acumulada. Se encontró que la curva media de precipitación de estaciones fue mayor en la zona de montaña en todo el período de análisis (Figura 5.10 y Figura 5.12), en contraste, la curva media de precipitación de satélite no presentó diferencias importantes entre las zonas de valle y montaña en los 19 años (Figura 5.11 y Figura 5.13). Este resultado sugiere que la información proveniente de satélite no tiene en cuenta la orografía de la zona de estudio.

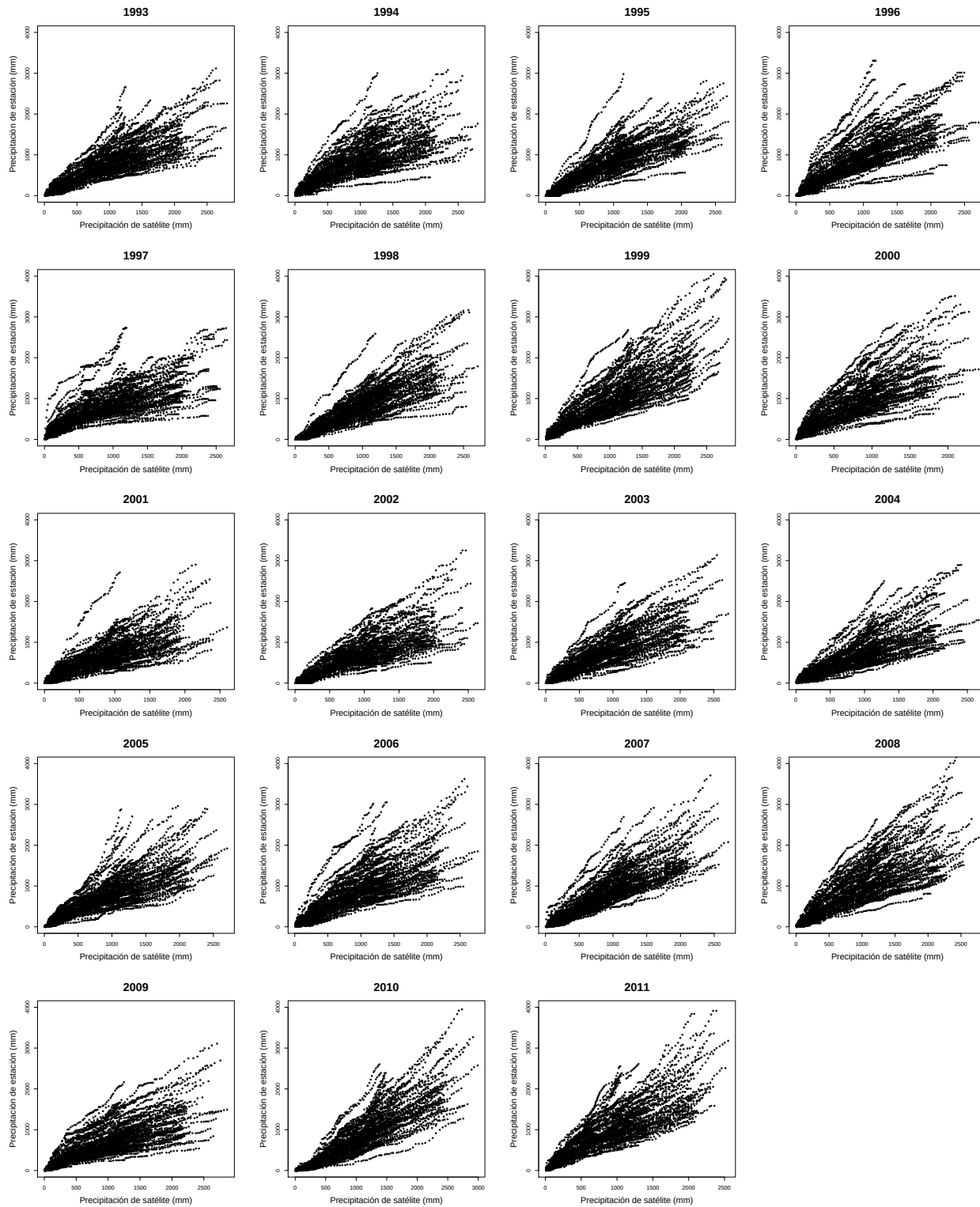


Figura 5.7: Relación entre los datos observados de precipitación de estaciones climáticas y de satélite, en el período 1993-2011.

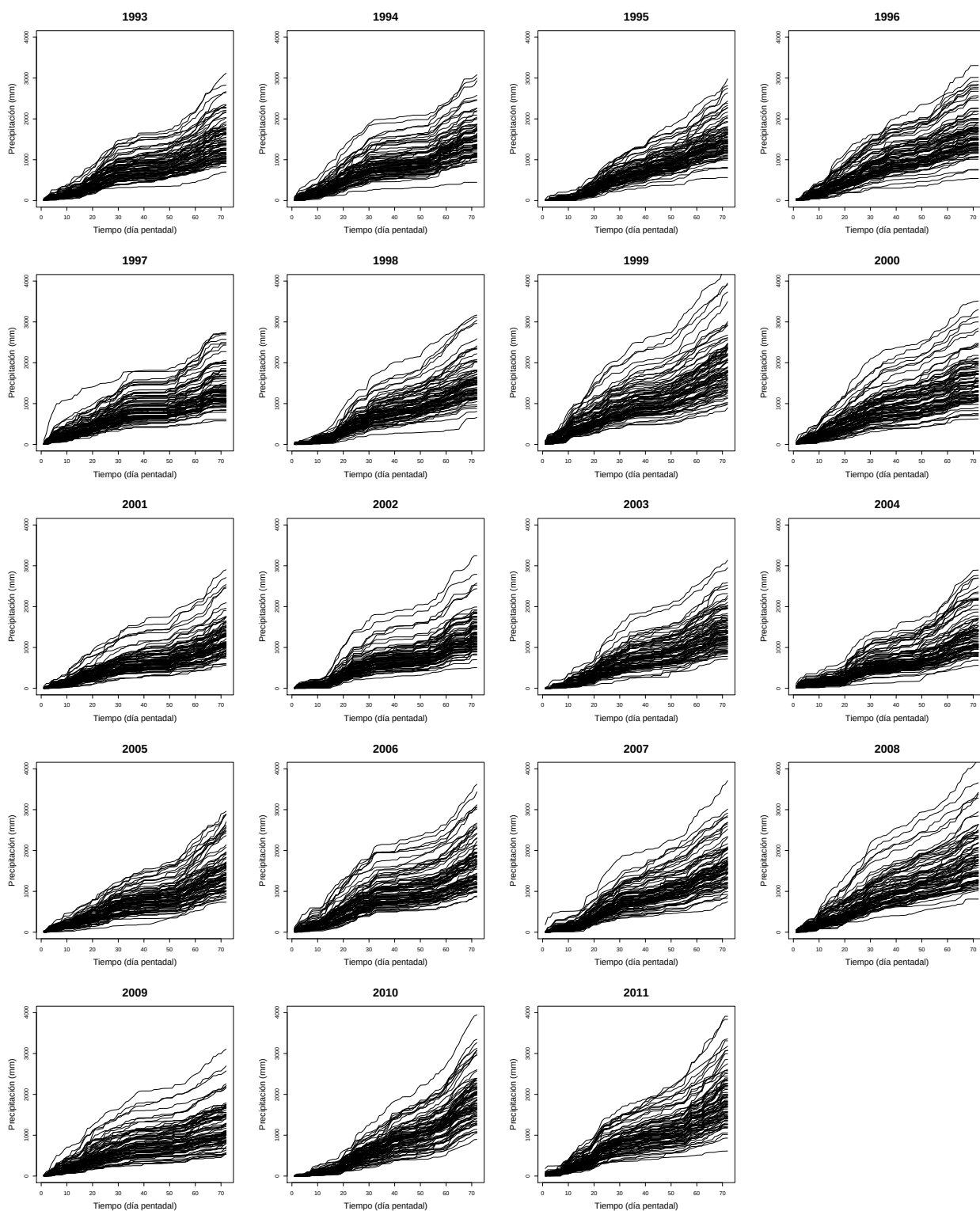


Figura 5.8: Precipitación de estaciones climáticas, en el período 1993-2011.

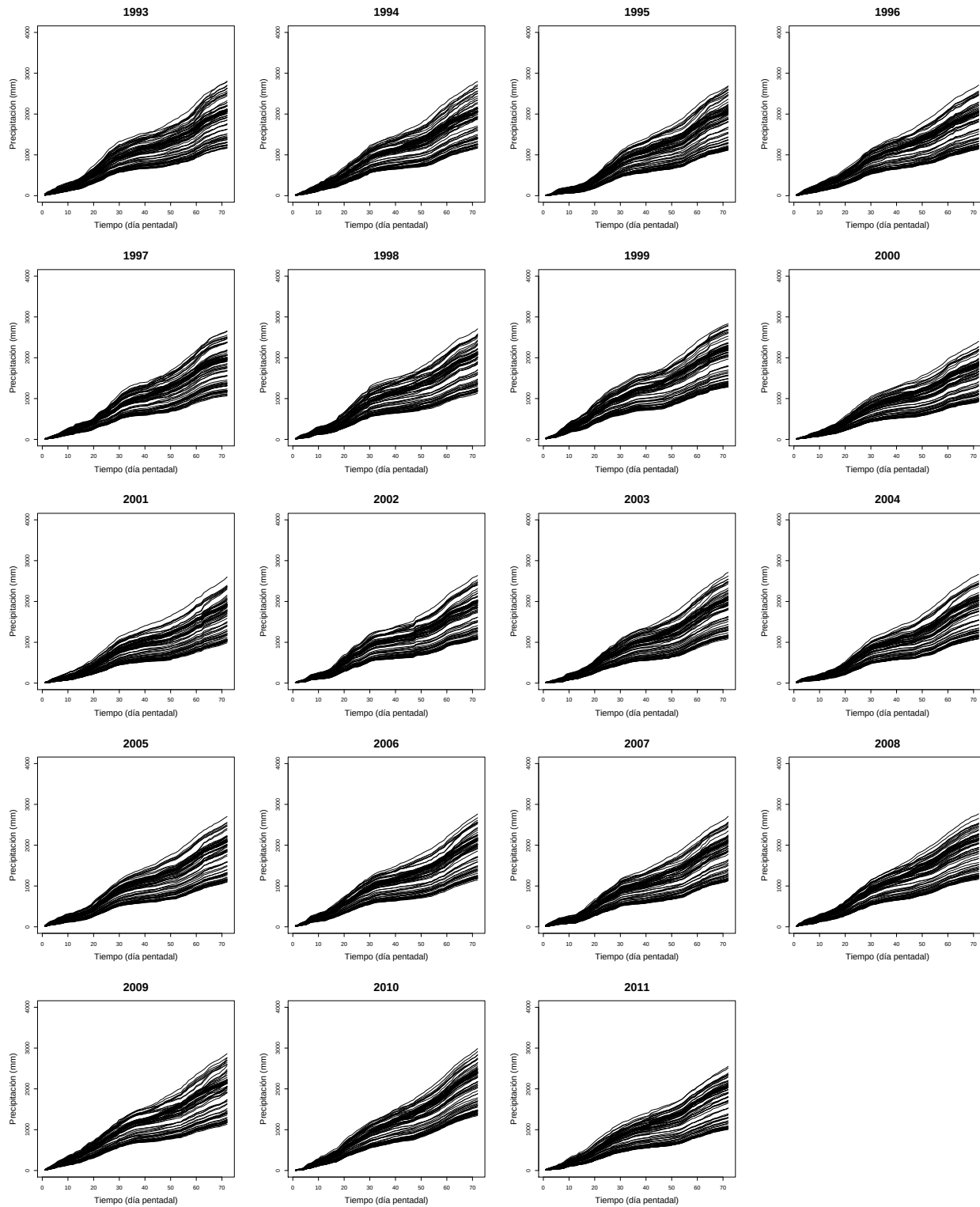


Figura 5.9: Precipitación de satélite, en el período 1993-2011.

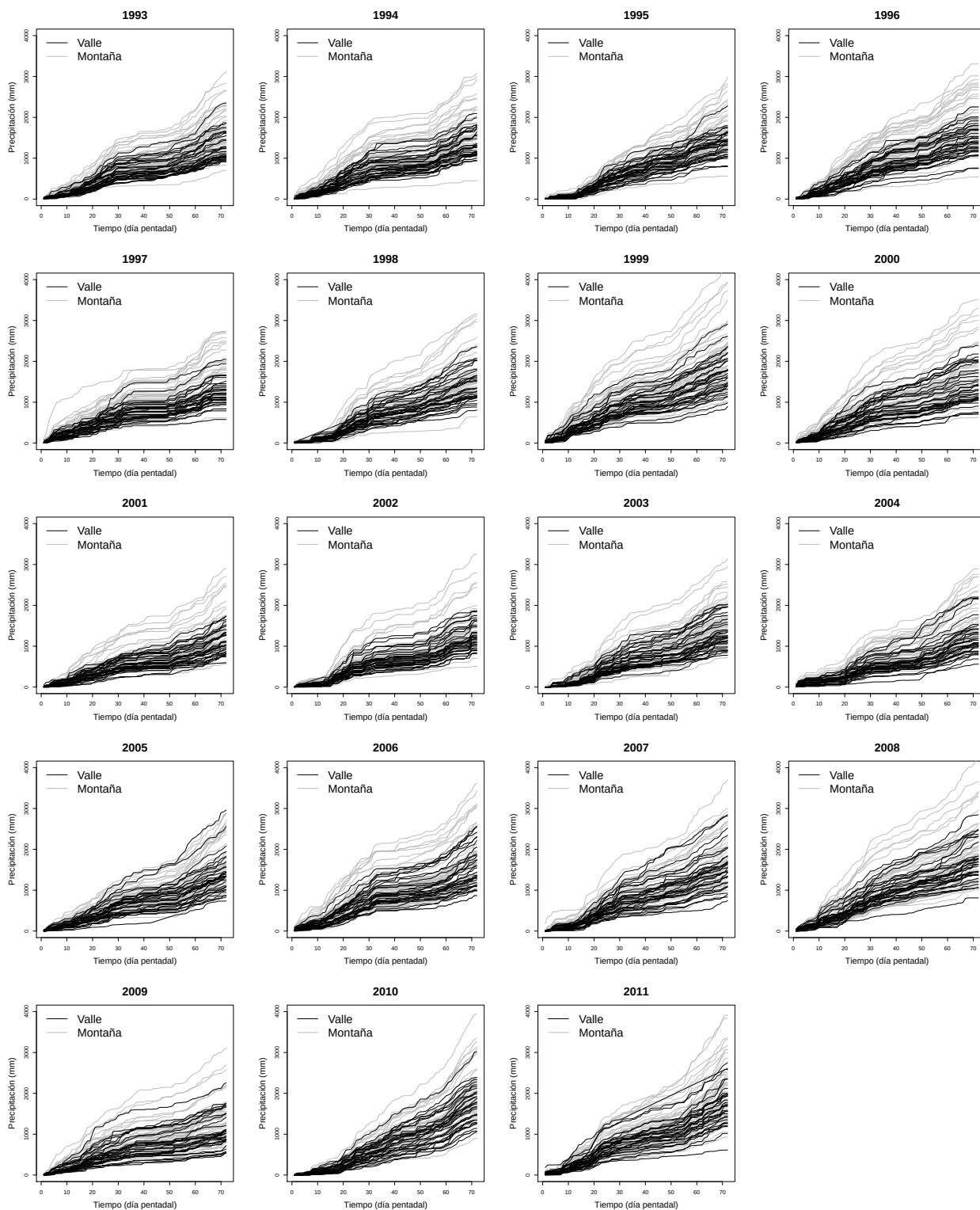


Figura 5.10: Precipitación de estaciones climáticas por región geográfica, en el período 1993-2011.

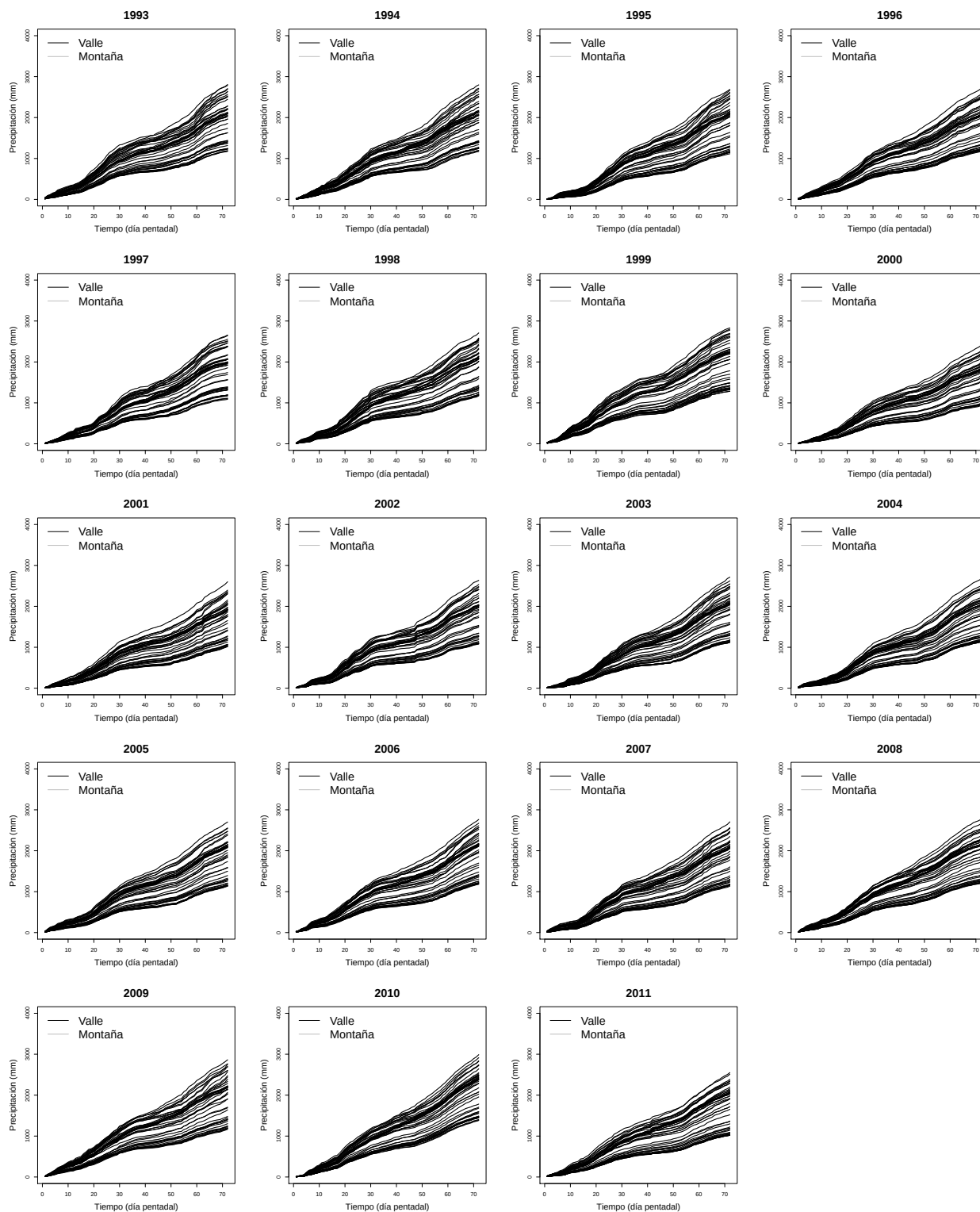


Figura 5.11: Precipitación de satélite por región geográfica, en el período 1993-2011.

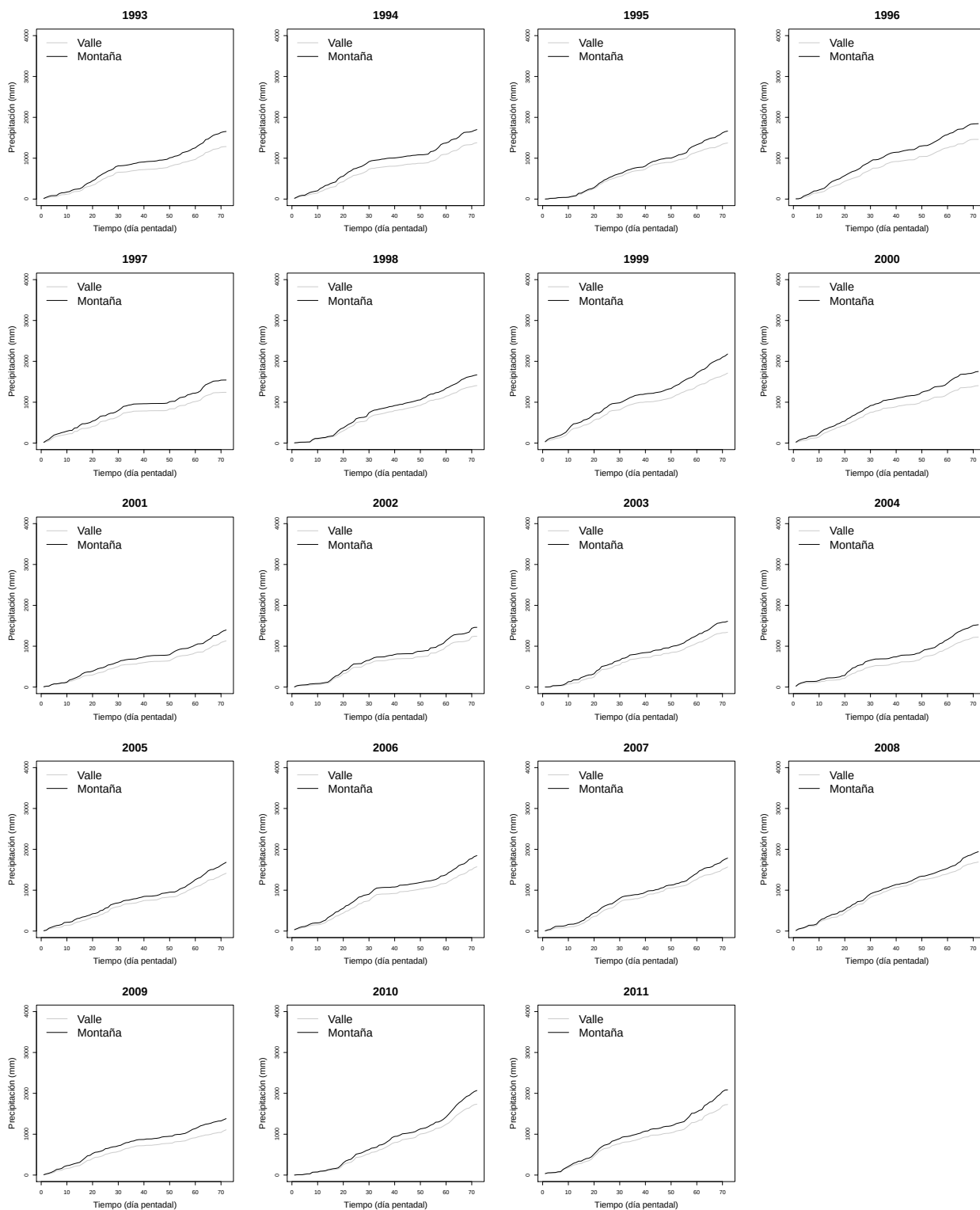


Figura 5.12: Precipitación promedio de estaciones climáticas por región geográfica, en el período 1993-2011.

