



# Análisis Geoestadístico Espacio Tiempo Basado en Distancias y Splines con Aplicaciones

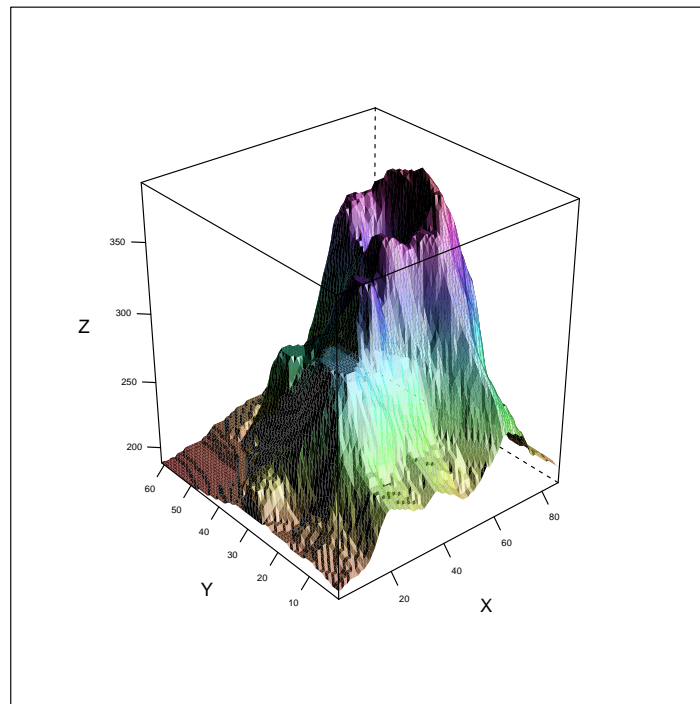
Carlos Eduardo Melo Martínez

**ADVERTIMENT.** La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX ([www.tdx.cat](http://www.tdx.cat)) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

**ADVERTENCIA.** La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR ([www.tdx.cat](http://www.tdx.cat)) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

**WARNING.** On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX ([www.tdx.cat](http://www.tdx.cat)) service has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized neither its spreading and availability from a site foreign to the TDX service. Introducing its content in a window or frame foreign to the TDX service is not authorized (framing). This rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

# Análisis Geoestadístico Espacio Tiempo Basado en Distancias y Splines con Aplicaciones



Carlos Eduardo Melo Martínez





# Análisis Geoestadístico Espacio Tiempo Basado en Distancias y Splines con Aplicaciones

MEMORIA PRESENTADA POR:

**Carlos Eduardo Melo Martínez**

PARA OPTAR AL TÍTULO DE DOCTOR POR LA UNIVERSIDAD DE BARCELONA

DOCTORANDO:

---

**Carlos Eduardo Melo Martínez**

DIRECTOR:

TUTOR:

---

**Dr. Jorge Mateu Mahiques**

Departamento de Matemáticas  
Universidad Jaume I de Castellón

---

**Dr. Antonio Monleón Getino**

Departamento de Estadística  
Facultad de Biología  
Universidad de Barcelona

Universidad de Barcelona  
Facultad de Biología  
Programa de Doctorado en Estadística  
Departamento de Estadística  
Barcelona, Mayo de 2012



## Agradecimientos

A mi director de tesis, el profesor Jorge Mateu, por ser un motivador permanente, por su constante interés, apoyo y por haber dedicado parte de su valioso tiempo guiándome en la realización de este trabajo. Mi admiración y sincera gratitud.

A mi hermana Sandra, compartimos como compañeros de estudios en el doctorado y fue más fácil la adaptación y estadía en Barcelona tan lejos de nuestra familia. Y a mi hermano Oscar con quien a lo largo de la vida hemos compartido y trabajado en infinidad de cosas, siendo así el doctorado una excusa mas para trabajar en equipo y compartir. Una enorme gratitud por su valiosa colaboración y apoyo en mis estudios.

A mi madre por enseñarme a escribir y por guiarme siempre hacia el buen camino. Todos sus sacrificios hicieron posible llegar hasta este punto. Gracias por ser la mejor mamá del mundo.

En la Universidad de Barcelona a mis profesores, en el año de docencia en especial a los profesores Carles Cuadras y Jordi Ocaña, recibí de ellos lo mejor. Y en el periodo de investigación, al profesor Antonio Monleón quien fue también tutor en esta tesis, por su colaboración y apoyo en mis estudios doctorales muchas gracias.

A la Universidad Distrital Francisco José de Caldas, por su valiosa ayuda a lo largo de mi vida como estudiante y como profesor, ya que allí fue donde me surgió el deseo de querer aprender más para ser un mejor profesional, docente y persona. Les quedaré por siempre agradecido por haberme dado esta invaluable oportunidad.

A la Universidad Nacional de Colombia y a la Universidad de Barcelona por haberme apoyado con la excepción del pago de matrícula en los periodos de docencia e investigación por medio de su convenio interinstitucional.

A los revisores y editores anónimos por sus valiosos comentarios sobre lo escrito, ya que este trabajo es también producto de las correcciones realizadas por ellos, en los diferentes artículos sometidos.