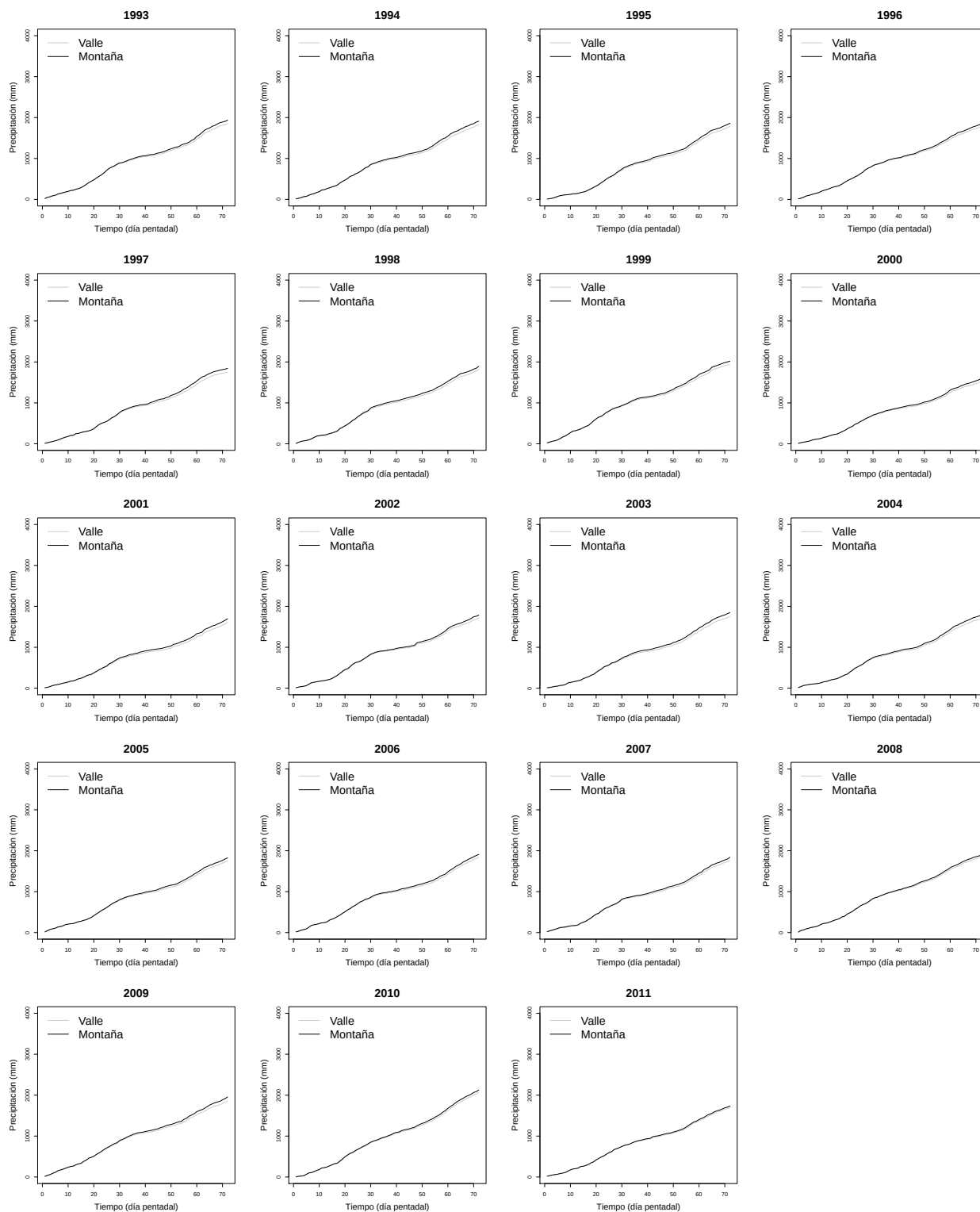


Figura 5.13: Precipitación promedio de satélite por región geográfica, en el período 1993-2011.

5.5. Conclusiones

En general, los datos de precipitación puntual mostraron un comportamiento que refleja el régimen bimodal de la precipitación característico de la región (Figura 5.4 y Figura 5.5), fácilmente denotado para los datos de satélite, no obstante, este comportamiento en los datos provenientes de estaciones climáticas es menos evidente debido a la variabilidad que presentan.

Las cantidades de precipitación registradas en las estaciones son mucho mayores a las registradas por el satélite. Lo anterior se debe a que las estaciones están midiendo constantemente la cantidad de precipitación a lo largo del día, mientras que el valor de precipitación de satélite se limita al momento en que el sensor captura la imagen.

Se encontraron asociaciones lineales importantes entre las 82 observaciones de estaciones climáticas y de satélite (Tabla 5.7) con un porcentaje de asociaciones significativas superiores al 50 % en todos los años. Finalmente, se observó una relación directa entre ambas fuentes de información para todo el período de estudio (Figura 5.7).

En cuanto al análisis de la precipitación por zona geográfica, se observaron aumentos en los valores de precipitación en la zona de montaña para los datos de estaciones climáticas (Figura 5.10), sin embargo, este comportamiento no fue observado con los datos provenientes de satélite (Figura 5.11).

Capítulo 6

Análisis descriptivo de datos funcionales

En este capítulo se hace un resumen de la bibliografía revisada de los conceptos preliminares del ADF. Ésta es una rama de la estadística, que estudia y analiza la información contenida en curvas, superficies, o cualquier elemento que varía sobre un continuo, generalmente el tiempo (Ramsay & Silverman, 2005). Las particularidades de estos objetos, las funciones, hacen que surja todo un aparato de definiciones, teoremas y herramientas para poder estudiarlas correctamente (Manteiga & Vieu, 2007). Por ello, el primer paso será definir y comentar los conceptos básicos, las características y peculiaridades de este tipo de datos.

La explosión tecnológica ha hecho del ADF una disciplina emergente dentro de la estadística con multitud de publicaciones y estudios. Prueba de este interés es la creciente presencia en revistas especializadas y la creación de un grupo de trabajo específico dentro del ERCIM (European Research Consortium for Informatics and Mathematics). Entre la literatura disponible sobresalen como referencias básicas los libros de Ramsay & Silverman (2005) y Ferraty & Vieu (2006). En el primer caso, tratan muchos de los problemas básicos de la estadística funcional, en el segundo caso abordan la metodología desde el punto de vista no-paramétrico. Ramsay y Silverman también tienen un segundo libro de carácter más aplicado en el que se estudian soluciones a problemas sobre conjuntos de datos concretos (Ramsay & Silverman, 2002).

6.1. Conceptos preliminares

Antes de analizar las propiedades y las dificultades del ADF es necesario establecer algunos conceptos de una manera más formal. Para ello se ha estudiado las referencias del tema en Ramsay & Silverman (2005) y Ferraty & Vieu (2006).

En primer lugar se debe definir el concepto de variable funcional. En general, cualquier observación que varíe de forma continua se puede concebir por un dato funcional, desde un electrocardiograma hasta la precipitación de una región. En la práctica, la información es recogida por dispositivos que toman muestras de una determinada variable aleatoria en distintos momentos de tiempo en un rango específico (t_{min}, t_{max}) . De esta forma, una observación puede expresarse mediante la familia $\{X(t_j)\}_{j=1, \dots, J}$. En la actualidad, los avances en los sistemas de información hacen posible hacer la periodicidad mas corta en la que es registrado el dato, siendo este uno de los motivos importantes el uso del ADF, de modo que se podría considerar estas muestras como observaciones es un espacio continuo $\chi = \{X(t); t \in (t_{min}, t_{max})\}$. Por lo tanto, una variable funcional es aquella variable aleatoria χ que toma valores en un espacio funcional, es decir, de dimension infinita, donde una observación de χ se denomina un dato funcional (Ferraty & Vieu, 2006).

Por otro lado son necesarias herramientas para extraer la información de las funciones. Estas herramientas vienen supeditadas por el espacio de la función.

- **Espacio Funcional:** Un espacio funcional $\mathcal{F}$, es un espacio hecho de funciones. Cada función en el espacio puede ser pensado como un punto (Tomaso Poggio & Frogner, 2010), por ejemplo:

1. $C[a, b]$, el conjunto de todas las funciones continuas con valores reales en el intervalo $[a, b]$.
2. $L_1[a, b]$, el conjunto de todas las funciones reales cuyo valor absoluto es integrable en el intervalo $[a, b]$.
3. $L_2[a, b]$, el conjunto de todas las funciones con valores reales cuadrado integrable en el intervalo $[a, b]$.

A continuación se presenta las definiciones y propiedades de distancia, precisados en Kolmogorov et al. (1957) para entender la métrica entre dos elementos en el espacio funcional L_2 .

Definición 1

Sea $\mathcal{F}$ un espacio funcional, es un espacio métrico si en $\mathcal{F}$ existe una distancia $d : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$ que $\forall f, g$ de $\mathcal{F}$ cumple:

1. $d(f, g) = 0$ y $d(f, g) = 0$ si y sólo si $f = g$, la distancia por 0 es 0.
2. $d(f, g) = d(g, f)$. Es decir que es simétrica.
3. $d(f, g) \leq d(f, h) + d(h, g)$. Es decir que la suma de las distancias de dos lados cualquiera es siempre mayor a la longitud del lado restante.

Definición 2

Sea $\mathcal{F}$ un espacio vectorial funcional, es también un espacio normado, si en $\mathcal{F}$ se encuentra definida su norma $\| \cdot \|: \mathcal{F} \rightarrow \mathbb{R}$ que $\forall f, g$ de $\mathcal{F}$ cumple:

1. $\| f \| \geq 0$ y $\| f \| = 0$ si, y sólo si, $f = 0$
2. Para todo λ , $\| \lambda f \| = |\lambda| \| f \|$
3. $\| f + g \| \leq \| f \| + \| g \|$.

Por lo tanto, todo espacio normado se puede convertir en un espacio métrico con la distancia inducida por la norma $d(f, g) = \| f - g \|$

Definición 3

Sea $\mathcal{F}$ un espacio funcional, este es un espacio euclidiano sobre $\mathbb{K}$ ($\mathbb{R}$ ó $\mathbb{C}$) si en $\mathcal{F}$ existe el producto escalar $\langle \cdot, \cdot \rangle: \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{K}$, que para cualquier elementos f, g, h de $\mathcal{F}$ se cumple:

1. $\langle f, f \rangle \geq 0$
2. Para todo $a, b \in \mathbb{K}$ $\langle af + bg, h \rangle = a \langle f, h \rangle + b \langle g, h \rangle$
3. $\langle f, g \rangle = \overline{\langle g, f \rangle}$

De igual forma, si el espacio es completo, es decir que toda sucesión de Cauchy converge con la norma, se dice que es un espacio de Bannach (Kolmogorov et al., 1957).

Todo espacio euclidiano se convierte en un espacio normado con la norma inducida por el producto escalar, $\| f \| = \sqrt{\langle f, f \rangle}$. De esta manera, si el espacio es además completo se dice que es un espacio de Hilbert. Esto es un caso particular de los $L - espacios$ definidos como $L_p(S, \mu) = \left\{ f : S \rightarrow \mathbb{K} \mid f \text{ medible } \| f \|_p^p = \int |f|^p d\mu < \infty \right\}$ donde (S, μ) es el espacio de medida. Estos espacios cuando $p = 2$ son los espacios de Hilbert o también conocidos como los espacios cuadrado integrable (Noguerales et al., 2010).

6.2. Base de funciones

En la práctica, los datos funcionales son obtenidos de forma discreta, siendo necesario convertirlos a funciones sobre las cuales se pueda trabajar, para esto pueden emplearse diferentes tipos de ajustes. Teniendo en cuenta lo anterior, la representación de estos datos como datos funcionales a partir de una base ortonormal brinda una ventaja tanto teórica como práctica, esta base se utiliza para transformar datos discretos en datos funcionales.

A continuación se presenta una revisión teórica del concepto de bases de funciones y las bases más utilizadas para la construcción de datos funcionales. (Ramsay & Silverman, 2002, 2005, Ferraty & Vieu, 2006, Olaya, 2012)

- **Bases:** es un conjunto de funciones conocidas $\{\phi_k\}_{k \in \mathbb{N}}$ tales que cualquier función puede ser aproximada, tan bien como se requiera, mediante una combinación lineal de K de ellas con K suficientemente grande.
- **Bases de Fourier:** la extensión de Fourier de $\chi(t)$ es de la forma

$$\hat{\chi}(t) = c_0\phi_0 + \sum_r c_{2r-1}\phi_{2r-1}(t) + c_{2r}\phi_{2r}(t)$$

donde $\phi_0(t) = \frac{1}{\sqrt{T}}$, $\phi_{2r-1}(t) = \frac{\sin(r\omega t)}{\sqrt{T/2}}$, $\phi_{2r}(t) = \frac{\cos(r\omega t)}{\sqrt{T/2}}$ forman una base periódica de período $T = \frac{2\pi}{\omega}$.

Algunas características importantes de estas bases son la fácil diferenciabilidad y la periodicidad. Dado que esta formada por senos y cosenos, siendo factible conseguir la expresión derivada en cualquier orden, además debido a sus características periódicas, se ajusta mejor a los datos que presenten esta particularidad (Ramsay & Silverman, 2005).

La expansión de Fourier es útil para funciones estables, sin grandes variaciones y con una curvatura mas o menos constante. Igualmente es conveniente que los datos muestren una periodicidad. Otra característica de este tipo de bases es que genera expansiones suaves en el sentido de diferenciabilidad, por lo que no será apropiada usarla cuando las funciones presenten discontinuidades (Ramsay & Silverman, 2005).

- **Bases B-Splines:** estas bases combinan la eficiencia computacional de los polinomios con una mayor flexibilidad, que permite elegir un número menor de funciones (K) con respecto a las bases anteriormente mencionadas, para obtener buenos resultado (Ramsay & Silverman, 2005). En primer lugar para definir un Spline es dividir el intervalo $T = [a, b]$ en L subintervalos separados por los puntos $a = t_0, t_1, \dots, t_L = b$, en cada intervalo, el Spline es un polinomio de grado m que trata de ajustar la curva. El orden del polinomio m hace referencia al número de coeficientes que hacen falta para definirlo, es decir, estos polinomios deben coincidir en los nodos de la partición hasta la derivada de orden $m - 2$, por lo que a excepción de los polinomios de orden 2, las uniones serán suaves en el sentido de diferenciabilidad. Con estas restricciones

el número de grados de libertad del Spline será $m + L - 1$. Por tanto, un elemento importante será definir el orden de los polinomios, el número de particiones y el punto en que se fijen los nodos * (Eubank, 1999).

Las bases de $B - splines$ permiten aproximar todas estas funciones facilitando así su manejo. Sus características se basan en las siguientes propiedades:

- Cada elemento de la base $\phi_k(t)$ será una función Spline de orden m y partición τ .
- Cualquier combinación lineal de funciones Spline es una función Spline.
- Cualquier función Spline de orden m sobre la partición τ se puede expresar como combinación lineal de las funciones de la base.

Sea $B_k(t, \tau)$ la representación del valor del k - ésimo elemento de la base sobre la partición τ en el instante t , la función Spline $S(t)$ se representa como:

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau)$$

- **Bases Wavelets:** la transformada wavelet, se basa en la representación de funciones mediante ondas. Los wavelets son un sistema de representación extendido al manejo de señales, utilizándose la transformada discreta para la codificación de estas, pero también ha ganado terreno en el campo de la genética, sismología y la clasificación funcional (Ramsay & Silverman, 2005).

La idea básica de esta base es que cualquier función de L_2 puede representarse mediante una función ψ apropiada, llamada wavelet madre, y considerando todas sus traslaciones de la forma $\psi_{jk}(t) = 2^{j/2} \psi(2^j t - k)$, con j y k enteros. La expansión de una función f , se basa en la observación de señales a distintas escalas de resolución, en el sentido de que el coeficiente ψ_{jk} aporta información sobre f cerca de la posición $2^{-j}k$ en una escala 2^{-j} . Una vez elegida la base, es decir, fijado el $j = J$ se tiene que

$$f(t) = \sum_k c_{J,k} \phi_{J,k}(t) + \sum_{j=J}^{\infty} \sum_k d_{j,k} \psi_{j,k}(t)$$

donde $c_{J,k}$ es el coeficiente de aproximación o señal suave

*Son puntos de control de tal manera que la función y sus dos primeras derivadas son continuas en cada punto y por lo tanto en todo su intervalo (Olaya, 2012).

$$c_{J,k} = \langle f, \phi_{J,k} \rangle = \int f(t) \phi_{J,k}(t) dt$$

y $d_{j,k}$ es el coeficiente de detalle

$$d_{j,k} = \langle f, \psi_{j,k} \rangle = \int f(t) \psi_{j,k}(t) dt$$

- **Bases con restricción:** estas bases tienen en cuenta restricciones sobre la suavización de las curvas (Ramsay & Silverman, 2005). Una función suavizada positiva x se puede definir a partir de la función exponencial:

$$x(t) = e^{W(t)} \quad (6.1)$$

donde $W(t)$ es una función conformada por bases de funciones, que dependiendo de las características de los datos observados se pueden usar bases Fourier o B-spline (Ramsay & Silverman, 2005).

En ocasiones se quiere suavizar funciones estrictamente creciente por ejemplo la precipitación acumulada. En este caso Ramsay & Silverman (2005) expresa la función suave a partir la velocidad Dx como el exponencial de una función sin restricciones W :

$$Dx(t) = e^{W(t)} \quad (6.2)$$

Integrando ambos lados de esta ecuación, se obtiene:

$$x(t) = C + \int_{t_0}^t \exp[W(u)] du, \quad (6.3)$$

donde C es una constante estimada a partir de los datos.

6.3. Construcción de datos funcionales

En Análisis de Datos Funcionales (ADF) los datos en lugar de ser un conjunto de vectores, como en análisis multivariado clásico, son un conjunto de curvas. En la mayoría de las aplicaciones, las curvas observadas proceden de la observación de un proceso estocástico en tiempo continuo (Aguilera, 2009).

El dato funcional se define a partir de una base de funciones $\{\phi_k\}_{k \in \mathbb{N}}$, tal que cualquier función puede representarse mediante una combinación lineal de ella, es decir, $x(t) = \sum_{k=1}^K c_k \phi_k(t)$. La expresión de forma matricial queda expresada como,

$$\mathbf{X}(t) = \mathbf{c}^T \Phi = \Phi^T \mathbf{c} \quad (6.4)$$

donde los c_k son los coeficientes en la nueva base, Φ es una matriz que contiene los valores de $\phi_k(t)$ y K será el número de bases. En este contexto es importante en la selección del valor de K , dado que la expresión de la función seguirá siendo infinita o demasiado grande, por lo que habrá que delimitar la serie de datos. De aquí la importancia de la elección del número de bases K , este se comporta como otro parámetro del modelo midiendo el grado de suavización de la función (Ramsay & Silverman, 2005).

De la expresión 6.4 se denota que el problema consiste en estimar el vector de coeficientes $\mathbf{c}$, este vector pondera cada función de la base sobre la función de interés $x(t)$, Ramsay & Silverman (2005) estimaron el vector $\mathbf{c}$ a partir de mínimos cuadrados ordinarios. A continuación se presenta el proceso de estimación propuesto:

Sea y_j el valor observado de la variable Y en el tiempo $t_j \forall j = 1, \dots, n$, tal que puede ser expresados como:

$$y_j = x(t_j) + \epsilon_j = \phi^T(t_j)c + \epsilon_j, \quad (6.5)$$

donde $x(t_j)$ es el valor de la función ajustada con (6.4) en el tiempo t_j y ϵ_j es el error de estimación en el momento t_j . Suponiendo que los errores se distribuyen normal, tienen media cero, varianza constante e independientes. La expresión por mínimos cuadrados ordinarios a minimizar es (Ramsay & Silverman, 2005):

$$SSE(c) = \sum_j^n \left[y_j - \sum_k^K c_k \phi_k(t_j) \right]^2 = \sum_j^n [y_j - \phi^T(t_j)c]^2, \quad (6.6)$$

Sea $\mathbf{Y}$ un vector conformado por todos los valores ajustados y_j , Φ^T una matriz de dimensión $n \times K$ compuesta por los valores de las K funciones base $\phi_k(t_j)$ y ϵ el vector de residuales de tamaño $n \times 1$, por lo tanto la expresión (6.5) es expresada de forma matricial como (Ramsay & Silverman, 2005):

$$\mathbf{Y} = \Phi^T \mathbf{c} + \epsilon \quad (6.7)$$

de esta manera, la expresión (6.6) se puede expresar como:

$$SSE(c) = (\mathbf{Y} - \Phi^T \mathbf{c})^T (\mathbf{Y} - \Phi^T \mathbf{c}), \quad (6.8)$$

resolviendo el sistema para estimar $\mathbf{c}$, se obtiene el estimador $\hat{\mathbf{c}}$ que minimiza la suma de cuadrados de (6.8), y esta dado por:

$$\hat{\mathbf{c}} = (\Phi \Phi^T)^{-1} \Phi \mathbf{Y} \quad (6.9)$$

Finalmente los valores ajustados son expresados como:

$$\hat{\mathbf{Y}} = \Phi^T (\Phi \Phi^T)^{-1} \Phi \mathbf{Y} \quad (6.10)$$

6.4. Estimación de la la varianza

Se puede estimar la varianza de un dato funcional (Ramsay & Silverman, 2005).

De la ecuación (9.8), una expresión de la varianza es,

$$\text{Var} [\hat{\mathbf{Y}}] = (\Phi^T \Phi)^{-1} \text{Var} [\hat{\mathbf{c}}] \Phi^T \quad (6.11)$$

donde,

$$\text{Var} [\hat{\mathbf{c}}] = (\Phi^T \Phi)^{-1} \Phi^T \Sigma_e \Phi (\Phi^T \Phi)^{-1} \quad (6.12)$$

asumiendo $\Sigma_e = \sigma^2 \mathbf{I}$, la matriz de varianzas y covarianzas del vector de errores ϵ , de esta manera la expresión resultante es:

$$\text{Var} [\hat{\mathbf{c}}] = \sigma^2 (\Phi^T \Phi)^{-1} \quad (6.13)$$

y finalmente,

$$\text{Var} [\hat{\mathbf{Y}}] = \sigma^2 \Phi (\Phi^T \Phi)^{-1} \Phi^T \quad (6.14)$$

6.5. Número óptimo de funciones de la base (K)

Ramsay & Silverman (2005) plantea que para seleccionar el número K se debe tener en cuenta tanto el sesgo,

$$\text{Sesgo}[\hat{x}(t)] = x(t) - E[\hat{x}(t)] \quad (6.15)$$

y la varianza de la estimación,

$$\text{Var}[\hat{x}(t)] = E[\{\hat{x}(t) - E[\hat{x}(t)]\}^2] \quad (6.16)$$

Por lo tanto, un indicador que permite controlar tanto el sesgo como la varianza es el Error Cuadrático Medio (ECM),

$$ECM[\hat{x}(t)] = \text{Sesgo}^2[\hat{x}(t)] + \text{Var}[\hat{x}(t)] \quad (6.17)$$

Esta relación indica que vale la pena tolerar un poco de sesgo, siempre y cuando haya un gran reducción de la varianza (Ramsay & Silverman, 2005). Adicional a este proceso, Ramsay & Silverman (2005) proponen la utilización de la validación cruzada generalizada (VCG).

- **Validación cruzada (VC):** es una metodología usada para las elecciones de parámetros que controlen el equilibrio entre la bondad del ajuste y la capacidad predictiva (Delicado, 2008). El método consiste en extraer de la muestra secuencialmente cada una de las observaciones x_i , estimar el modelo con los datos restantes y predecir el dato ausente. Y finalmente, comparar la predicción con el dato real. Si se realiza esto para cada valor de K , entonces se pueden construir la función

$$ECM_{VC}(K) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{m}_{(i)}(x_i)]^2 \quad (6.18)$$

que mide el error de la predicción del estimador fuera de la muestra para cada valor de K . El valor que minimice esa función, K_{VC} , será el valor del parámetro elegido.

En forma matricial se tiene que

$$\hat{Y} = SY$$

donde los vectores columnas Y e $\hat{Y}$ tienen elementos y_i e $\hat{y}$, respectivamente y la matriz S va a ser una matriz de estimación. Esta matriz es análoga a la matriz HAT $H = X(X^T X)^{-1} X^T$ en regresión lineal múltiple:

$$\hat{Y}_L = X(X^T X)^{-1} X^T Y = HY \quad (6.19)$$

y se puede demostrar que

$$ECM_{VC}(K) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - s_{ii}} \right)^2, \quad (6.20)$$

al igual que en regresión múltiple, no es necesario ajustar n regresiones no paramétricas, solo hay que ajustar la que hace intervenir todos los datos y anotar la matriz S (Delicado, 2008).

- **Validación cruzada generalizada (VCG):** Estos modelos puede hacerse una modificación a la validación cruzada. A esta metodología se le conoce como Validación Cruzada Generalizada (VCG). Esta modificación se basa en reemplazar los valores s_{ii} de la diagonal principal de la matriz S por su valor promedio. (Delicado, 2008).

$$ECM_{VCG}(K) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - \frac{\nu}{n}} \right)^2, \quad (6.21)$$

donde $\nu = \text{Traza}(S)$

Reexpresando la ecuación (6.21) se tiene que

$$ECM_{VCG}(K) = \frac{n\hat{\sigma}_\varepsilon^2}{n - \nu} \quad (6.22)$$

donde

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n - \nu} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6.23)$$

es una estimación de la varianza de los residuales del modelo. Y el valor de K_{VCG} que minimiza esa función, es el parámetro elegido por el criterio de validación cruzada generalizada (Delicado, 2008).

6.6. Estadística descriptiva funcional

Las estadísticas descriptivas clásicas para datos univariados se aplica igualmente a los datos funcionales. Suponiendo que un dato funcional para i réplicas, llega como un conjunto de valores discretos $x_{i1}, \dots, x_{in}$, la primera tarea es convertir estos valores, utilizando métodos

no paramétricos, a una función $X(t)$ calculable para cualquier valor de t (Ramsay & Silverman, 2005). Teniendo en cuenta que un primer análisis estadístico de un conjunto de datos debe empezar por un análisis exploratorio, se presentan a continuación algunas características descriptivas funcionales, propuestas en Ramsay & Silverman (2005):

Sea $x_i(t)$ $i = 1, 2, \dots, N$ una muestra de curvas o datos funcionales, la media, varianza, covarianza y correlación funcional están dadas por:

- Promedio funcional:

$$\bar{x}(t) = N^{-1} \sum_{i=1}^N x_i(t) \quad (6.24)$$

- Varianza funcional:

$$Var_x(x(t)) = (N-1)^{-1} \sum_{i=1}^N (x_i(t) - \bar{x}(t))^2 \quad (6.25)$$

- Covarianza funcional:

$$Cov_x(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^N (x_i(t_1) - \bar{x}(t_1)) (x_i(t_2) - \bar{x}(t_2)) \quad (6.26)$$

- Correlación funcional:

$$Corr_x(t_1, t_2) = \frac{Cov_x(t_1, t_2)}{\sqrt{Var_x(t_1) Var_x(t_2)}} \quad (6.27)$$

- Covarianza cruzada funcional:

$$Cov_{x,y}(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^n (x_i(t_1) - \bar{x}(t_1)) (y_i(t_2) - \bar{y}(t_2)) \quad (6.28)$$

- Correlación cruzada funcional:

$$Corr_{x,y}(t_1, t_2) = \frac{Cov_{x,y}(t_1, t_2)}{\sqrt{Var_x(t_1) Var_y(t_2)}} \quad (6.29)$$

6.7. Resultados de la construcción y análisis exploratorio de datos funcionales

En éste capítulo se muestra la aplicación de la metodología para la construcción de curvas de precipitación acumulada de estaciones climáticas y procedentes de satélite en el Valle del

Cauca, período 1993-2011.

Se construyeron los datos funcionales a partir de una base Bspline monótona creciente, sin parámetro de penalización, es decir, que la estimación de las curvas es realizada a partir de mínimos cuadrados ordinarios (Ramsay & Silverman, 2005). El número óptimo de funciones de base K se estimó por año a partir de validación cruzada generalizada. En la Tabla 6.1 se observa el número de funciones de bases necesarias para construir los datos funcionales de estaciones climáticas y de satélite en todo el período de análisis. De acuerdo a los resultados, se observó una recurrencia en el número óptimo de bases de funciones para ambos tipos de datos, siendo $K = 10$ para transformar los datos de estaciones climáticas y $K = 20$ para la construcción de los datos funcionales de satélite.

Tabla 6.1: Número de funciones para datos de estaciones y de satélite, en el período 1993-2011.

Año	K		Año	K	
	Estación	Satélite		Estación	Satélite
1993	10	20	2003	10	20
1994	6	20	2004	10	20
1995	10	20	2005	10	20
1996	10	20	2006	10	20
1997	10	20	2007	10	20
1998	6	20	2008	10	20
1999	10	20	2009	10	20
2000	10	20	2010	10	20
2001	10	20	2011	10	20
2002	18	20			

Para ilustrar la suavización de los datos, a continuación se muestra el proceso de construcción de los datos funcionales de precipitación procedentes de estaciones climáticas para el año 2011. La Tabla 6.2 muestra los coeficientes del ajuste usando bases con restricción y en la Figura 6.1 se observa el comportamiento de las curvas suavizadas para las primeras cuatro estaciones, se destaca un excelente ajuste con respecto a los datos observados, ya que para todas las estaciones, las curvas pasan justo por los puntos que se desean estimar.

Tabla 6.2: Coeficientes del ajuste de los datos funcionales, para 82 estaciones climáticas en el año 2011.

Estación	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
1	20.18	264.04	-51.61	1042.87	1352.91	1518.52	1811.2	1977.69	2625.11	2734.39
2	19.88	73.59	127.89	601.82	890.94	1096.41	1228.43	1676.27	1880.73	1978.98
3	74.9	100.47	233.09	547.1	922.03	1098.51	1309.58	1586.34	1916.34	1922.88
4	5.92	4.17	57.96	231.33	449.21	507.35	610.77	714.19	943.16	1035.38
5	82.42	-122.61	561.94	1094.64	1435.81	1667.57	2024.58	2287.01	2522.5	2606.12
6	10.93	-37.14	271.91	670.65	880.1	804.58	859.73	988.99	1225.81	1247.67
7	18.44	0.45	146.78	324.93	621.23	818.47	740.12	895.69	1221.95	1289.34
8	26.25	-24.6	268.26	677.91	913.05	804.46	949.46	1214.77	1477.24	1500.3
9	26.4	-87.38	276.26	548.99	742.6	763.04	735.72	1153.26	1526.66	1463.52
10	97.48	-80.09	507.68	907.53	1152.12	1203.46	1248.69	1840.66	2277.03	2404.64
11	55.27	-21.16	261.1	694.06	915.81	1022.33	1166.01	1598.9	2104.75	2166.8
12	7.51	17.96	200.06	623.3	841.99	882.8	961.74	1321.26	1686.92	1655.63
13	28.09	-50.46	223.8	639.67	879.82	967.38	1016.59	1566.38	1659.61	1734.92
14	49.34	-51.4	332.49	624.56	898.54	905.55	1052.06	1528.72	2184.44	2282.76
15	103.02	32.04	404.69	620.61	1176.21	1152.61	1517.04	1836.27	2403.91	2575.1
16	40.8	149.69	127.14	613.33	888.69	1003.67	1328.36	1476.64	2176.69	2216.05
17	68.82	19.88	304.22	558.99	974.63	1034.94	1292.17	1775.72	2112.62	2221.46
18	21.09	5.44	196.39	543.65	1051.9	1085.82	1348.49	1442.59	2239.74	2385.18
19	32.18	14.65	147.29	508.31	871.81	950.47	1070.35	1380.16	1741.5	1820.18
20	-1.4	10.48	122.65	322.42	533.2	569.69	716.9	1016.74	1344.86	1348.01
21	17.55	25.93	69.28	589.94	956.08	1035.28	1139.23	1549	1806.48	1828.73
22	27.09	149.09	4.64	634.75	1037.32	1374.6	1618.97	1742.19	2483.24	2455.51
23	23.01	129.03	58.82	628.81	1028.52	993.6	1283.78	1477.6	1769.04	1828.29
24	53.59	-52.99	481.86	831.86	1593.49	1628.88	1779.22	2358.02	2914.47	3079.14
25	9.95	68.14	100.14	517.37	950.68	959.08	1044.85	1382.15	1677.21	1764.6
26	37.09	41.93	121.39	464.56	790.33	890.74	1118.85	1292.9	1518.64	1544.14
27	41.27	28.09	222.48	405.78	554.89	840.43	815.65	1163.88	1401.63	1416.11
28	64.78	26.64	294.7	560.58	1125.16	1235.61	1406.97	1764.4	2062.94	2402.06
29	48.35	14.15	222.14	596.41	866.81	1056.87	1096.88	1509.98	1954.23	2152.57
30	-0.18	-47.58	214.71	390.71	767.85	753.54	850.88	1170.32	1417.64	1410.92
31	65.86	109.97	465.03	826.47	1230.18	1426.98	1734.71	2172.69	2597.72	2889.18
32	29.14	195.79	304.79	963.46	1412.36	1749.21	2315.63	2675.59	3059.1	3166.11
33	38.51	-23.18	219.72	466.84	946.55	954.36	1098.87	1478.52	1795.63	2002.77
34	45.53	-7.52	458.36	881.61	1686.04	1523.55	1794.21	2123.73	2732.23	3026.48
35	126.36	39.01	448.76	1101.26	1829.8	1878.87	2010.91	3057.56	3660.95	3924.76
36	92.15	39.7	360.14	721.34	1350.07	1260.67	1516.36	1727.72	2293.78	2248.27
37	163.98	31.91	483.21	896.87	1662.83	1659.44	1963.39	2368.03	3061.34	3332.95
38	89.57	88.02	278.35	425.5	740.07	833.71	917.04	1259.57	1536.63	1711.46
39	-3.02	236.3	72.43	637.32	1005.3	1096.59	1177.5	1809.39	2250.74	2315.19
40	53.79	16.86	327.3	575.58	1062.42	1137.73	1359.81	1600.97	2072.8	2206.61
41	85.18	12.13	117.49	927.16	1426.17	1493.98	1713.55	2036.15	2526.87	2579.53
42	41.18	20.21	155.03	421.36	682.69	935.42	973.02	1397.49	1662.56	1698.46

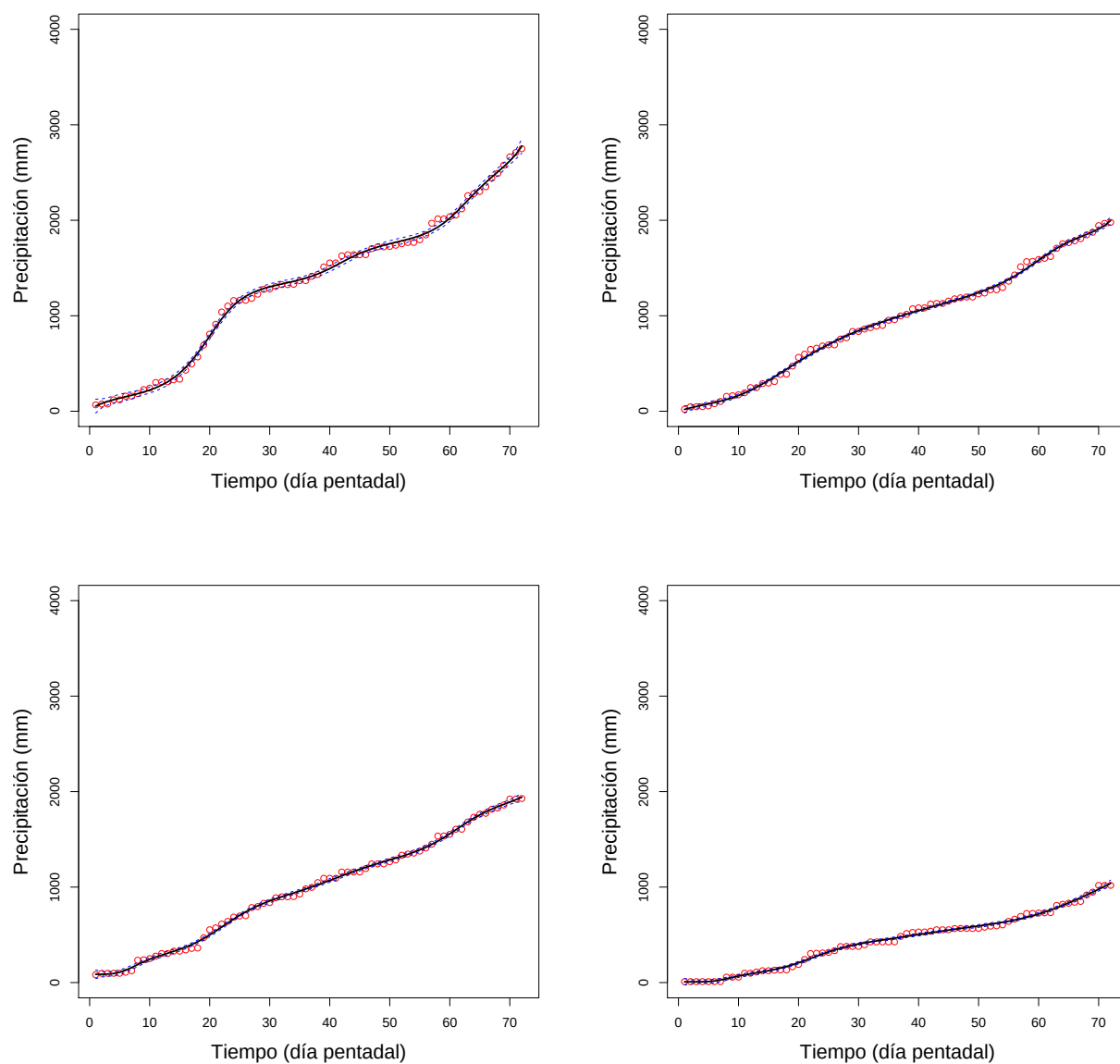


Figura 6.1: Comportamiento de los datos observados (puntos), datos suavizados (línea) e intervalo del 95 % de confianza (línea azul) de las primeras cuatro estaciones, para el año 2011.

En la Figura **6.2** se muestran los datos funcionales de estaciones climáticas y en la Figura **6.3** los datos funcionales de satélite por año. Al igual que los datos observados, los datos funcionales de precipitación proveniente de estaciones climáticas presentaron mayor variabilidad con respecto a la precipitación funcional proveniente de satélite, ya que esta información es obtenida a partir de un proceso de estimación en toda el área de cobertura, sin tener en cuenta el relieve de la superficie.

La curva promedio y desviación estándar de ambas fuentes de datos, fué mayor en los datos de precipitación provenientes de estaciones climáticas (Figura **6.2**) en comparación a la información de satélite (Figura **6.3**); en esta última, hubo un comportamiento más homogéneo en todo el período de estudio. Una razón, por la cual la variabilidad se incrementa en las curvas de estaciones y satélite, es el proceso acumulativo utilizado para la construcción de los datos observados.

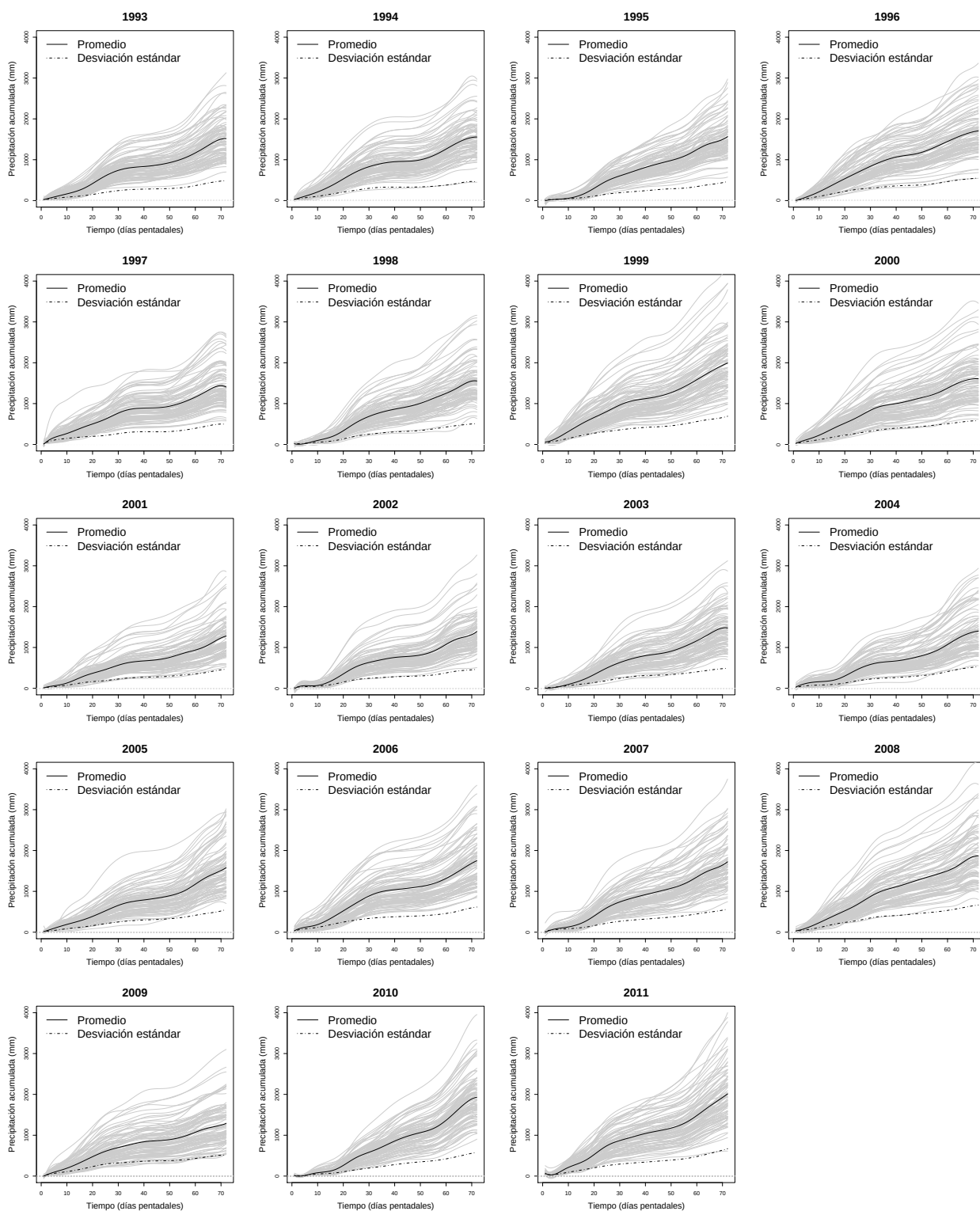


Figura 6.2: Datos funcionales de precipitación de estaciones climáticas, promedio y desviación estándar en el período 1993-2011.

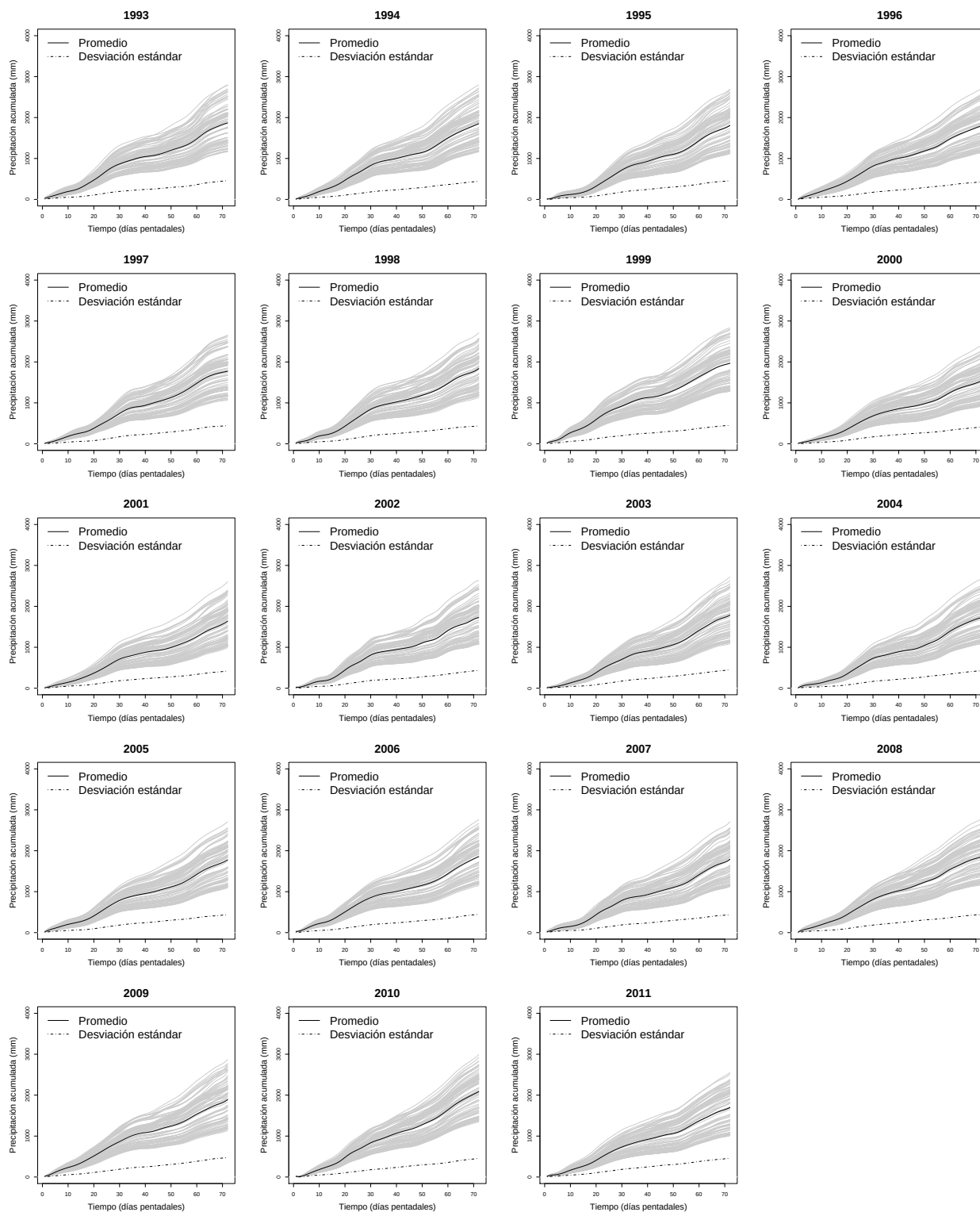


Figura 6.3: Datos funcionales de precipitación de satélite, promedio y desviación estándar en el período 1993-2011.

6.8. Conclusiones

Se pudo establecer el número óptimo de funciones para la construcción de los datos funcionales de estaciones climáticas ($K = 10$) y de satélite ($K = 20$) a partir de validación cruzada generalizada (Tabla **6.1**).

El uso de una base monótona creciente, fue apropiado para construir los datos de precipitación para estaciones climáticas y de satélite, ya que permitió restringir las curvas estimadas de acuerdo al origen de los datos observados. Se observó que la función estimada sigue el comportamiento del dato funcional (Figura **6.1**).

Los datos funcionales tanto de estaciones climáticas como de satélite reflejaron el comportamiento de los datos observados. La curva media y desviación estándar mostraron un comportamiento creciente debido al proceso acumulativo de los datos observados (Figura **6.2** y Figura **6.3**).

Capítulo 7

Modelo de regresión lineal para datos funcionales

En este capítulo se presenta el ajuste del modelo de regresión funcional asumiendo independencia en los errores, es decir, sin asumir la dependencia espacial debido a la localización entre las estaciones climáticas, con el objetivo de extraer los residuales para identificar la estructura de correlación espacial para cada año de estudio. De igual manera, se muestran los resultados del ajuste del modelo con datos de precipitación, donde la variable de respuesta funcional es constituida por las curvas de precipitación provenientes de estaciones climáticas y la covariable funcional por las curvas de precipitación procedentes de imágenes de satélite.

7.1. Modelo de regresión funcional concurrente

Para el modelo con respuesta funcional y variables regresoras funcionales no se dispone en general de tantas herramientas fuera de L^2 . Este modelo esta representando en bases restringidas tanto la respuesta como la variable regresora (Ramsay & Silverman, 2005).

El modelo se expresa como:

$$Y_i(t) = \beta_0(t) + \beta_1(t)X_{i1}(t) + \beta_2(t)X_{i2}(t) + \cdots + \beta_q(t)X_{iq}(t) + \epsilon_i(t) \quad (7.1)$$

En notación matricial se denota de la siguiente manera:

$$\mathbf{Y}(t) = \mathbf{X}(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$$

Donde β_j son determinados por:

$$\beta_j(t) = \sum_k^{K_j} b_{kj} \theta_{kj}(t) = \boldsymbol{\theta}_j^T(t) \mathbf{b}_j(t)$$

en términos de K_j funciones de base $\boldsymbol{\theta}_j^T(t)$.

Definiendo $K_\beta = \sum_j^q K_k$, se construye el vector $\mathbf{b}$ de tamaño K_β

$$\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_q^T)$$

denotando la matriz en función de $\boldsymbol{\theta}$ de dimensión q por K_β

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\theta}_1^T & 0 & \cdots & 0 \\ & \boldsymbol{\theta}_2^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\theta}_q^T \end{pmatrix}$$

De esta manera el modelo se expresa

$$\mathbf{Y}(t) = \mathbf{X}(t)\boldsymbol{\Theta}(t)\mathbf{b} + \boldsymbol{\epsilon}(t) \quad (7.2)$$

Si se define

$$\mathbf{X}^*(t) = \mathbf{X}(t)\boldsymbol{\Theta}(t),$$

entonces

$$\mathbf{Y}(t) = \mathbf{X}^*(t)\mathbf{b} + \boldsymbol{\epsilon}(t) \quad (7.3)$$

Donde el modelo de regresión funcional tiene K_β parámetros.

La estimación de los parámetros del modelo se realiza por mínimos cuadrados ordinarios (Ramsay & Silverman, 2005),

$$MCO[\mathbf{b}] = \int [\mathbf{Y}(t) - \mathbf{X}^*(t)\mathbf{b}]^T [\mathbf{Y}(t) - \mathbf{X}^*(t)\mathbf{b}] dt$$

Diferenciando e igualando a cero

$$\begin{aligned} \left(\int_T \mathbf{X}^{*T}(t)\mathbf{X}^*(t)dt \right) \hat{\mathbf{b}} &= \int_T \mathbf{X}^{*T}(t)\mathbf{Y}(t)dt \\ \hat{\mathbf{b}} &= \left(\int_T \mathbf{X}^{*T}(t)\mathbf{X}^*(t)dt \right)^{-1} \left(\int_T \mathbf{X}^{*T}(t)\mathbf{Y}(t)dt \right) \end{aligned} \quad (7.4)$$

7.2. Intervalos de confianza en el modelo de regresión funcional

Basados en las expresiones expuestas en Ramsay & Silverman (2005), Ramsay et al. (2011), primero se definen los residuales del modelo expresado en (7.2):

$$\mathbf{r}(t) = \mathbf{Y}(t) - \mathbf{X}(t)\boldsymbol{\Theta}(t)\hat{\mathbf{b}} \quad (7.5)$$

y la matriz de varianzas y covarianzas de los residuales funcionales queda expresada como:

$$\boldsymbol{\Sigma}_e^* = \frac{1}{N} \mathbf{r}(t)\mathbf{r}^T(t) \quad (7.6)$$

donde, $\mathbf{r}$ es la matriz de residuales. Sea $\mathbf{C}$ la matriz de coeficientes tal que $\mathbf{Y}(t) = \mathbf{C}\boldsymbol{\phi}(t)$ y sustituyendo en la expresión 7.4, entonces

$$\begin{aligned} \left(\int_T \mathbf{X}^{*T}(t)\mathbf{X}^*(t)dt \right) \hat{\mathbf{b}} &= \int_T \mathbf{X}^{*T}(t)\mathbf{C}\boldsymbol{\phi}(t)dt \\ \hat{\mathbf{b}} &= \left(\int_T \mathbf{X}^{*T}(t)\mathbf{X}^*(t)dt \right)^{-1} \left(\int_T \boldsymbol{\phi}(t)^T \otimes \mathbf{X}^{*T}(t)dt \right) \text{vec}(\mathbf{C}) \end{aligned} \quad (7.7)$$

donde, $\otimes$ representa el producto Kroneker y $\text{vec}(\mathbf{C})$ es la vectorización de la matriz $\mathbf{C}$ (Ramsay et al., 2011). Ahora, si se denota a:

$$\text{c2bMap} = \left(\int_T \mathbf{X}^{*T}(t) \mathbf{X}^*(t) dt \right)^{-1} \left(\int_T \boldsymbol{\phi}(t)^T \otimes \mathbf{X}^{*T}(t) dt \right) \text{vec}(\mathbf{C}) \quad (7.8)$$

y

$$\text{y2cMap} = (\boldsymbol{\phi}^T \boldsymbol{\phi})^{-1} \boldsymbol{\phi}^T \quad (7.9)$$

Finalmente, se puede expresar la varianza de $\hat{\mathbf{b}}$ como:

$$\text{Var}[\hat{\mathbf{b}}] = \text{c2bMap} \text{y2cMap} \boldsymbol{\Sigma}_e^* \text{y2cMap}^T \text{c2bMap}^T \quad (7.10)$$

7.3. Bondad de ajuste del modelo

La metodología para evaluar el ajuste del modelo establecido se realizará a partir de validación cruzada. Esta consiste en predecir la curva de precipitación excluyendo una ubicación donde se conozca el dato funcional y de esta manera predecir la curva; posteriormente compararla con el dato funcional conocido correspondiente a la ubicación excluida y determinar el poder predictivo. Otros métodos de evaluación son el coeficiente de determinación funcional $R^2(t)$, propuesto en Ramsay & Silverman (2005).

$$R^2(t) = 1 - \int \{\hat{y}_i(t) - y_i(t)\}^2 dt / \int \{y_i(t) - \bar{y}(t)\}^2 dt \quad (7.11)$$

7.4. Modelo de regresión funcional entre precipitación y altura

Para evaluar el aporte de la altura en el ajuste de la precipitación pentadal de estaciones climáticas. Se ajustó un modelo con la información de satélite, la altura y la altura al cuadrado como variables explicativas, ya que en el trabajo de Cruz & Males (2012) mostraron que este parámetro y su componente cuadrática presentaron una contribución importante en la modelación de la precipitación en el Valle del Cauca. A continuación, se expresa el modelo previamente descrito:

$$\begin{aligned} PrecipEstPent_i(t) = & \beta_0(t) + \beta_1(t) PrecipSatPent_i(t) + \beta_2(t) Altura_i(t) \\ & + \beta_3(t) Altura_i^2(t) + \epsilon_i(t) \end{aligned} \quad (7.12)$$

donde, $PrecipEstPent_i(t)$ es la variable respuesta funcional que corresponde a los datos funcionales de precipitación procedentes de estaciones climáticas, $PrecipSatPent_i(t)$ es la variable explicativa funcional que corresponde a los datos funcionales de precipitación procedentes de satélite, $Altura_i(t)$ y $Altura_i^2(t)$ corresponden a la altura y altura al cuadrado de las estaciones utilizadas en el análisis (estas fueron tratadas como variables funcionales) y $\epsilon_i(t)$ es la componente aleatoria del modelo correspondiente al error funcional, $i = 1, 2, \dots, 82$ correspondiente a 82 estaciones climáticas por año y t representa el tiempo en días pentadales de un año que está en un rango de 1 a 72.

La variable altura fue extraída para cada estación climática incluida en el estudio. Para cada año, esta variable fue constante, es decir, que se tenían 82 series constantes de altura y su componente cuadrática. Posteriormente, fueron transformadas a datos funcionales, generando 82 objetos funcionales $Altura_i(t)$ y $Altura_i^2(t)$ donde $i = 1, 2, \dots, 82$. Finalmente, ambas variables funcionales fueron incluidas en el modelo de regresión funcional.

En la Figura 7.1, se muestra la estimación del coeficiente funcional asociado a la altura con sus respectivas bandas de confianza (95%), se denota un mínimo aporte en la modelación de la precipitación de las estaciones climáticas, ya que las bandas de confianza siempre contuvieron al cero en todo el período de tiempo, excepto en los años 2009, 2010 y 2011. En la Figura 7.2, aparece el comportamiento del coeficiente funcional estimado, asociado a la altura al cuadrado. Así mismo, se observan fluctuaciones extremadamente pequeñas alrededor de cero, es decir, que no tuvo una contribución para explicar las variaciones de los datos de precipitación de las estaciones climáticas.

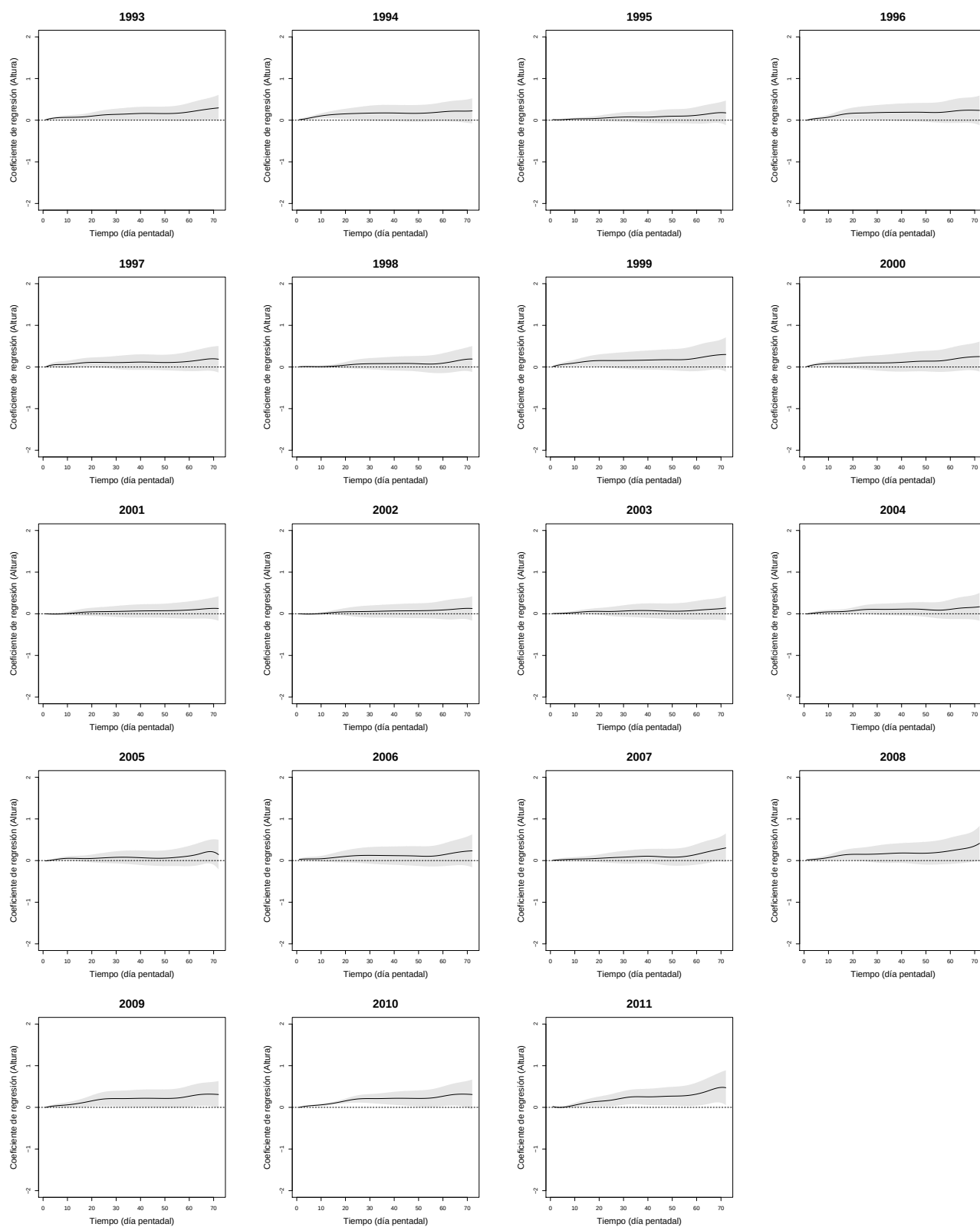


Figura 7.1: Coeficiente de regresión asociado a la altura, en el período 1993-2011.

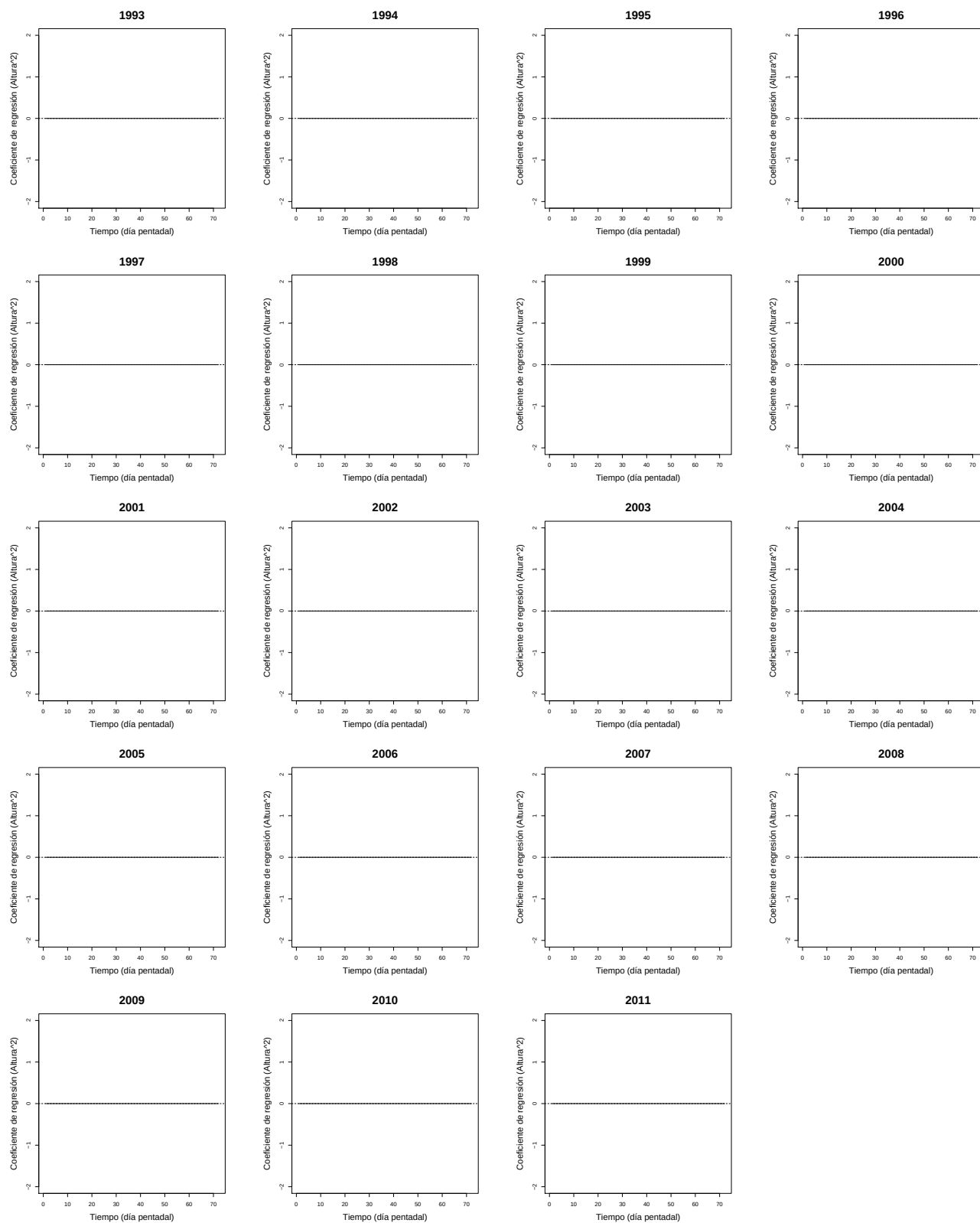


Figura 7.2: Coeficiente de regresión asociado a la altura al cuadrado, en el período 1993-2011.

7.5. Conclusiones

El ajuste del modelo de regresión funcional incluyendo la altura al cuadrado no mostró un aporte en la explicación de los datos funcionales de precipitación de estaciones climáticas (Figure 7.1 y Figure 7.2). Una razón posible, es que no se cuenta con un modelo de elevación digital que permita tener un valor de altura en cualquier punto de la región de estudio, ya que esto posibilitaría incluir en el modelo una curva de altitud de cualquier localización en el Valle del Cauca y no sólo limitarlo a las localizaciones de las estaciones en tierra.

Capítulo 8

Estimación de la estructura de la correlación espacial de los errores

En este capítulo se propone estimar la estructura de correlación espacial de los residuales del modelo de regresión funcional concurrente ajustado para cada año de estudio. A continuación se presentan las etapas del proceso de estimación:

1. Se ajustó un modelo de regresión funcional con variable explicativa funcional (datos de estaciones climáticas) y covariable funcional (datos de satélite).
2. Se extrajeron los residuales del modelo para cada año y se convirtieron a datos funcionales. En este proceso se seleccionó el número óptimo de bases a partir de validación cruzada generalizada, se utilizaron bases Bspline con parámetro de penalización igual a cero.
3. Con las estimaciones de las traza variogramas simples por año en función de la matriz de distancias euclidianas, se ajustó un modelo paramétrico conocido (Esférico, Exponencial, Gaussiano o Matern) para obtener la estructura de dependencia espacial.

A continuación se presentan los aspectos teóricos de la geoestadística extendida al campo funcional (Giraldo, 2009).

8.1. Geoestadística para datos funcionales

Sea $\{\chi_s(t) : s \in R^d, t \in T = [a,] \subset R\}$ un proceso aleatorio tal que $\chi_s(t)$ es una variable aleatoria funcional. Sean $s_1, s_2, \dots, s_n$ puntos en el espacio D , en los cuales se observa una realización del proceso aleatorio $\{\chi_s(t) : \chi_{s_1}(t), \chi_{s_2}(t), \dots, \chi_{s_n}(t)\}$.

El proceso es estacionario de segundo orden e isotrópico si (Giraldo, 2009):

1. $E(\chi_s(t)) = m(t), \forall t \in T; s \in D$.

2. $V(\chi_s(t)) = \sigma^2(t), \forall t \in T; s \in D.$
3. $Cov(\chi_{s_i}(t), \chi_{s_j}(t)) = C(h; t) = C_{s_i s_j}(t), \forall s_i, s_j \in D; \text{ con } h = \|s_i - s_j\|.$
4. $\frac{1}{2}V(\chi_{s_i}(t) - \chi_{s_j}(t)) = \gamma_{s_i s_j}(t); \forall s_i, s_j \in D; t \in T; \text{ con } h = \|s_i - s_j\|.$

Dado que uno de los objetivos dentro del campo de la geoestadística es la predicción de una curva en un sitio no observado s_0 , se define la versión funcional del predictor kriging ordinario como (Giraldo, 2009):

$$\hat{\chi}_{s_0}(t) = \sum_{i=1}^n \lambda_i \chi_{s_i}(t), \lambda_1, \dots, \lambda_n \in R. \quad (8.1)$$

Los pesos λ_i óptimos para la predicción se obtienen al solucionar el sistema de ecuaciones:

$$\begin{pmatrix} \int_t \gamma_{s_1, s_1}(t) d(t) & \cdots & \int_t \gamma_{s_1, s_n}(t) d(t) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \int_t \gamma_{s_n, s_1}(t) d(t) & \cdots & \int_t \gamma_{s_n, s_n}(t) d(t) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \int_t \gamma_{s_0, s_1}(t) dt \\ \vdots \\ \int_t \gamma_{s_0, s_n}(t) dt \\ 1 \end{pmatrix}$$

donde,

$$\gamma(h) = \frac{1}{2} \int_t V(\chi_{s_i} - \chi_{s_j}(t)) dt, h = \|s_i - s_j\| \quad (8.2)$$

es denominado la traza-variograma (Giraldo, 2009).

Usando el teorema de Fubini la traza-variograma se puede expresar como:

$$\frac{1}{2} \int_t E \left([\chi_{s_i}(t) - \chi_{s_j}(t)]^2 \right) dt = \frac{1}{2} E \left[\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt \right] \quad (8.3)$$

por lo tanto, el estimador por el método de los momentos está expresado como (Giraldo, 2009):

$$\gamma(\hat{h}) = \frac{\sum_{i,j \in N(h)} \int_t [\chi_{s_i}(t) - \chi_{s_j}(t)]^2 dt}{2N(h)} \quad (8.4)$$

dónde $N(h)$ es el número de pares de sitios en función de una distancia h .

Con la traza-variograma estimada para el parámetro de interés, se puede ajustar un modelo paramétrico (Esférico, Exponencial, Gaussiano o Matern), a los sitios $(h_i, \hat{\gamma}(h_i))$ para obtener $\int_t \gamma_{s_i, s_j}(t) dt, i = 1, 2, \dots, n; j = 1, 2, \dots, n.$

8.2. Identificación de la estructura de correlación espacial

Utilizando como criterio la suma de cuadrados minimizados, se identificó el modelo óptimo de Traza-Variograma para cada año. Posteriormente se incorporó la estructura del modelo para estimar la covarianza en función de la distancia entre las estaciones.

En el contexto funcional define la covarianza espacial funcional como (Giraldo, 2009):

$$\int_t C(h; t) = \int_t \sigma^2(t) - \int_t \gamma(h; t) \quad (8.5)$$

entonces una estimación sería:

$$\int_t \hat{C}(h; t) = \int_t \hat{\sigma}^2(t) - \int_t \hat{\gamma}(h; t) \quad (8.6)$$

$$\Omega = \begin{pmatrix} \hat{\sigma}^2(t) & \hat{C}_{12} & \hat{C}_{13} & \cdots & \hat{C}_{1n} \\ \hat{C}_{21} & \hat{\sigma}^2(t) & \hat{C}_{23} & \cdots & \hat{C}_{2n} \\ \hat{C}_{31} & \hat{C}_{32} & \ddots & \cdots & \hat{C}_{3n} \\ \vdots & & & \ddots & \vdots \\ \hat{C}_{n1} & \cdots & \cdots & \cdots & \hat{\sigma}^2(t) \end{pmatrix}_{N \times N} \quad (8.7)$$

donde Ω será la matriz de varianzas y covarianzas que se integrará en la estimación de los coeficientes para la construcción de los parámetros funcionales del modelo expuesta en 8.7. En la Tabla 8.1 se exponen las estructuras de covarianza espacial utilizadas en el proyecto de investigación (Littell et al., 2006).

Tabla 8.1: Estructuras de covarianza espacial.

Modelo	Estructura
Exponencial	$\sigma^2 [\exp(-d_{ii'}^2/\rho)]$
Esférico	$\begin{cases} \sigma^2 \left[1 - \left(\frac{3d_{ii'}}{2\rho} - \frac{d_{ii'}^3}{2\rho^3} \right) \right], & d_{ii'} \leq \rho \\ \sigma^2, & d_{ii'} > \rho \end{cases}$
Gaussiano	$\sigma^2 [\exp(-d_{ii'}^2/\rho^2)]$
Matern	$\sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\sqrt{2\nu} \frac{d_{ii'}}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{d_{ii'}}{\rho} \right)$

Donde, ρ es el rango, $d_{ii'}$ es la distancia entre pares de localizaciones y $K_\kappa\left(\frac{d_{ii'}}{\rho}\right)$ es la función de Bessel modificada de la tercera clase de orden κ .

8.3. Resultados de la modelación de la correlación espacial de los residuales funcionales

Usando el criterio de la suma de cuadrados minimizada, se eligió el modelo teórico más apropiado para representar la estructura de correlación espacial de los residuales funcionales del modelo de regresión funcional propuesto por Ramsay & Silverman (2005). En la Tabla 8.2 se muestran los resultados obtenidos y se resaltan los valores mínimos que corresponden al modelo seleccionado. Se evaluaron los modelos: Exponencial, Esférico, Gaussiano y Matern. Con los resultados anteriores, se construyó la Traza-Variograma bin para cada año de estudio.

Tabla 8.2: Suma minimizada de cuadrados según modelo de covarianza espacial, en el período 1993-2011.

Año	Exponencial	Esférico	Gaussiano	Matern
1993	5.440E+12	5.441E+12	5.446E+12	5.443E+12
1994	1.219E+11	1.308E+11	1.308E+11	1.308E+11
1995	2.433E+11	2.433E+11	2.437E+11	2.436E+11
1996	1.531E+11	1.571E+11	1.534E+11	1.534E+11
1997	1.726E+09	1.726E+09	1.729E+09	1.727E+09
1998	6.181E+11	6.181E+11	6.190E+11	6.179E+11
1999	1.493E+10	1.501E+10	1.504E+10	1.499E+10
2000	1.148E+10	1.148E+10	1.150E+10	1.144E+10
2001	1.801E+10	1.805E+10	1.807E+10	1.806E+10
2002	4.076E+10	4.076E+10	4.081E+10	4.079E+10
2003	7.234E+09	7.234E+09	7.235E+09	7.235E+09
2004	3.682E+09	3.682E+09	3.685E+09	3.682E+09
2005	2.158E+09	2.158E+09	2.159E+09	2.159E+09
2006	7.372E+08	7.372E+08	7.382E+08	7.304E+08
2007	2.213E+09	2.213E+09	2.210E+09	2.211E+09
2008	7.794E+08	7.795E+08	7.795E+08	7.766E+08
2009	3.449E+08	3.449E+08	3.448E+08	3.448E+08
2010	5.751E+09	5.750E+09	5.730E+09	5.744E+09
2011	5.161E+08	5.161E+08	5.148E+08	5.148E+08

Con los residuales funcionales contruidos (sección 4.4) y con la matriz de distancias creada a través de las coordenadas de los 82 localizaciones de las estaciones climáticas, se estimó la Traza-Variograma por año. En la Figura 8.1 se observa la Traza-Variograma bin, que corresponde la estimación realizada con los promedios de las diferencias para cada distancia $d_{ii'}$ de la precipitación pentadal en cada año. En la Tabla 8.3 se muestra el ajuste del mejor modelo de correlación por año, se destaca que en 9 de los 11 años de análisis, el modelo Exponencial fue el que mejor contribución tuvo en la descripción del comportamiento de la correlación espacial de los errores funcionales.

Tabla 8.3: Resultados del modelo ajustado de correlación espacial, en el período 1993-2011.

Año	Modelo	C_0	σ^2	ϕ	κ
1993	Exponencial	23211.03	242726.89	2644.6	-
1994	Esférico	3445.54	40670.75	6960.84	-
1995	Exponencial	4868.33	54619.99	2918.93	-
1996	Exponencial	4162.76	49171.68	3730.98	-
1997	Exponencial	456.78	2514.25	3061.01	-
1998	Matern	5877.17	107876.27	5653.91	0.41
1999	Exponencial	0	1440.4	11.61	-
2000	Matern	427.86	910.98	18.76	0.54
2001	Exponencial	153.59	1327.31	10.08	-
2002	Exponencial	1238.56	25078.25	2399.98	-
2003	Esférico	806.44	9630.14	5675.3	-
2004	Exponencial	611.15	6181.04	2404.39	-
2005	Exponencial	555.49	2930.98	4527.58	-
2006	Matern	0	337.86	8.21	0.78
2007	Gaussiano	459.36	6501.15	648.57	-
2008	Matern	0	329.62	10.2	0.64
2009	Gaussiano	171.81	2353	613.24	-
2010	Gaussiano	761.15	11152.64	601.58	-
2011	Gaussiano	235.93	2567.99	557.56	-

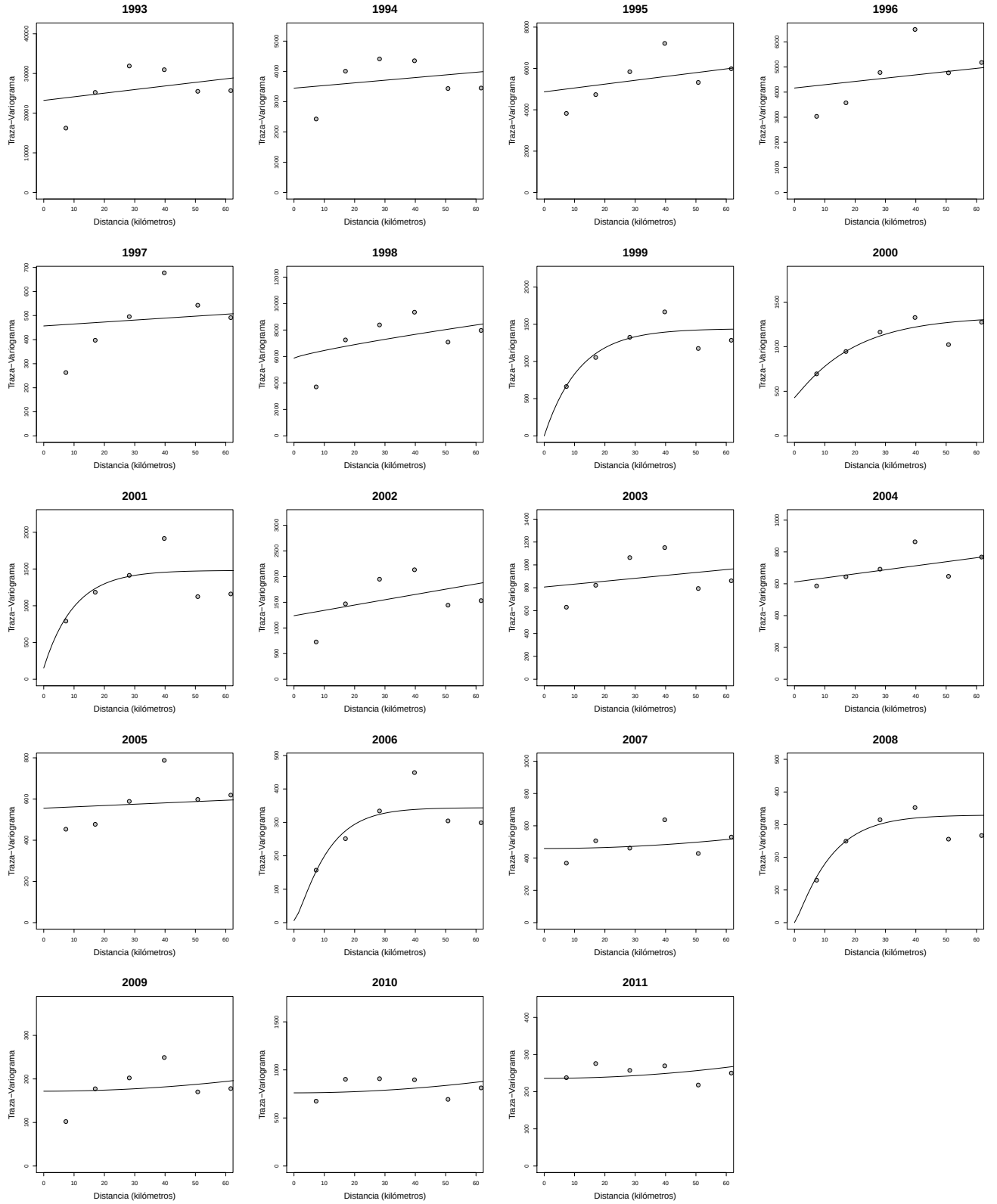


Figura 8.1: Traza-Variograma bin de los residuales funcionales, en el período 1993-2011.

8.4. Conclusiones

La caracterización de la correlación espacial de los residuales, mostró que el modelo exponencial reflejó el comportamiento de la estructura de correlación espacial en la mayor parte del período de análisis (Tabla **8.3**).

También fue evidente la presencia de autocorrelacion espacial. La Traza-Variograma bin muestra una semivarianza menor en valores cercanos de cero en el eje de la semivarianza, es decir, que las estaciones climáticas que se encuentran a menor distancia presentaron valores similares de precipitación, mientras que para distancias mayores (30 a 50 km), se encontraron valores más altos de semivarianza, indicando que las estaciones climáticas más alejadas tendrán valores de precipitación menos parecida (Figura **8.1**).

Capítulo 9

Modelo de regresión lineal funcional con estructura de correlación espacial en los errores

Con base en la teoría expuesta y en los resultados obtenidos en los capítulos anteriores, en este capítulo se expone la propuesta del presente trabajo de investigación, el cual consiste en ajustar un modelo de regresión funcional concurrente con errores correlacionados incorporando la estructura de correlación espacial identificada y modelada, a través de geoestadística funcional e integrada en la estimación de los parámetros funcionales utilizando mínimos cuadrados generalizados.

9.1. Definición del modelo y estimación de los parámetros funcionales

Para la estimación de los parámetros funcionales, se plantea emplear mínimos cuadrados generalizados usando la matriz de varianzas y covarianzas estimada en la sección 5.3.

El modelo que se va a ajustar es de la forma:

$$\mathbf{Y}(t) = \beta_0(t) + \beta_1(t)\mathbf{X}(t) + \boldsymbol{\epsilon}(t), \quad (9.1)$$

$$\mathbf{Y}(t) = \begin{pmatrix} 1 & X_1(t) \\ \vdots & \vdots \\ 1 & X_n(t) \end{pmatrix} \begin{pmatrix} \beta_0(t) \\ \beta_1(t) \end{pmatrix} + \boldsymbol{\epsilon}(t), \quad (9.2)$$

donde,

$\mathbf{Y}(t)$ es la variable de respuesta funcional del modelo.

$\mathbf{X}(t)$ es la variable predictora funcional del modelo.

$\beta_0(t)$ es el intercepto funcional.

$\beta_1(t)$ es el parámetro funcional asociado a la variable predictora funcional del modelo.

Usando base de funciones, los parámetros son expresados como

$$\sum_{j=1}^K b_{0j} \theta_j(t) = \boldsymbol{\theta}^T(t) \mathbf{b}_0, \beta_0(t) = \boldsymbol{\theta}^T(t) \mathbf{b}_0 \quad (9.3)$$

$$\sum_{j=1}^K b_{1j} \theta_j(t) = \boldsymbol{\theta}^T(t) \mathbf{b}_1, \beta_1(t) = \boldsymbol{\theta}^T(t) \mathbf{b}_1 \quad (9.4)$$

donde,

$\mathbf{b}_0$ es el vector de coeficientes correspondiente al parámetro funcional $\beta_0(t)$.

$\mathbf{b}_1$ es el vector de coeficientes correspondiente al parámetro funcional $\beta_1(t)$.

$\theta_j(t)$ son las bases de funciones de b_0 y b_1 .

Definiendo

$$\mathbf{b} = [(b_{01}, \dots, b_{0K}), (b_{11}, \dots, b_{1K})], \text{ y}$$

$$\Theta = \begin{pmatrix} \theta^T(t) & 0 \\ 0 & \theta^T(t) \end{pmatrix} = \begin{pmatrix} \theta_1(t) \dots \theta_K(t) & 0 \dots 0 \\ 0 \dots 0 & \theta_1(t) \dots \theta_K(t) \end{pmatrix}$$

el modelo se puede expresar matricialmente como:

$$\mathbf{Y}(t) = \mathbf{X}(t) \Theta(t) \mathbf{b} + \boldsymbol{\epsilon}(t) \quad (9.5)$$

$$\mathbf{Y}(t) = \begin{pmatrix} 1 & X_1(t) \\ \vdots & X_n(t) \end{pmatrix}_{(n \times 2)} \begin{pmatrix} \theta_1(t) \dots \theta_K(t) & 0 \dots 0 \\ 0 \dots 0 & \theta_1(t) \dots \theta_K(t) \end{pmatrix}_{(2 \times 2K)} \begin{pmatrix} b_{01} \\ \vdots \\ b_{0K} \\ b_{11} \\ \vdots \\ b_{1K} \end{pmatrix}_{(2K \times 1)} + \boldsymbol{\epsilon}(t)_{n \times 1}$$

$$\mathbf{Y}(t) = \begin{pmatrix} \theta^T(t) [c_{11} \dots c_{1K}] & \begin{pmatrix} \theta_1(t) \\ \vdots \\ \theta_K(t) \end{pmatrix} [\theta_1(t) \dots \theta_K(t)] \\ \vdots \\ \theta^T(t) [c_{n1} \dots c_{nK}] & \begin{pmatrix} \theta_1(t) \\ \vdots \\ \theta_K(t) \end{pmatrix} [\theta_1(t) \dots \theta_K(t)] \end{pmatrix} \begin{pmatrix} b_{01} \\ \vdots \\ b_{0K} \\ b_{11} \\ \vdots \\ b_{1K} \end{pmatrix} + \boldsymbol{\epsilon}(t)$$

reexpresando la ecuación 9.5 se tiene que

$$\mathbf{Y}(t) = \mathbf{X}^*(t)\mathbf{b} + \boldsymbol{\epsilon}(t) \quad (9.6)$$

Por mínimos cuadrados generalizados se deben encontrar los $\beta(t)$ que minimizan:

$$\begin{aligned} MCG(\mathbf{b}) &= \int [\mathbf{Y}(t) - \mathbf{X}^*(t)\mathbf{b}]^T \Omega^{-1} [\mathbf{Y}(t) - \mathbf{X}^*(t)\mathbf{b}] \\ &= \int \left[\mathbf{Y}^T(t)\Omega^{-1}\mathbf{Y}(t) - 2\mathbf{Y}^T(t)\Omega^{-1}\mathbf{X}^*(t)\mathbf{b} + (\mathbf{X}^*(t)\mathbf{b})^T \Omega^{-1}\mathbf{X}^*(t)\mathbf{b} \right] dt \end{aligned}$$

dónde la matriz Ω es la matriz de varianza y covarianza de los residuales $\epsilon(t)$ del modelo inicial estimada a partir de geoestadística funcional, cuya estructura se encuentra descrita en la Tabla 8.1 y su estimación en la Tabla 8.3 por año.

Diferenciando con respecto a $\mathbf{b}$ e igualando a cero, se obtiene

$$\int [\mathbf{X}^{*T}(t)\Omega^{-1}\mathbf{X}^*(t)dt] \mathbf{b} = \int [\mathbf{X}^{*T}(t)\Omega^{-1}\mathbf{Y}(t)] dt \quad (9.7)$$

La estimación de los coeficientes para la construcción de los coeficientes funcionales del modelo se expresa como:

$$\hat{\mathbf{b}} = \left[\int \mathbf{X}^{*T}(t)\Omega^{-1}\mathbf{X}^*(t)dt \right]^{-1} \left[\int \mathbf{X}^{*T}(t)\Omega^{-1}\mathbf{Y}(t)dt \right] \quad (9.8)$$

Finalmente la estimación de los parámetros funcionales del modelo, a partir de mínimos cuadrados generalizados queda expresado como:

$$\hat{\beta}(t) = \Theta \hat{\mathbf{b}} \quad (9.9)$$

9.2. Ajuste del modelo integrando la estructura de correlación espacial

En ésta sección se presentan los resultados de la aplicación de un modelo de regresión lineal funcional para modelar la precipitación de estaciones climáticas en función de la precipitación de satélite, integrando la estructura de correlación espacial de los errores funcionales debida a las localizaciones de las estaciones climáticas, en el período 1993-2011.

El modelo con los datos de estudio queda expresado como:

$$PrecipEstPent_i(t) = \beta_0(t) + \beta_1(t) PrecipSatPent_i(t) + \epsilon_i^\Omega(t) \quad (9.10)$$

Donde, $PrecipEstPent_i(t)$ es la variable respuesta funcional que corresponde a los datos funcionales de precipitación procedentes de estaciones climáticas, $PrecipSatPent_i(t)$ es la variable explicativa funcional que corresponde a los datos funcionales de precipitación procedentes de satélite, $\epsilon_i^\Omega(t)$ son los errores funcionales del modelo con la estructura de correlación espacial estimada en la sección 5.3 que minimizaron la suma de cuadrados (Tabla 5.2), $i = 1, 2, \dots, 82$ hace referencia a las 82 estaciones climáticas por año y t representa el tiempo en días pentadales de un año, en el período 1993-2011.

En la Figura 9.1 se muestra el comportamiento de la curva media funcional y el intercepto del modelo. Se destaca que en todos los años de estudio, el intercepto tuvo una tendencia creciente en el transcurso de cada año. Este tuvo un mayor aporte en los períodos en los años donde hubo un episodio de La Niña (Tabla 9.1 y Figura 9.1), esto fué en los años 1995, 1998, 1999, 2000, 2005, 2008, 2010 y 2011 (Weather Services, 2015).

En la Figura 9.2 se muestra el coeficiente de regresión debido a los datos de satélite para el modelo ajustado, en el período 1993-2011. Se observan aportes importantes en los años 1993 durante los pentadales 20 y 55 que representa la mitad del mes de Marzo y Septiembre. En el año 1998, posterior al pentadal 10 correspondiente al mes de Febrero, es decir, en todo el período de análisis la imagen de satélite proporcionó un aporte significativo para explicar la precipitación de estaciones climáticas. En 1999 después del pentadal 5, que se refiere al mes de Enero. En el 2000, después del pentadal 12 que comprende inicios de Marzo. En el 2001 entre el pentadal 1 y el pentadal 4 correspondiente al mes de Enero. En los años 2003 y 2006 después del pentadal 25 que comprende el mes de Abril. En el año 2007, después del pentadal 15 que comprende la mitad del mes de Marzo. En el 2008, 2009 y 2010, posterior al pentadal 10 que comprende la mitad de Febrero respectivamente. En el año 2011, el aporte de la precipitación proveniente de la imagen de satélite fué importante en todo el período de estudio. En la Tabla 9.1, se observan los aportes significativos por día pentadal, los años de estudio y si hubo episodio de El Niño o La Niña. En los años referenciados con (*), se destacan aportes importantes que coinciden con episodios de La

Niña (Weather Services, 2015).

Los residuales estudentizados observados mostraron un patrón alto de variabilidad en los años 1993, 1995 y 1998. En el primer año, la variable explicativa funcional no tuvo un aporte importante, mientras que en los dos últimos se caracterizaron por tener un aporte relevante debido a la precipitación procedente de imagen de satélite y además en esos años se presentaron episodios de La Niña. Así mismo, se denota que siempre hubo una tendencia creciente a lo largo del tiempo debido al proceso acumulativo de la precipitación para formar tanto la respuesta como la variable explicativa (Figura 9.4). Finalmente, con las curvas residuales se caracterizó la estructura de correlación espacial, este proceso fue realizado para cada año de estudio.

Se observó una relación directa entre los datos de estaciones climáticas y de satélite en los primeros días del año, ya que la curva de $\beta_1(t)$ fué creciente en esa parte del año, sin embargo, en los siguientes días pentadales, la relación entre ambos tipos de datos se mantuvo constante (Figura 9.2). Evidenciando un aporte significativo en los años donde el régimen de lluvia fué mayor, es decir, que la imagen de satélite puede ser un complemento importante para ajustar el dato en tierra en los períodos húmedos (Pentadales 18 a 30 que corresponden al período Marzo-Mayo y los pentadales 54 a 66 que corresponden al período Septiembre- Noviembre).

Tabla 9.1: Aporte de la precipitación de satélite, en el período 2003-2011.

Año	Tiempo (día pentadal)	período	Episodio	
1993	20 - 25	Marzo - Septiembre	-	
1994	-	-	NIÑO	
1995	-	-	NIÑA	
1996	-	-	-	
1997	-	-	NIÑO	
1998	>10	Posterior a Febrero	NIÑA	*
1999	>5	Posterior a Enero	NIÑA	*
2000	>12	Posterior a Marzo	NIÑA	*
2001	01-04	Enero	NIÑA	*
2002	-	-	NIÑO	
2003	>25	Posterior a Abril	-	
2004	-	-	NIÑO	
2005	-	-	NIÑA	*
2006	-	-	NIÑO	
2007	>15	Posterior a Marzo	NIÑA	*
2008	>10	Posterior a Febrero	NIÑA	*
2009	>10	Posterior a Febrero	NIÑO	
2010	>10	Posterior a Febrero	NIÑA	*
2011	01-72	Todo el período	NIÑA	*

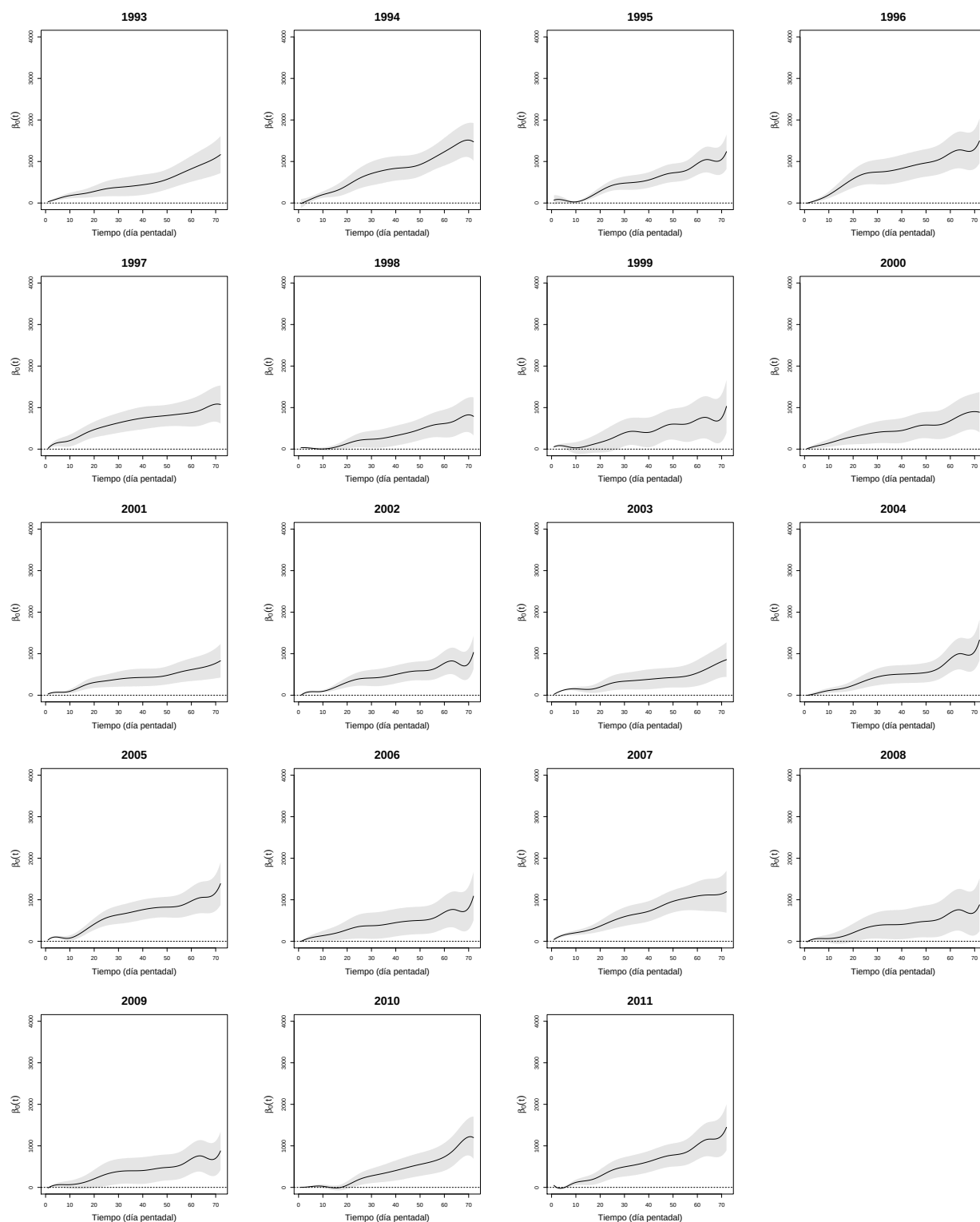


Figura 9.1: Intercepo funcional, en el período 1993-2011.

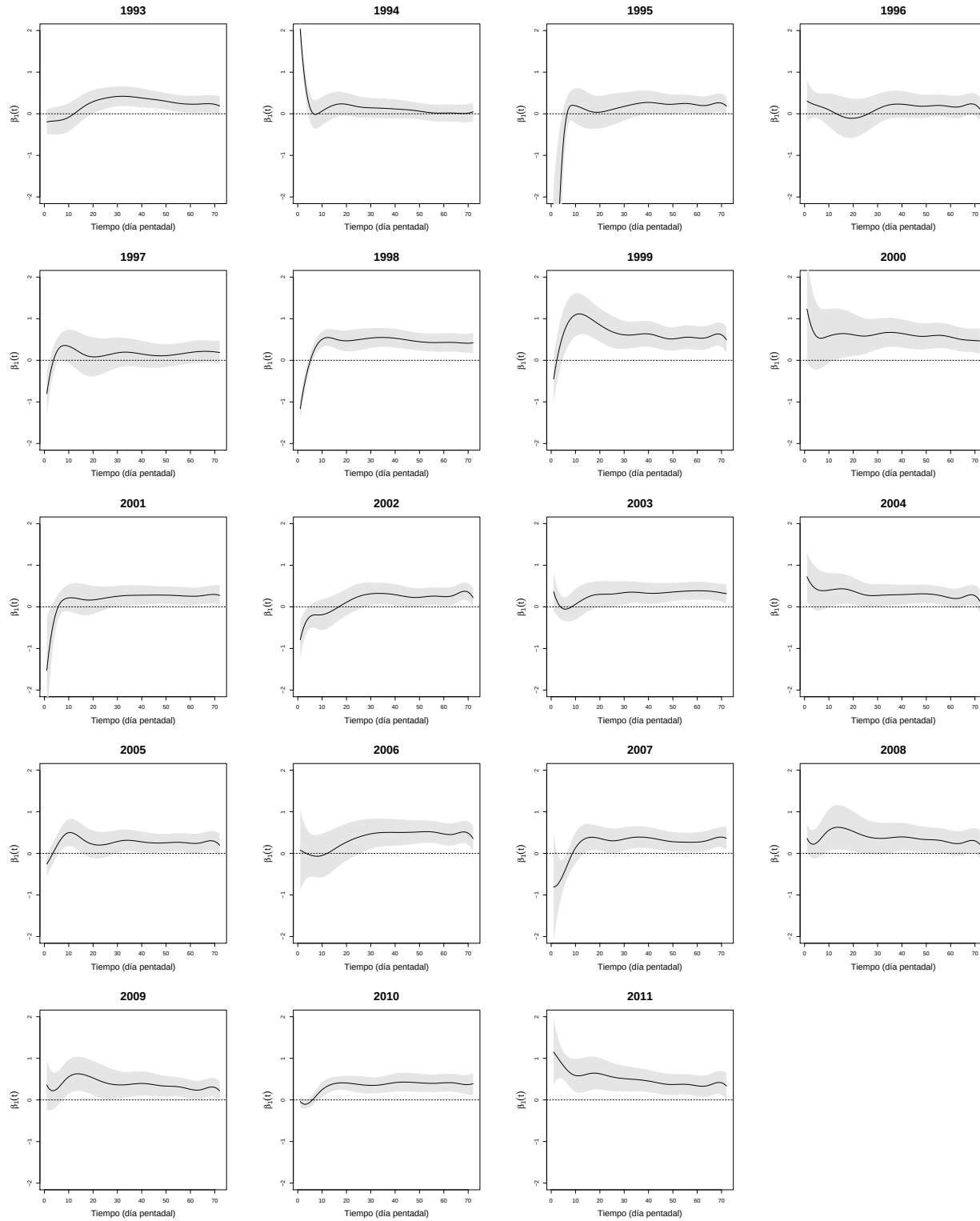


Figura 9.2: Coeficiente de regresión asociado a los datos de satélite, en el período 1993-2011.

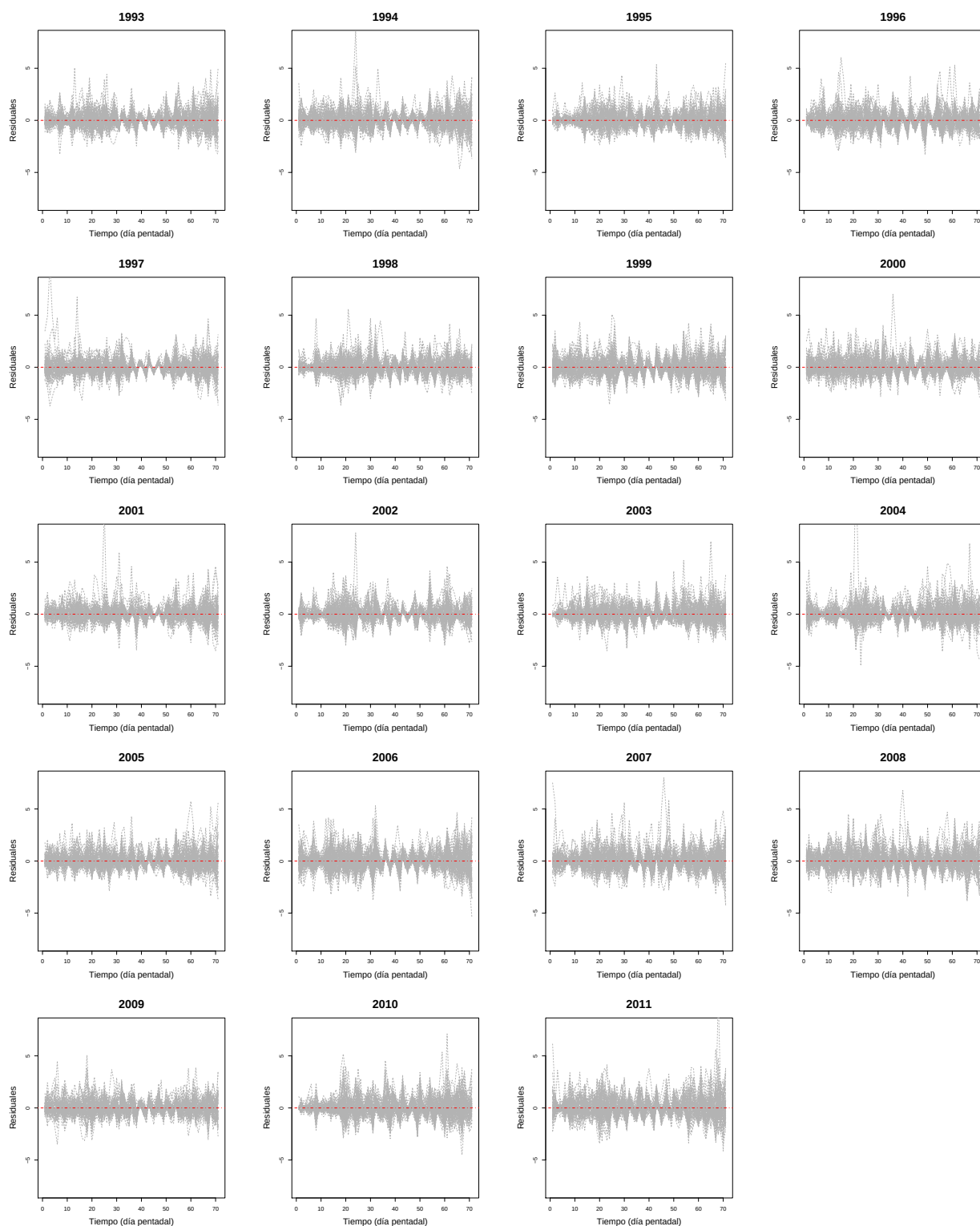


Figura 9.3: Residuales estudentizados del modelo ajustado, en el período 1993-2011.

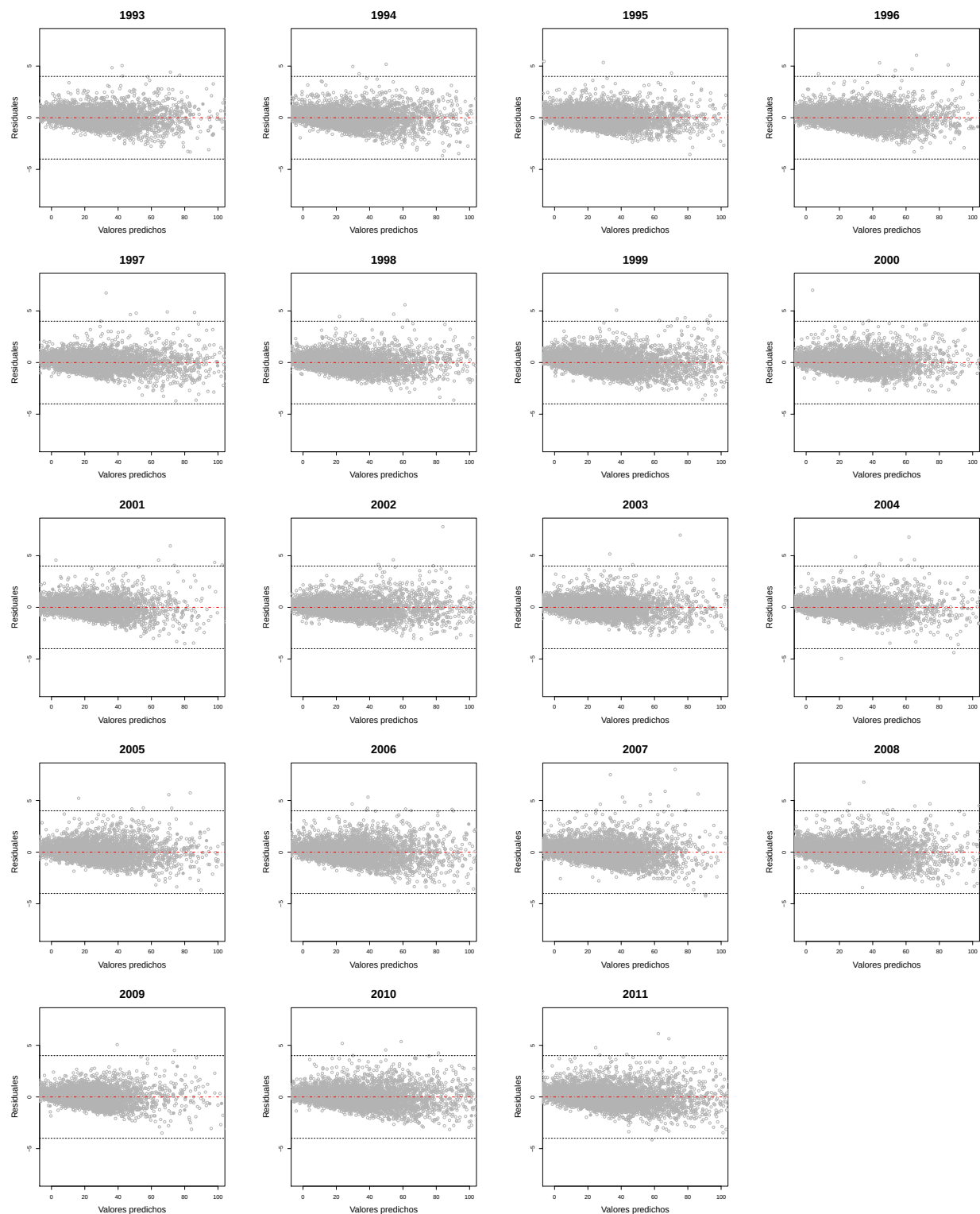


Figura 9.4: Residuales versus valores predichos, en el período 1993-2011.

9.3. Conclusiones

De acuerdo con los resultados obtenidos en este capítulo, se observaron contribuciones importantes de la información procedente de los sensores remotos, cuando se incorporó al modelo la estructura de correlación espacial de los errores (Figura 9.2).

Teniendo como referencia los episodios del fenómeno de el NIÑO y la NIÑA, se destacan aporte significativos cuando hubo episodios de la NIÑA, es decir, que hubo una mayor contribución del modelo propuesto cuando hay un aumento en los registros de precipitación (Tabla 9.1).

Los residuales estudentizados del modelo ajustado estuvieron alrededor de cero, sin embargo, en los años 1993, 1995, 1996 y 1998 mostraron un incremento al final del período, este comportamiento puede ser causado por el proceso acumulativo de los datos de precipitación (Figura 9.3 y Figura 9.4).

Capítulo 10

Predicción espacial

En este capítulo se muestran los resultados del nivel predictivo del modelo ajustado, a partir del coeficiente de determinación funcional y la Suma de Cuadrados del Error (SCE_{VC}) obtenida por validación cruzada, en el período 1993-2011. En la segunda parte, se muestra el ajuste de las curvas y los mapas de precipitación acumulada para el último año de análisis, utilizando el modelo propuesto en el capítulo 9.

10.1. Evaluación del nivel predictivo del modelo

En la Figura 10.1 se muestra el comportamiento del coeficiente de determinación funcional $R^2(t)$ para los modelos con correlación espacial en los errores y sin correlación espacial en los errores. Se destacan incrementos en $R^2(t)$ en el modelo ajustado con correlación espacial, después del pentadal 20. A nivel general observó que fue mayor el coeficiente de correlación funcional en los años: 1993, 1994, 1995, 1998, 1999, 2000, 2001, 2003, 2004, 2006, 2007, 2010. Es decir, que se mejoró el nivel predictivo del modelo en 12 de los 19 años de estudio (63.1 %).

En la Tabla 10.1 se muestra el comportamiento de la SCE_{VC} obtenidos por validación cruzada en el período 1993-2011, comparando entre el modelo con correlación espacial en los errores y el modelo sin correlación espacial en los errores. En una primera fase, se obtuvo la SCE_{VC} para las 82 curvas y se calcularon las medianas para cada año, posteriormente se compararon los centramientos con la prueba de Mann–Whitney, donde se observaron valores menores de la SCE_{VC} con contribuciones importantes ($p < 0.05$) del modelo con correlación en los años: 1994 ($p = 0.009$), 1995 ($p = 0.05$), 1997 ($p = 0.028$), 1999 ($p = 0.039$), 2002 ($p = 0.029$), 2003 ($p = 0.023$), 2004 ($p = 0.070$), 2007 ($p = 0.032$) y 2011 ($p = 0.041$). Este resultado sugiere que se mejoró el nivel predictivo del modelo, cuando se incluyó la correlación espacial en los errores.

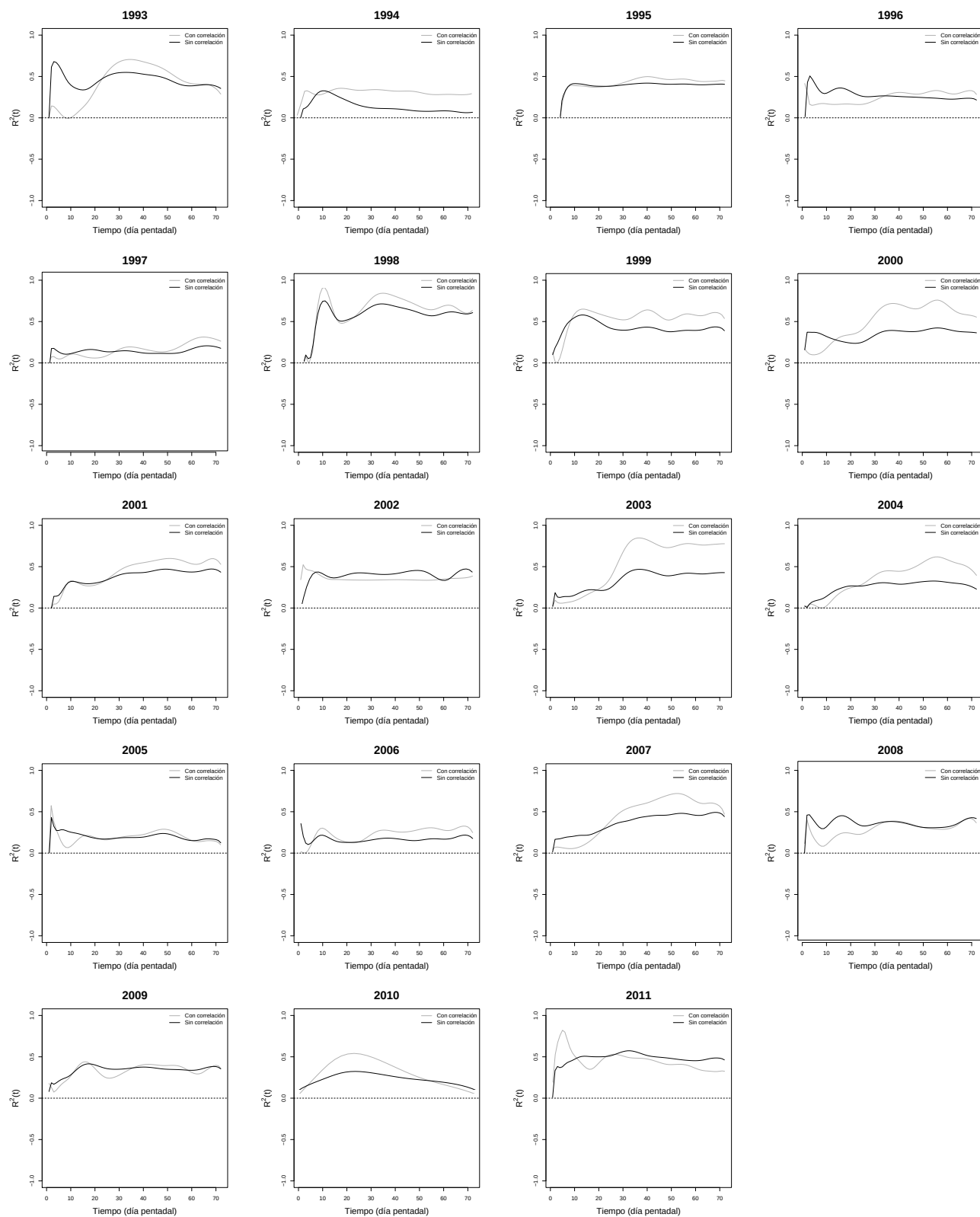


Figura 10.1: Curva de $R^2(t)$ funcional, con correlación espacial en los errores y sin correlación espacial en los errores, en el período 1993-2011.

Tabla 10.1: Comportamiento del SCE_{VC} (*Mediana*) con correlación espacial en los errores y sin correlación espacial en los errores , en el período 1993-2011.

Año	Con correlación	Sin Correlación	p	
1993	3.18	4.09	0.083	
1994	3.86	4.14	0.017	*
1995	3.04	1.55	0.259	
1996	3.96	7.41	0.005	*
1997	3.01	5.47	0.043	*
1998	3.21	2.33	0.271	
1999	4.62	8.12	0.036	*
2000	3.74	6.60	0.016	*
2001	2.60	2.95	0.179	
2002	2.17	2.70	0.256	
2003	2.90	3.33	0.032	*
2004	2.68	2.79	0.555	
2005	3.19	2.96	0.061	*
2006	4.73	6.75	0.021	*
2007	4.40	4.65	0.308	
2008	5.26	8.00	0.041	*
2009	3.49	6.06	0.039	*
2010	3.16	6.60	0.039	*
2011	4.93	4.57	0.182	

La Figura **10.2** muestra el gráfico de dispersión de las 82 curvas ajustadas por validación cruzada y observadas, en el período 1993 a 2011. Se observaron buenas estimaciones en la mayoría de las estaciones de estudio, sin embargo, hubo algunas localizaciones en las que no se evidencia un comportamiento lineal entre los datos observados y ajustados.

En términos generales, las predicciones obtenidas por validación cruzada muestran el mismo comportamiento que las curvas observadas, indicando un buen desempeño del modelo propuesto, el cual puede ser utilizado para la modelación de series históricas de precipitación, a partir de la metodología propuesta.

Los datos funcionales de la precipitación de estaciones climáticas (Observados) y los datos funcionales predichos con validación cruzada (Estimados) con correlación espacial en los errores, para 20 estaciones climáticas en el año 2011 mostraron el mismo comportamiento (Figura **10.3**).

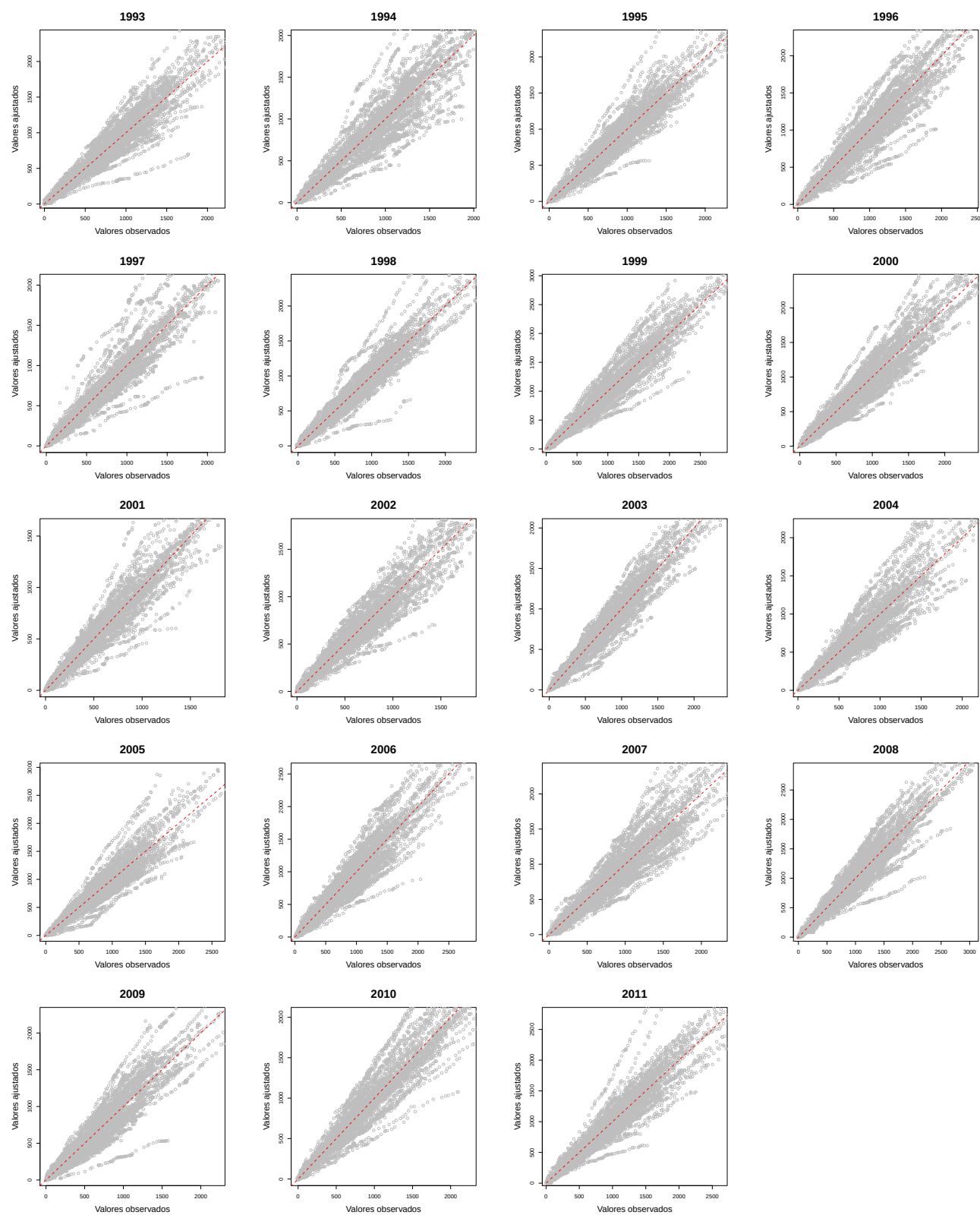


Figura 10.2: Comportamiento de las curvas observadas y ajustadas, en el período 1993-2011.

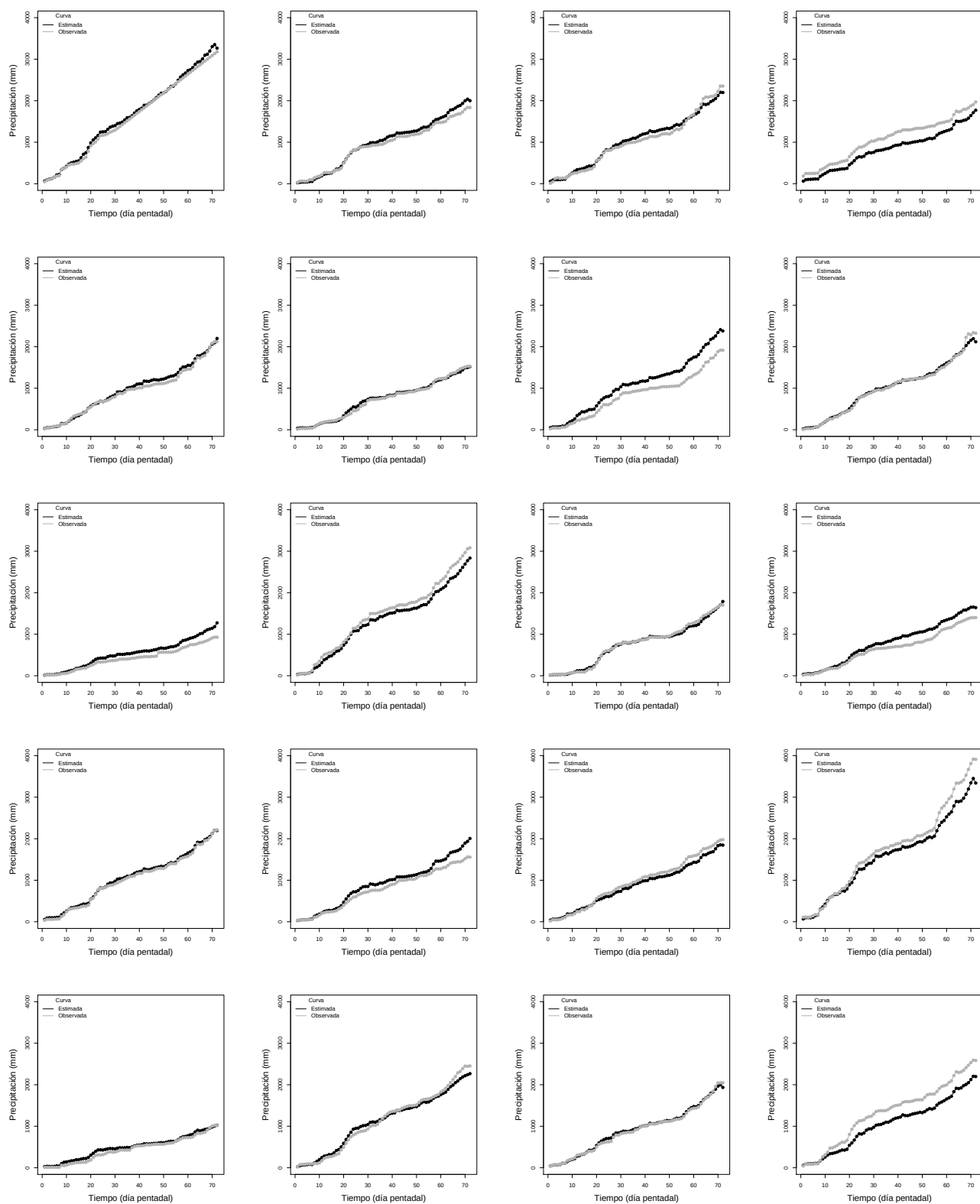


Figura 10.3: Comparación entre las curvas observadas y estimadas para 20 estaciones climáticas, para el año 2011.

10.2. Mapas de predicción con el modelo propuesto

A continuación se presentan los mapas de predicción para cada estación climática en el período 1993-2011, a partir de la Raíz del Error Cuadrático Medio (*RECM*).

En la Figura 10.4 se muestra el comportamiento de la predicción del modelo para el período 1993-1998. Se destacan *RECM* alrededor de 0.7mm y 2.2mm con respecto a las series de precipitación observadas, si comparamos estas diferencias con respecto al rango de variación de la precipitación puntual (0 a 300 mm), se obtiene un error de predicción del 0.23 % y 0.73 % respectivamente. Las diferencias más altas fueron encontradas en los años 1995, 1996 y 1997, con diferencias máximas de 1.7mm (0.57 %), 2.2mm (0.73 %) y 2.2mm (0.73 %) respectivamente.

En el período 1994-2004, la mayoría de las estaciones tuvieron valores de *RECM* alrededor 0.7mm (0.23 %) y 1mm (0.73 %), es decir, denotando un buen desempeño del modelo propuesto en este período. Además, se encontraron diferencias máximas en los años 2001 y 2005 con valores de $RECM = 2.2$ y $RECM = 2.4$ respectivamente (Figura 10.5).

Finalmente, en el período 2005-2011 los resultados mostraron un comportamiento óptimo en el nivel predictivo, con valores de *RECM* alrededor de 0.71mm (0.24 %) y 1.2mm (0.4 %) en la mayoría de las estaciones climáticas evaluadas (Figura 10.6).

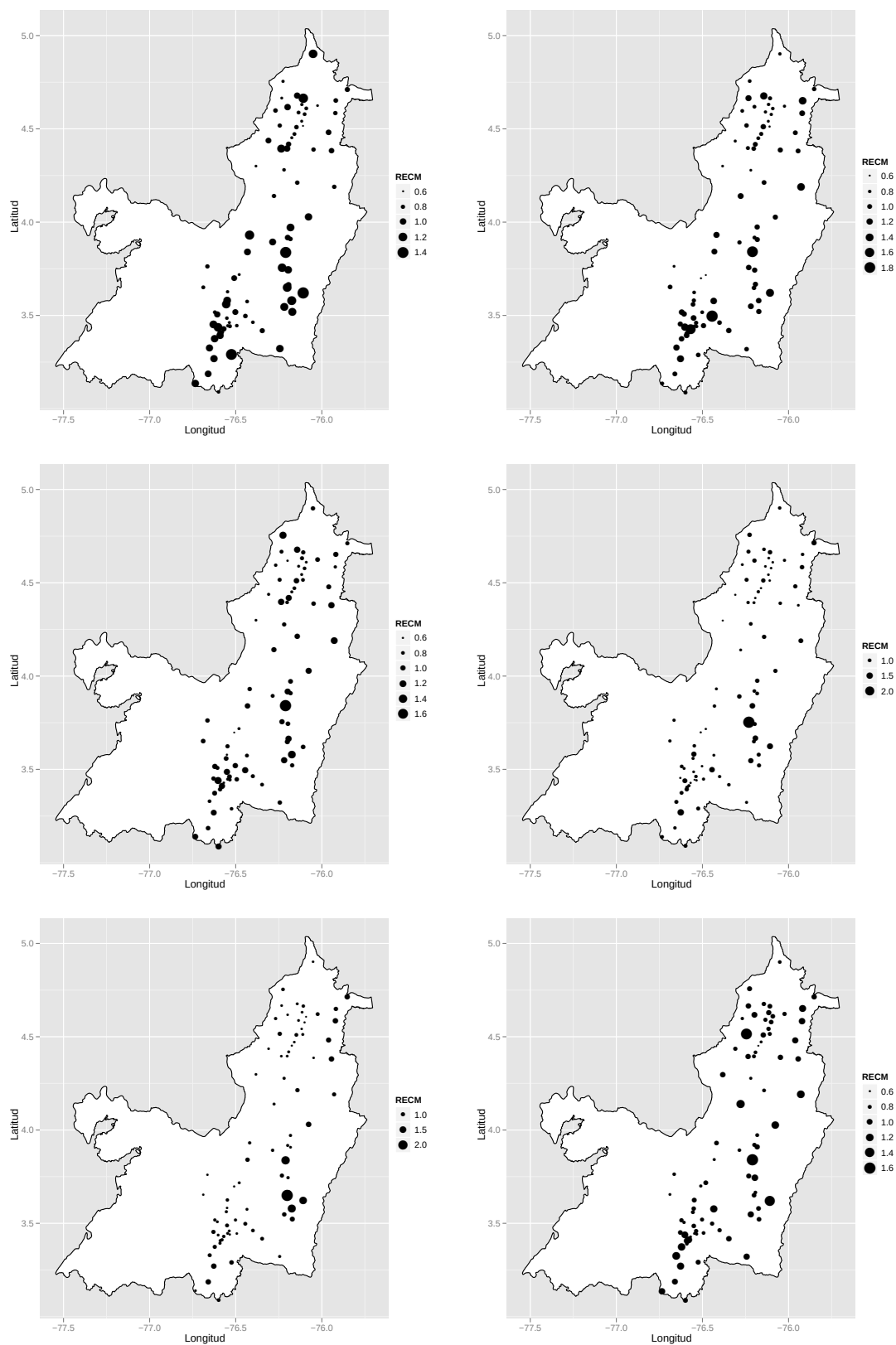


Figura 10.4: Comportamiento del *RECM*, en el período 1993-1998.

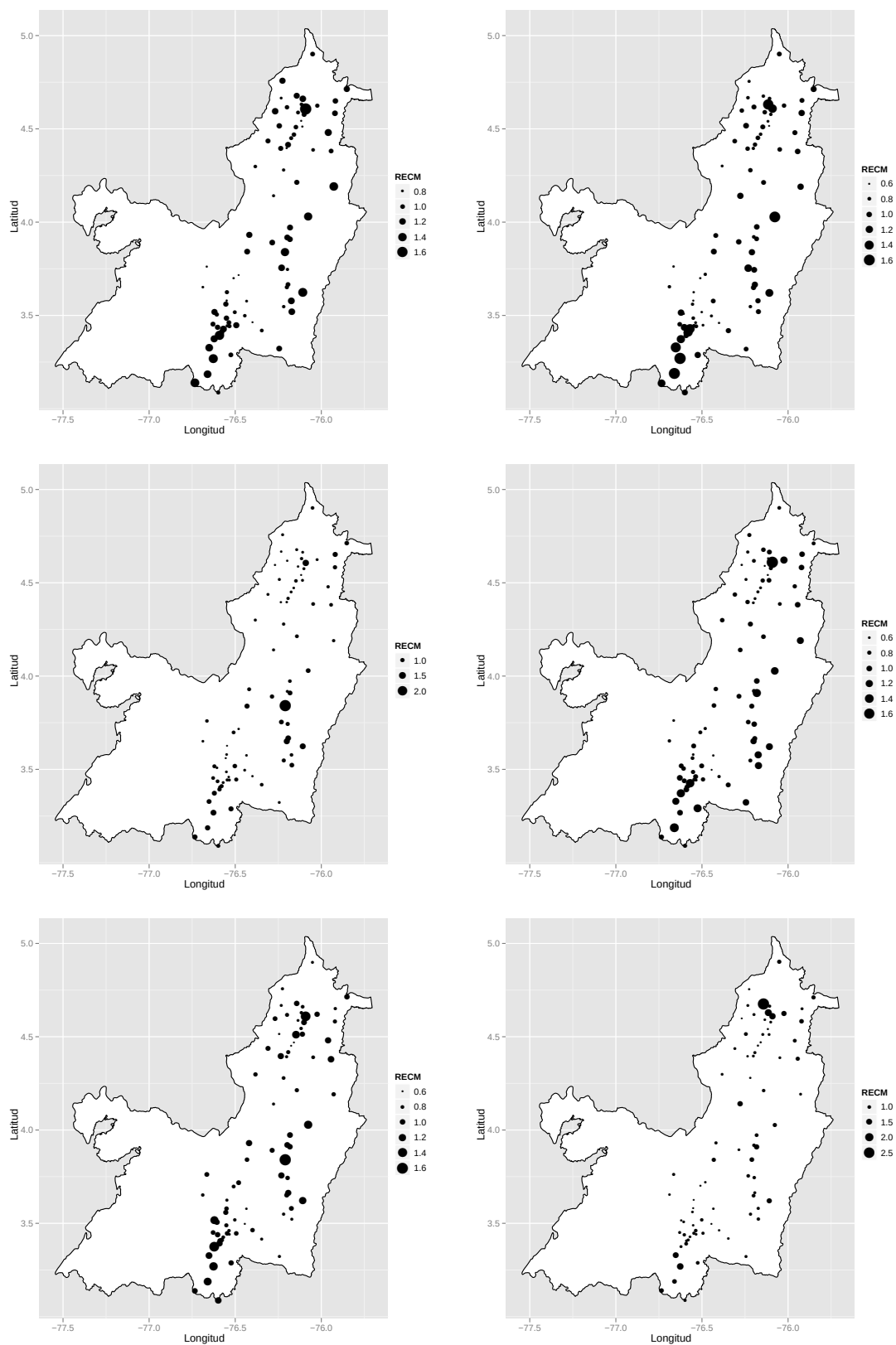


Figura 10.5: Comportamiento del *RECM*, en el período 1999-2004.

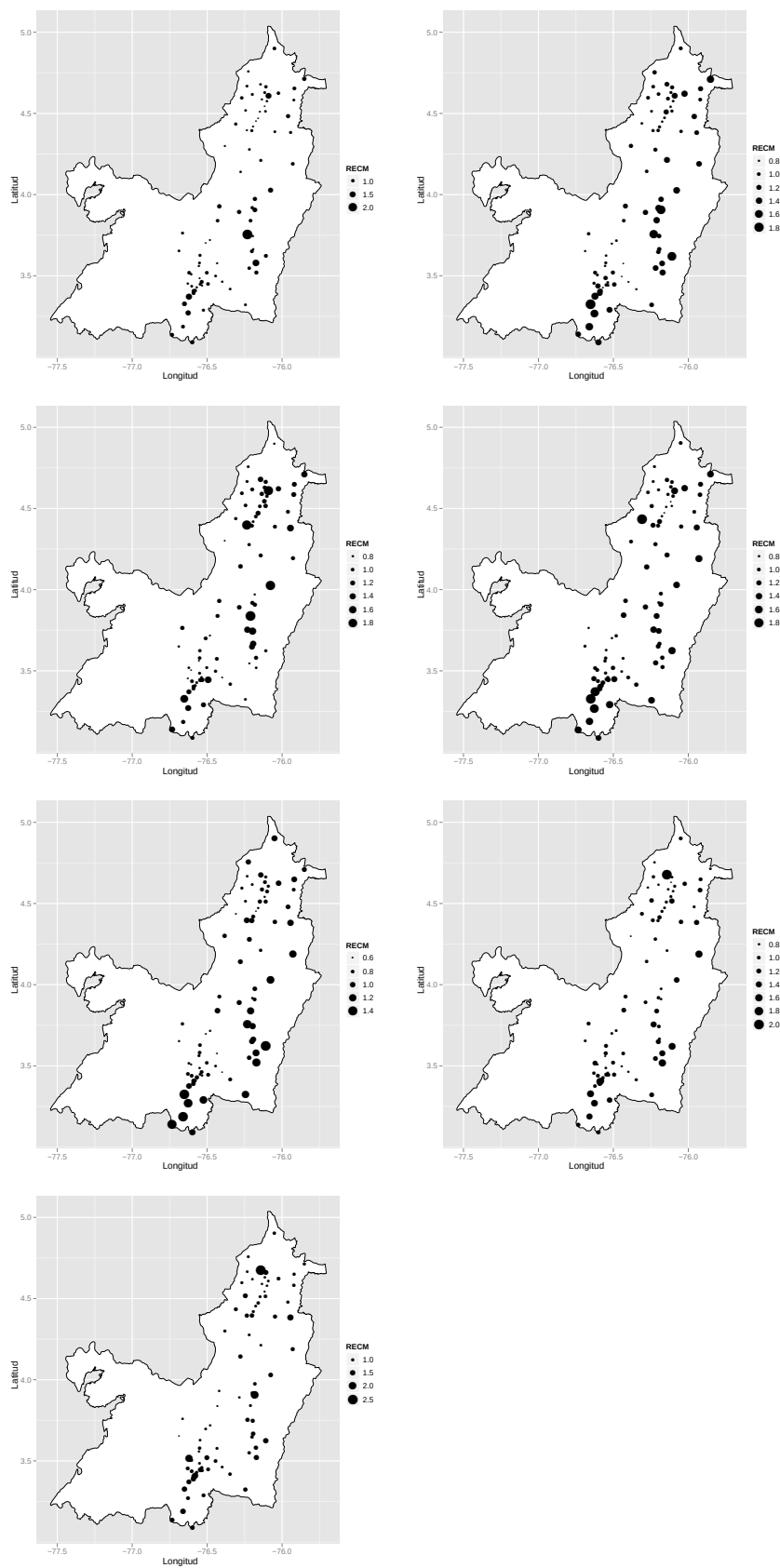


Figura 10.6: Comportamiento del *RECM*, en el período 2005-2011.

10.3. Conclusiones

El modelo propuesto tuvo un mejor desempeño en el primer mes del año 2011, esto se evidencia con el coeficiente de determinación funcional $R^2(t)$ que tuvo mejor desempeño cuando se incorporó la correlación espacial en los errores (Figura 10.1).

Al tener en cuenta la correlación espacial de los errores, se mejoró la capacidad predictiva en el 52.6 % de los años de estudio, ya que hubo una disminución importante en los valores de SCE_{VC} por año (Tabla 10.1).

En términos generales las mayoría de las estaciones mostraron una relación lineal entre las series de precipitación observadas y ajustadas con el modelo propuesto (Figura 10.3).

Los mapas de predicción de la precipitación por estación climática, mostraron un mejor desempeño en la zona central del Valle del Cauca, al incorporar la información del satélite. Igualmente, el *RECM* mostró variaciones pequeñas en la mayoría de las estaciones climáticas. Por otro lado, los resultados del *RECM* no mostraron incrementos de acuerdo a la dirección de las estaciones climáticas, indicando que el modelo propuesto no es afectado por la localización de las estaciones climáticas evaluadas (Figura 10.4, Figura 10.5 y Figura 10.6).

Capítulo 11

Conclusiones

En este proyecto se desarrolla una metodología para la estimación la precipitación, a partir del ajuste de un modelo de regresión funcional con estructura de correlación espacial en los errores, para la precipitación pentadal acumulada en el Valle del Cauca, durante el período 1993-2011, a partir de la integración de la información puntual disponible (variable respuesta) e imágenes de satélite (variable predictora). El modelo propuesto permite avanzar en el entendimiento de la variable de estudio y en general a la identificación de patrones de comportamiento que ella presenta.

11.1. Discusión

Incorporar la información de precipitación que provee el satélite contribuyó a mejorar la estimación de precipitación en la mayor parte del período de estudio, proporcionando una herramienta adicional en la estimación de la precipitación en sitios no observados en el Valle del Cauca. Estos resultados son de suma importancia en zonas emergentes cuando las redes de monitoreo en tierra son escasas, ya que la medición a través de satélite logra tener en tiempo real información de precipitación con una cobertura total sobre la región de estudio.

Las metodologías utilizadas para la modelación de la precipitación basadas en métodos de interpolación Grimes et al. (1999), Teo & Grimes (2007), Méndez Venegas (2008), Alvarez-Villa et al. (2011), Verdin et al. (2015) y basadas en modelos de regresión lineal mixto Andrade et al. (2012), sólo tienen en cuenta la información de las estaciones climáticas y realizan la predicción en toda la zona de estudio, además utilizan valores resúmenes como el promedio mensual o promedio anual para caracterizar el comportamiento en un largo período de tiempo, esto tiene como consecuencia la pérdida de información sobre la variable de estudio y el aumento en el número de modelos que deben ajustar, siendo menos eficiente cuando el objetivo es la predicción de la serie de precipitación completa en diferentes localizaciones de una zona de estudio.

Dentro del contexto del ADF, la metodología propuesta en este trabajo de investigación, es un aporte en la investigación de la modelación de datos funcionales con estructuras de correlación. Trabajos como Scheipl et al. (2015) abordan la modelación de la precipitación a partir del modelo funcional aditivo mixto considerando la correlación espacial, mientras que Suhaila et al. (2011) asumen independencia entre las curvas de precipitación evaluadas para diferentes localizaciones. En contraste, este trabajo de investigación propone un modelo de regresión funcional, integrando la información de sensores remotos y la correlación espacial de los errores debido a las estaciones climáticas.

En esta investigación se propone un marco teórico general para el modelo de regresión con errores funcionales correlacionados, permitiendo incorporar el efecto de la estructura de correlación espacial debido a las localizaciones de las estaciones climáticas. Dichas estructuras de dependencia se pueden modelar de forma implícita mediante la extensión de la geoestadística para datos funcionales desarrollada en Giraldo (2009). La estimación del parámetro funcional asociado a la variable predictora funcional es realizada a través de la teoría de mínimos cuadrados generalizados, permitiendo integrarla con la robustez y flexibilidad de los algoritmos desarrollados para la estimación del modelo de regresión funcional clásico propuesto en Ramsay & Silverman (2005).

En esta investigación el modelo funcional lineal se ha extendido de una manera directa al caso de predictores funcionales y respuesta funcional con errores correlacionados espacialmente, aplicado a datos de precipitación de estaciones climáticas y de satélite. En este contexto, se encontró una interesante relación entre la información suministrada por los sensores remotos y estaciones en tierra: hay una contribución mayor por parte de la imagen de satélite cuando aumentan los niveles de precipitación, sin embargo, se presenta un efecto en la variabilidad espacial debido a las estaciones, puesto que el nivel de correlación espacial se hace menos evidente a medida que aumentan los niveles de precipitación.

11.2. Conclusiones generales

El Análisis de Datos Funcionales es una metodología que puede representar datos de precipitación acumulada en forma de funciones suaves o curvas que dan información significativa. En este estudio, se empleó la metodología del ADF para representar datos de precipitación en forma de funciones suaves para la región del Valle del Cauca, en el período 1993-2011. Los resultados de este estudio mostraron que diez y veinte funciones de base son necesarias para representar los patrones de precipitación proveniente de estaciones climáticas y de satélite respectivamente, reflejando una excelente adaptabilidad de las curvas estimadas al comportamiento de los datos observados (Tabla 6.1).

Por otra parte, la implementación del ADF para describir el comportamiento de la

precipitación en el Valle del Cauca permitió realizar un análisis exploratorio de las series completas por año y estudiar de forma conjunta dicho comportamiento, demostrando que el análisis descriptivo funcional puede emplearse en la evaluación y caracterización de curvas de precipitación en el Valle del Cauca (Figura 6.2 y Figura 6.3).

En general las curvas estimadas representaron muy bien el comportamiento de los datos observados, ya que las bandas de confianza siempre estuvieron constantes alrededor de la curva estimada (Figura 6.1). Los incrementos en la desviación estándar funcional se debe al proceso de acumulación por el cual se generaron los datos observados de estaciones climáticas y de satélite (Figura 7.1 y Figura 7.2).

Tener en cuenta la altura como una variable funcional sobre el modelo propuesto no generó una contribución importante para explicar la precipitación pentadal, sin embargo, este resultado se puede originar porque la información de la imagen se lleva toda la contribución en la explicación de la variable respuesta del modelo, ya que no tiene en cuenta la elevación del terrero.

En cuanto a la identificación de la estructura de correlación espacial más apropiada por año, se encontró que utilizar Geoestadística funcional es importante para caracterizar un patrón conjunto de correlación espacial en los errores funcionales, en lugar de ajustar modelos de semivariograma para cada estación y año de estudio. La importancia de aplicar este tipo de metodología en el presente trabajo de investigación fue que permitió identificar un patrón de correlación espacial de manera conjunta. Mostrando un modelo teórico apropiado para explicar la estructura de correlación espacial para cada año (Tabla 8.3). En general, se evidenció la presencia de correlación espacial hasta un rango de 30 a 50 kilómetros en el Valle del Cauca con la red de estaciones utilizadas para todo el período de estudio y el modelo más recurrente para identificar la correlación espacial fue el modelo exponencial (Figura 8.1 y Tabla 8.2).

Los resultados obtenidos del nivel predictivo del modelo propuesto, a través de validación cruzada mostraron un mejor desempeño al incorporar la correlación debida a las estaciones climáticas (Tabla 10.1), esto indica que puede ser adoptado como método válido para el modelación de la precipitación utilizando la regresión de datos funcionales con errores espacialmente correlacionados (Figura 10.2 y Figura 10.3). Los mapas mostraron un comportamiento aleatorio en el desempeño del *RECM*, esto indica que el nivel predictivo del modelo presentó un buen comportamiento independiente de la localización de las estaciones climáticas (Figura 10.4, Figura 10.5 y Figura 10.6).

11.3. Trabajo futuro

Se propone ajustar un modelo, discretizando la respuesta funcional en varios modelos escalares y aplicar a cada una de esas discretizaciones un modelo de regresión funcional con respuesta escalar y correlación en los errores.

Otro enfoque sería integrar al modelo funcional otras covariables que permitan mejorar la estimación de precipitación, tales como: imagen de temperatura y un modelo de elevación digital del Valle del Cauca.

Utilizando la información de este proyecto, se puede plantear mejorar la estimación de precipitación en tierra a partir de una técnica como el Cokring funcional y comparar las predicciones con los resultados obtenidos en este trabajo de investigación.

Bibliografía

- Aguilera, M. d. C. (2009). Estimación penalizada de datos funcionales. Tesis de Maestría - Universidad de Granada.
- Alvarez-Villa, O. D., Velez, J. I., & Poveda, G. (2011). Improved longterm mean annual rainfall fields for Colombia. *International Journal of Climatology*, 31(14), 2194–2212.
- Andrade, M., Conde, G., & Flórez, A. (2012). Modelación estadística de variables climáticas en el suroccidente colombiano. Proyecto Convocatoria Interna 2617. Universidad del Valle.
- Armstrong, M. (1998). *Basic Linear Geostatistics*. Berlín, Alemania: Springer. 155p.
- BIOPALMIRA (2008). Avance de los temas de investigación clima, biodiversidad y calidad del hábitad. http://www.idea.palmira.unal.edu.co/paginas/proyectos/paginas/avances_invest.pdf.
- Carvajal, Y., Jiménez, H., & Materón, H. (1998). Incidencia del fenómeno ENSO en la hidroclimatología del Valle del río Cauca-Colombia. *Bulletin de l'Institut Francais d'Études Andines*, 27(3), 743–751.
- Cerón, W. L., Trujillo, A. R., & Escobar, Y. C. (2011). Modelo para el monitoreo y seguimiento de indicadores de sostenibilidad del recurso hídrico en el sector agrícola. *Cuadernos de Geografía: Revista Colombiana de Geografía*, 20(2), 77–89.
- Chen, F.-W. & Liu, C.-W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (idw) in the middle of taiwan. *Paddy and Water Environment*, 10(3), 209–222.
- Coe, R. & Stern, R. (1982). Fitting models to daily rainfall data. *Journal of Applied Meteorology*, 21(7), 1024–1031.
- Cowpertwait, P., O'Connell, P., Metcalfe, A., & Mawdsley, J. (1996). Stochastic point process modelling of rainfall. i. single-site fitting and validation. *Journal of Hydrology*, 175(1), 17–46.

- Cressie, N. (1993). *Statistics for spatial data (Wiley Series in Probability and Statistics)*. New York, USA: Wiley-Interscience. 928p.
- 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.
- DANE (2011). Informe de coyuntura económica en el Valle del Cauca. Departamento Administrativo Nacional de Estadísticas (DANE). http://www.dane.gov.co/files/icer/2011/valledelcauca_icer__11.pdf.
- DANE (2013). Informe de coyuntura económica en el Valle del Cauca. Departamento Administrativo Nacional de Estadísticas (DANE). http://www.dane.gov.co/files/icer/2011/valledelcauca_icer_13.pdf.
- Delicado, P. (2008). Curso de modelos no paramétricos. *Departamento de Estadística Investigación Operativa, Universidad Politécnica de Cataluña*. http://www-eio.upc.es/~delicado/docencia/Apuntes_Models_No_Parametrics.pdf.
- Escobar Carmona, S. T. (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.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. New York, USA: CRC press. 360p.
- Ferraty, F. & Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. New York, USA: Springer. 260p.
- Funk, C. & Verdin, J. (2003). Comparing satellite rainfall estimates and reanalysis precipitation fields with station data for western Kenya. In *Proc. Int. Workshop on Crop Monitoring for Food Security in Africa, European Joint Research Centre UN Food and Agriculture Organization, Nairobi, Kenya* (pp. 28–30).
- Galindo, G., Cabrera, E., & Londoño, C. (2005). Spatial analysis to determine priority conservation areas of dry ecosystems in two interandean valleys of valle del cauca-colombia. *Lyonia 8 (2) 2005-Dry Forest Biodiversity and Conservation 1: Biodiversity*.
- Garcés, J. & Santana, L. M. (2005). Aplicación de tecnologías de información geográfica para el estudio de la variabilidad climática en la cuenca alta del río Cauca. reporte de investigación - Universidad del Valle. Universidad del Valle.

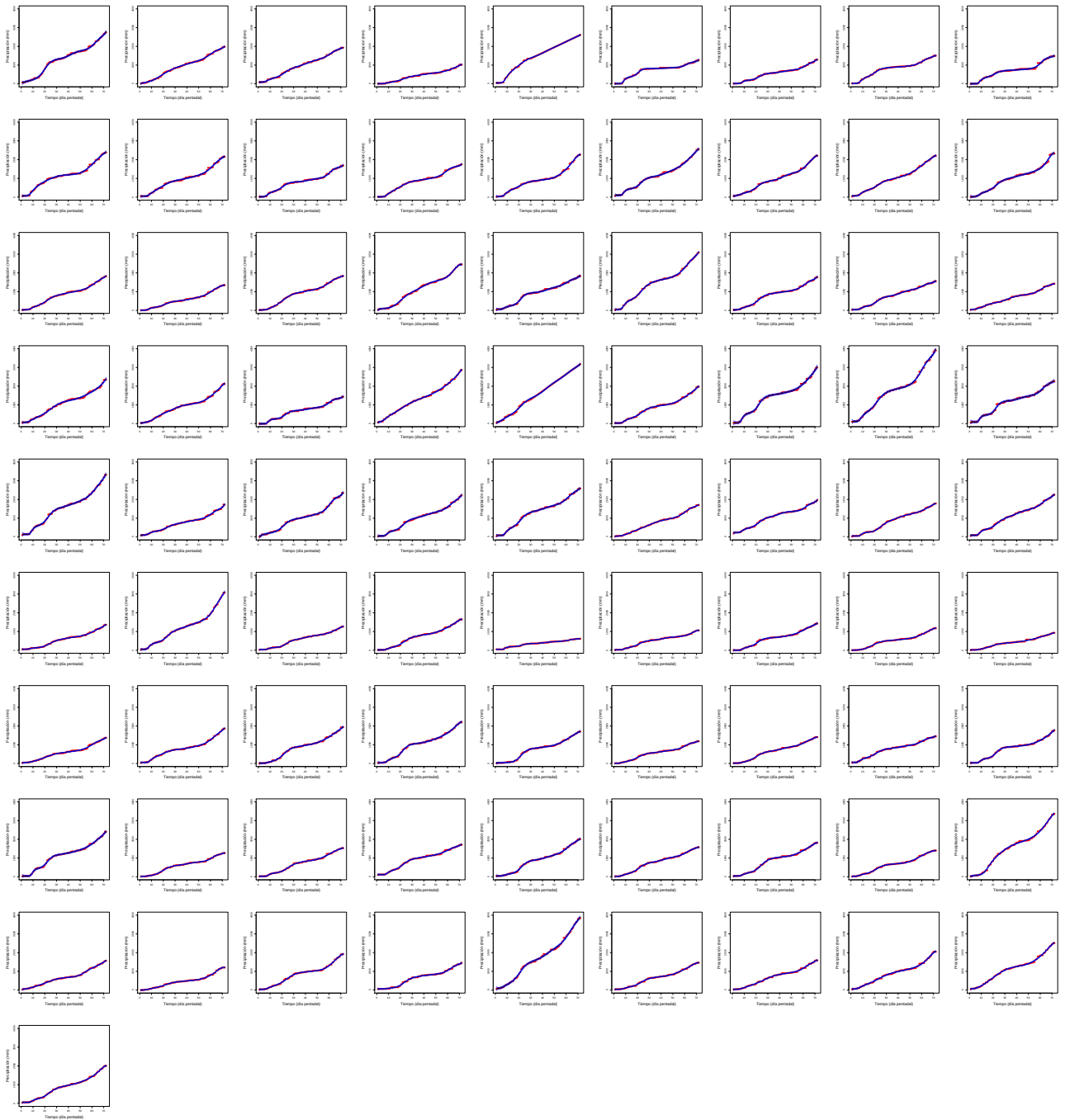
- Giraldo, H. (2002). Introducción a la geoestadística: Teoría y aplicación. *Universidad Nacional de Colombia, sede Bogotá, Bogotá DC*.
- Giraldo, R. (2009). *Geostatistical analysis of functional data*. PhD thesis, PhD thesis, Universitat Politècnica de Catalunya.
- Grimes, D., Pardo-Iguzquiza, E., & Bonifacio, R. (1999). Optimal areal rainfall estimation using raingauges and satellite data. *Journal of Hydrology*, 222(1), 93–108.
- Hughes, J. P., Guttorp, P., & Charles, S. P. (1999). A non-homogeneous hidden markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1), 15–30.
- IGAC (2012). Instituto Geográfico Agustín Codazzi (IGAC). http://geoportal.igac.gov.co/mapas_de_colombia/igac/mps_fisicos_deptales/2012/Valle_del_Cauca.pdf.
- Kolmogorov, A. N., Fomin, S. V., & Fomin, S. (1957). *Elements of the theory of functions and functional analysis: Volumen I metric and normed spaces*. New York, USA: Graylock Press. 280p.
- Lall, U., Rajagopalan, B., & Tarboton, D. G. (1996). A nonparametric wet/dry spell model for resampling daily precipitation. *Water Resources Research*, 32(9), 2803–2823.
- Littell, R. C., Stroup, W. W., Milliken, G. A., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models*. NC, USA: SAS institute. 795p.
- Manteiga, W. G. & Vieu, P. (2007). Statistics for functional data. *Computational Statistics & Data Analysis*, 51(10), 4788–4792.
- Méndez Venegas, J. (2008). Modelación de la distribución espacial de la precipitación en el valle de la ciudad de México usando técnicas geoestadísticas. Tesis de maestría - Colegio de Postgraduados.
- Moazami, S., Golian, S., Kavianpour, M. R., & Hong, Y. (2014). Uncertainty analysis of bias from satellite rainfall estimates using copula method. *Atmospheric Research*, 137(0), 145 – 166.
- Noguerales, J. L. T., González, A. C., & Fernández, C. S. C. (2010). Análisis de datos funcionales, clasificación y selección de variables. Tesis de maestría. Universidad Autónoma de Madrid.
- Norliyana, W. W. & Suhaila, J. (2015). Smoothing wind and rainfall data through functional data analysis technique. *Jurnal Teknologi*, 74(1).

- Olaya, J. (2012). *Métodos de regresión no paramétrica*. Cali, Colombia: Editorial: Universidad del Valle. 109p.
- OMM (2001). La enseñanza y formación profesional del personal en meteorología e hidrología operativa. Volumen I: Meteorología. Organización Meteorológica Mundial (OMM). ftp://ftp.wmo.int/Documents/MediaPublic/Publications/WMO_258_Volume_I/WMO_258_Volume_I_es.pdf.
- Pedrerros, D. (2012). Rainfall enhancement process. Famine Early Warning Systems Network (FEWS NET) Climate Hazard Group, UCSB. Presentación en International Center for Tropical Agriculture CIAT.
- 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.
- Puertas Orozco, O. L., Carvajal Escobar, Y., & Quintero Angel, M. (2011). Estudio de tendencias de la precipitación mensual en la cuenca alta-media del río Cauca, Colombia. *Dyna*, 78(169), 112–120.
- Ramsay, J. & Silverman, B. (2005). *Functional data analysis*. New York, USA: Springer. 428p.
- Ramsay, J., Wickham, H., Graves, S., & Hooker, G. (2011). Fda: functional data analysis. *R package version*, 2(6).
- Ramsay, J. O. & Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. New York, USA: Springer. 191p.
- Ravelo, A. & Santa, J. (2000). Estimación de las precipitaciones utilizando información satelital y terrestre en la provincia de Córdoba (Argentina). *AgriScientia*, 17.
- Rojas, J. E. F. (2003). Comparacion entre la precipitación medida por satélite y estaciones de superficie en la costa norte del Perú. *Compendio de Trabajos de Investigación. CNDG - Biblioteca Instituto Geofísico del Perú.*, 4, 19–24.
- Scheipl, F., Staicu, A.-M., & Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2), 477–501.
- 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.
- Teo, C.-K. & Grimes, D. I. (2007). Stochastic modelling of rainfall from satellite data. *Journal of Hydrology*, 346(1–2), 33 – 50.

- Weather Services (2015). El Niño and la Niña years and intensities based on Oceanic Niño Index (ONI). <http://ggweather.com/enso/oni.htm>.
- Tomaso Poggio, L. R. & Frogner, C. (2010). Statistical learning theory and applications, spring 2010. Module: Functional analysis. <http://www.mit.edu/~9.520/spring10/Classes/mathcamp01.pdf>.
- Urrutia, J. A., Palomino, R., & Salazar, H. D. (2010). Metodología para la imputación de datos faltantes en metereología. *Scientia et Technica*, 3(46), 44–49.
- 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.
- Woldemeskel, F. M., Sivakumar, B., & Sharma, A. (2013). Merging gauge and satellite rainfall with specification of associated uncertainty across australia. *Journal of Hydrology*, 499(0), 167 – 176.
- Xie, P., Janowiak, J. E., Arkin, P. A., Adler, R., Gruber, A., Ferraro, R., Huffman, G. J., & Curtis, S. (2003). Gpcp pentad precipitation analyses: An experimental dataset based on gauge observations and satellite estimates. *Journal of Climate*, 16(13), 2197–2214.
- Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, I. (2013). Comparison of missing value imputation methods in time series: the case of turkish meteorological data. *Theoretical and applied climatology*, 112(1-2), 143–167.
- Zakaria, R. & Moslim, N. H. (2015). Independent mixed-gamma variables for modelling rainfall. *Mathematical Methods and Systems in Science and Engineering*, 41, 239–244.
- Zape, D. K. & Castañda, K. J. (2012). Determinación de la correlación espacial y temporal de la precipitación mensual en el Valle del Cauca, en el Periodo (1971-2009). Tesis de pregrado - Universidad del Valle.
- Zquiza, E. P.-i. (1998). Comparison of geostatistical methods for estimating the areal average climatological rainfall mean using data on precipitation and topography. *Int. J. Climatol*, 18, 1031–1047.

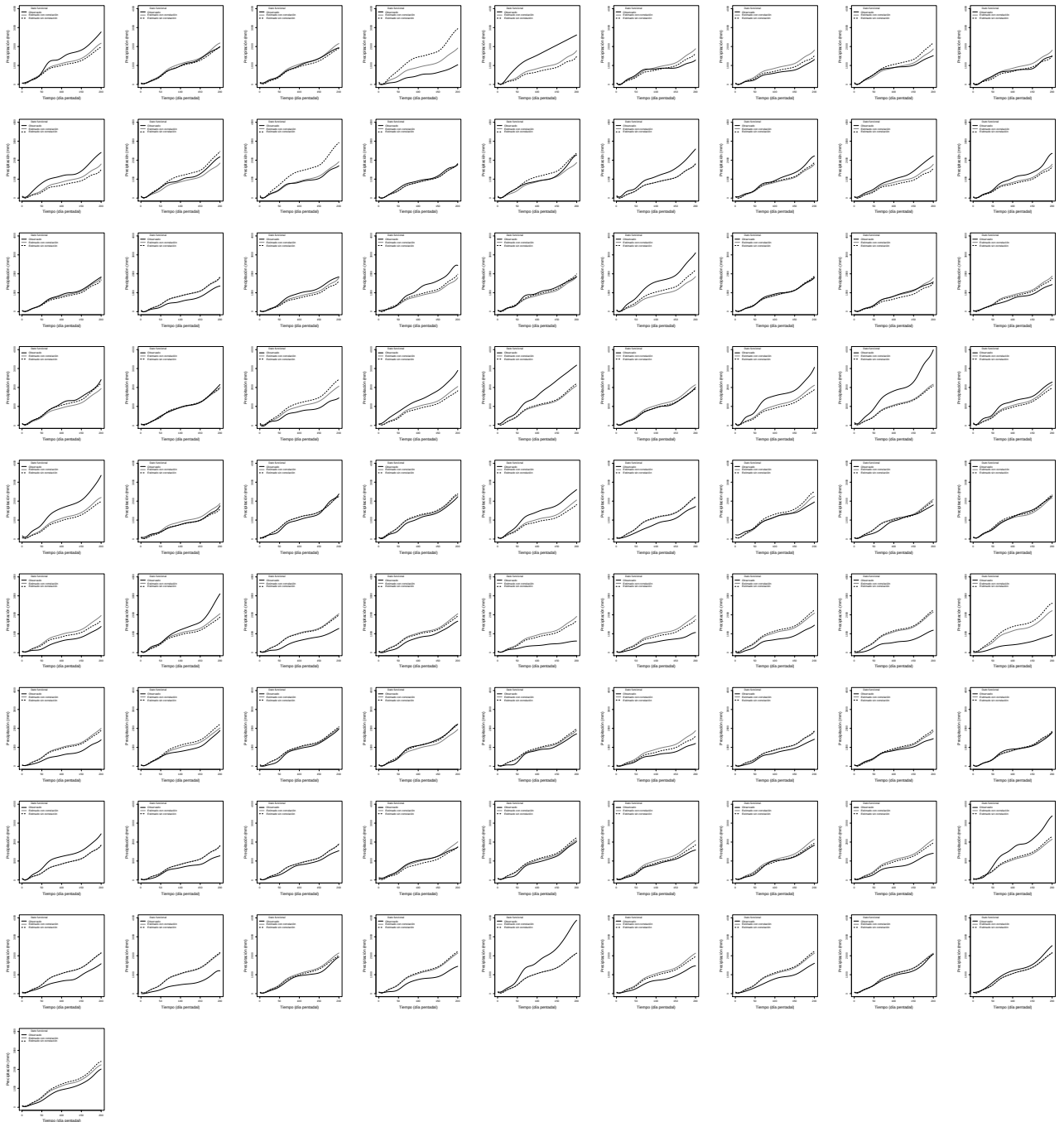
A. Curvas ajustadas y observadas de 82 estaciones, para el año 2011

En este capítulo se encuentran los resultados obtenidos de los datos funcionales de precipitación y las curvas de precipitación observadas de 82 estaciones climáticas analizadas en el año 2011. Las gráficas pueden obtenerse en el siguiente enlace https://drive.google.com/open?id=0B5w690V0_sa5RE1QUF9xNGxzc2M.



B. Modelo ajustado con correlación y sin correlación de 82 estaciones, para el año 2011

En este capítulo se encuentran los resultados obtenidos de las curvas de precipitación observadas y ajustadas sin correlación espacial en los errores y considerando la correlación espacial. Las gráficas pueden obtenerse en el siguiente enlace https://drive.google.com/open?id=0B5w690V0_sa5WlMtYVEzdUxiSUk.



C. Código en para estimar datos faltantes a través de geoestadística univariada

```
prediccion <- list()
precipitacion_promedio_geo <- list()
WheatherStation_llenada <- WheatherStation
prediccion <- 0

for(i in 1:19){
  precipitacion_promedio_geo[[i]]
  <- as.geodata(cbind(lat = PrecipitacionPromedio[[i]]$Lat,
                      lon = PrecipitacionPromedio[[i]]$Lon,
                      data = PrecipitacionPromedio[[i]]$prec_prom))
}

#Se imputan los datos faltantes en los 19 años
for(i in 1:19){
  for(j in 13:377){
    if(length(WheatherStation_llenada[[i]]
[is.na(WheatherStation_llenada[[i]][,j]),j]) != 0){
      prediccion <- 0
      prediccion <- kslide(precipitacion_promedio_geo[[i]],
                          coords = precipitacion_promedio_geo[[i]]$coords,
                          data = precipitacion_promedio_geo[[i]]$data,
                          cov.model="wave",
                          cov.pars=c(par_variogram[i,1], par_variogram[i,2]),
                          nugget = par_variogram[i,3],
                          locations =
                          cbind(WheatherStation[[i]][is.na(WheatherStation[[i]][,j]),12],
eval(parse(text=paste(
"WheatherStation_llenada[[i]][is.na(WheatherStation_llenada[[i]][,j]),j]",
"<-round(prediccion$predict, 0)", sep="")))
    }
  }
}
```

```
#Validación cruzada
ECM_Mod <- matrix(NA, nrow = 19, ncol = 4)

for(i in 1:19){
  cv_pr_Exp<- krige.cv(prec_prom~1, data = PrecipitacionPromedio[[i]],
    locations =~Lon+Lat,
    model = get(ls(pattern="vmod_Exp")[i]),
    nfold = nrow(PrecipitacionPromedio[[i]]))
  ECM_Mod[i,1] = mean(cv_pr_Exp$residual^2)

  cv_pr_Sph<- krige.cv(prec_prom~1, data = PrecipitacionPromedio[[i]],
    locations =~Lon+Lat,
    model = get(ls(pattern="vmod_Sph")[i]),
    nfold = nrow(PrecipitacionPromedio[[i]]))

  ECM_Mod[i,2] = mean(cv_pr_Sph$residual^2)

  cv_pr_Gau<- krige.cv(prec_prom~1, data = PrecipitacionPromedio[[i]],
    locations =~Lon+Lat,
    model = get(ls(pattern="vmod_Gau")[i]),
    nfold = nrow(PrecipitacionPromedio[[i]]))

  ECM_Mod[i,3] = mean(cv_pr_Gau$residual^2)

  cv_pr_Mat<- krige.cv(prec_prom~1, data = PrecipitacionPromedio[[i]],
    locations =~Lon+Lat,
    model = get(ls(pattern="vmod_Mat")[i]),
    nfold = nrow(PrecipitacionPromedio[[i]]))

  ECM_Mod[i,4] = mean(cv_pr_Mat$residual^2)
}
```

D. Código para la construcción de los datos funcionales de precipitación de estaciones climáticas y de satélite

```
preciMonEst      <- list(19)
WfdEst           <- list(19)
betafEst         <- list(19)
preciajusfdEst   <- list(19)
datos_evalEst    <- list(19)
DatosAjusEst     <- list(19)
preciMonSat      <- list(19)
WfdSat           <- list(19)
betafSat         <- list(19)
preciajusfdSat   <- list(19)
datos_evalSat    <- list(19)
DatosAjusSat     <- list(19)

#Datos funcionales - Estación climática
pdia             = 1:72
npdia            = length(pdia)
pdiaRng          = range(pdia)
nfine            = 72
pdiafine         = seq(pdiaRng[1], pdiaRng[2], length = nfine)
nindiv           = dim(DatosObsEst[[1]])[2]

# Orden y número de funciones de base
nbasesEst = 10
wbasesEst = create.bspline.basis(pdiaRng, nbasesEst)
nbasesSat = 20
wbasesSat = create.bspline.basis(pdiaRng, nbasesSat)

# Definición del parámetro de penalización para la función W
LfdobjEst      = 2          # penalización de la aceleración
LfdobjSat      = 2
lam            = 0.0001     # parámetro de suavización
```

```

cvecfEst      = matrix(0, nbasesEst, nindiv)
Wfd0Est       = fd(cvecfEst, wbasesEst)
precifdParEst = fdPar(Wfd0Est, LfdobjEst, lam)

cvecfSat      = matrix(0, nbasesSat, nindiv)
Wfd0Sat       = fd(cvecfSat, wbasesSat)
precifdParSat = fdPar(Wfd0Sat, LfdobjSat, lam)

for(i in 2:19){
  # Suavización monótona
  preciMonEst[[i]] = smooth.monotone(pdia, DatosObsEst[[i]],
    precifdParEst)
  WfdEst[[i]]      = preciMonEst[[i]]$Wfd
  betafEst[[i]]    = preciMonEst[[i]]$beta
  preciajusfdEst[[i]] = preciMonEst[[i]]$yhatfd
  datos_evalEst[[i]] = eval.monfd(pdia, WfdEst[[i]])
  DatosAjustEst[[i]] = betafEst[[i]][1,] +
    betafEst[[i]][2,] * t(datos_evalEst[[i]])
  DatosAjustEst[[i]] = t(DatosAjustEst[[i]])

  preciMonSat[[i]] = smooth.monotone(pdia, DatosObsSat[[i]], precifdParSat)
  WfdSat[[i]]      = preciMonSat[[i]]$Wfd
  betafSat[[i]]    = preciMonSat[[i]]$beta
  preciajusfdSat[[i]] = preciMonSat[[i]]$yhatfd
  datos_evalSat[[i]] = eval.monfd(pdia, WfdSat[[i]])
  DatosAjustSat[[i]] = betafSat[[i]][1,] +
    betafSat[[i]][2,] * t(datos_evalSat[[i]])
  DatosAjustSat[[i]] = t(DatosAjustSat[[i]])
}

```


E. Código para estimar un modelo de regresión funcional asumiendo independencia en los errores

```
Xt.fdAlt          <- list(19)
fRegressListAlt   <- list(19)
betaestlistAlt    <- list(19)

for(i in 1:9){
  Xt.fdAlt[[i]]    = list(const = rep(1,dim(WheatherStation_
llenada_final82est[[i]])[1]),
                           alt  = DataFrameLonLat82Estaciones$Alt,
                           alt2 = DataFrameLonLat82Estaciones$Alt^2,
                           preciMonSat.fd[[i]])
  betaafdPar       = fdPar(WfdEst[[i]], 2, 0)
  betalistAlt      = list(const = betaafdPar,
                           alt  = betaafdPar,
                           alt2 = betaafdPar,
                           sat  = betaafdPar)
  fRegressListAlt[[i]] = fRegress(preciMonEst.fd[[i]], Xt.fdAlt[[i]], betalistAlt)
}
```

F. Código para estimar las matrices de correlación espacial a partir de geoestadística funcional

```
M          <- list(19)
L2norm     <- list(19)
Trazavag   <- list(19)
Modelo     <- list(19)
#Matriz de distancias
M.dist <- max(as.matrix(dist(coord_utm,
  method="euclidean", diag=TRUE, upper=TRUE)))
for(i in 1:19){
  M[[i]]      = bsplinepen(bspl.basis[[i]] ,Lfdobj=0)
  L2norm[[i]] = l2.norm(n, datafd[[i]], M[[i]])
  Trazavag[[i]] = trace.variog(coord_utm, L2norm[[i]], bin=FALSE,
                                max.dist = M.dist, uvec="default",
                                breaks = "default", nugget.tolerance)}

#Traza-variograma clou
for(i in 1:19){
  pdf(paste("Graficos/TrazaVariograma_",1992+i, ".pdf", sep = ""))
  plot(Trazavag[[i]],
       xlab="Distancia",
       ylab="Traza-Variograma",
       col = "grey80",
       main = 1992+i,
       cex.lab = 1.5,
       cex.main = 2)
  dev.off()}
Modelo0 <- list()
for(i in 1:19){
  Modelo0[[i]] = fit.tracevariog(Trazavag[[i]],
    models = c("spherical","exponential","gaussian","matern"),
    sigma2.0 = ParametrosTrazaVariograma[i,2],
    phi.0 = ParametrosTrazaVariograma[i,3],
    fix.kappa = TRUE, max.dist.variogram = M.dist)}
```

G. Código para estimar un modelo de regresión lineal funcional con errores correlacionados

```
Y_est_fd <- list(19)
X_sat_fd <- list(19)
yhatfd_eval <- list(19)
yfd_eval <- list(19)
Diff_yfd_eval <- list(19)
Diff_yfd_eval <- list(19)
Diff_yhat_eval <- list(19)
Diff_yhat_eval <- list(19)

for(i in 1:19){

  X_sat <- DatosPentadalesSateliteAcumulado
  [DatosPentadalesSateliteAcumulado$YEAR == 1992 + i,13:(13+n-1)]
  X_sat <- t(X_sat)
  dim(X_sat)

  #Datos de estación
  Y_est <- DatosPentadalesEstacionAcumulados
  [DatosPentadalesEstacionAcumulados$YEAR == 1992 + i, 13:(13+n-1)]
  Y_est <- t(Y_est)
  dim(Y_est)

  tbasis = create.bspline.basis(rangval, nbasis = NumBasis, norder = 2)
  nbasis = NumBasis
  sbasis = tbasis
  intbasis = eval.penalty(tbasis, 0)
  Bases = create.bspline.basis(rangval, nbasis = NumBasis, norder = 2)

  Y_est_fd[[i]] <- Data2fd(Y_est, TiempoEval, BasesEjm)
  X_sat_fd[[i]] <- Data2fd(X_sat, TiempoEval, BasesEjm)
```

```

yfd = Y_est_fd[[i]]
xfd = X_sat_fd[[i]]
xphi = inprod(tbasis,xfd)
dim(xphi)
sxphi = apply(xphi,1,sum)
class(sxphi)
length(sxphi)

ypsi = inprod(tbasis,yfd)
class(ypsi)
dim(ypsi)
sypsi = apply(ypsi,1,sum)
class(sypsi)
length(sypsi)

dim(xphi%x%matrix(1,nbasis,1))
(matrix(1,nbasis,1)%x%ypsi)

yxphi = (xphi%x%matrix(1,nbasis,1))*(Omega[i]%x%ypsi)
dim(yxphi)
yxphi = apply(yxphi,1,sum) # dimensión k^2 x t
length(yxphi)

X = xphi%*%t(xphi) #dimensión k x k
dim(X)

insbasis = eval.penalty(sbasis,0)
inLsbasis = eval.penalty(sbasis,0)
inLtbasis = eval.penalty(tbasis,0)
Rs = intbasis%x%inLsbasis
Rt = inLtbasis%x%insbasis

lambda0 = 0
lambdas = 0
lambdat = 0

R0 = inLtbasis
penmat = rbind( cbind(lambda0*R0, matrix(0,nbasis,nbasis^2)),
                cbind(matrix(0,nbasis^2, nbasis),lambdas*Rs+lambdat*Rt) )

#class(penmat)
#dim(penmat)

Xmat = rbind( cbind( intbasis, t(sxphi)%x%Omega[i]%x%intbasis ),
              cbind( sxphi%x%intbasis, X%x%Omega[i]a%x%intbasis) )

```

```
#dim(Xmat)

ymat = c(sypsi,yxphi)
class(ymat)
length(ymat)
#betacoefs = solve(Xmat,ymat)
#library(MASS)
betacoefs = ginv(Xmat) %*% ymat
class(betacoefs)

beta0coefs = betacoefs[1:nbasis]
beta0fd = fd(beta0coefs,tbasis)


beta1coefs = betacoefs[nbasis+(1:(nbasis*nbasis))]
beta1Cmat = matrix(beta1coefs,nbasis,nbasis)
beta1fd = bffd(beta1Cmat,tbasis,sbasis)

beta1Cmat = matrix(betacoefs[nbasis+(1:(nbasis*nbasis))],nbasis,nbasis)

yhatfd = fd(betacoefs[1:nbasis]%x%matrix(1,1,N),tbasis) + fd(beta1Cmat%*%xphi,tbasis)
```