A mi familia, hermanos, sobrinas y en especial a mis padres Maria y Gustavo, quienes con su sacrificio y esfuerzo me iniciaron a muy temprana edad en la pasión por el conocimiento. Gracias por su apoyo incondicional durante toda mi vida, y por ser siempre motor, soporte y la razón de ser en todo lo que emprendo.





# Contenido

Lista de figuras	vi
Lista de tablas	x
Abreviaturas	1
Introducción	2
Objetivos	12
<b>1 Conceptos básicos del análisis de datos geoestadísticos</b>	<b>15</b>
1.1 Introducción . . . . .	15
1.2 Análisis geoestadístico tradicional . . . . .	16
1.2.1 Definiciones básicas . . . . .	17
1.2.2 El covariograma . . . . .	21
1.2.3 El variograma . . . . .	22
1.2.4 El correlograma . . . . .	23
1.2.5 Forma general de estas funciones . . . . .	23
1.3 Estimación del variograma y del covariograma . . . . .	25
1.3.1 Estimador clásico . . . . .	25
1.3.2 Estimador robusto . . . . .	26

1.4	Principales modelos de variogramas y covariogramas isotrópicos	26
1.5	Estimación de los parámetros del variograma . . . . .	27
1.5.1	Estimación por mínimos cuadrados . . . . .	30
1.5.2	Estimación mediante máxima verosimilitud . . . . .	31
1.6	Predicción espacial . . . . .	33
1.6.1	Kriging ordinario . . . . .	35
1.6.2	Kriging universal . . . . .	38
1.7	Diagnóstico mediante validación cruzada . . . . .	40
<b>2</b>	<b>Conceptos básicos del análisis espacio-temporal, de distancias y funciones de bases radial</b>	<b>43</b>
2.1	Introducción . . . . .	43
2.2	Geoestadística espacio-temporal . . . . .	46
2.3	Estimación del variograma y del covariograma . . . . .	51
2.4	Modelos de covarianza espacio-temporales . . . . .	52
2.4.1	Modelo métrico . . . . .	52
2.4.2	Modelo producto . . . . .	52
2.4.3	Modelo suma . . . . .	53
2.4.4	Modelo producto-suma . . . . .	53
2.4.5	Modelo Cressie-Huang . . . . .	54
2.5	Modelización de procesos espacio-temporales . . . . .	55
2.6	Predicción de procesos espacio-temporales . . . . .	56
2.6.1	Kriging ordinario espacio-temporal . . . . .	56
2.6.2	Kriging Universal espacio-temporal . . . . .	58
2.7	Regresión basada en distancias . . . . .	59
2.7.1	Distancia y similaridad . . . . .	59

2.7.2	Modelo de regresión basado en distancias . . . . .	63
2.8	Funciones de base radial . . . . .	64
2.8.1	Multicuadrática (MQ) . . . . .	65
2.8.2	Multicuadrática inversa (IM) . . . . .	66
2.8.3	Spline con tensión (ST) . . . . .	66
2.8.4	Spline capa delgada (TPS) . . . . .	66
2.8.5	Completamente regularizada spline (CRS) . . . . .	67
2.8.6	Gaussiana (GAU) . . . . .	67
<b>3</b>	<b>Modelo DB para la predicción espacial con tendencia</b>	<b>69</b>
3.1	Introducción . . . . .	69
3.2	Modelo basado en distancias con tendencia . . . . .	71
3.2.1	Kriging universal basado en distancias (DBUK) . . . . .	75
3.2.2	Medidas de evaluación . . . . .	80
3.3	Estudio de simulación y discusión . . . . .	81
3.3.1	Caso 1: Tendencia basada en variables mixtas sin omisión de variables explicativas . . . . .	82
3.3.2	Caso 2: Tendencia como en el caso 1, pero omitiendo una variable explicativa . . . . .	85
3.4	Aplicaciones . . . . .	87
3.4.1	Temperatura media diaria en Croacia . . . . .	89
3.4.2	Contenido de Calcio . . . . .	94
<b>4</b>	<b>Modelo DB para la predicción espacial utilizando RBF</b>	<b>99</b>
4.1	Introducción . . . . .	99
4.2	Modelo geoestadístico basado en distancias con funciones de base radial . . . . .	102

4.2.1	Predicción espacial basada en distancias con funciones de base radial . . . . .	105
4.3	Estudio de simulación y discusión . . . . .	110
4.4	Aplicación . . . . .	117
<b>5</b>	<b>Modelo DB para la predicción espacio-temporal usando funciones de base radial</b>	<b>121</b>
5.1	Introducción . . . . .	121
5.2	Modelo espacio-temporal basado en distancias con tendencia lineal local . . . . .	124
5.2.1	Tendencia basada en distancias con funciones de base radial . . . . .	126
5.2.2	Predicción espacio-temporal usando funciones de base radial basada en distancias . . . . .	131
5.3	Estudio de simulación y discusión . . . . .	135
5.4	Aplicación . . . . .	140
<b>6</b>	<b>Funciones geoestadísticas y funciones de base radial en el programa R: Paquete geospt</b>	<b>147</b>
6.1	Introducción . . . . .	147
6.2	Implementación de funciones geoestadísticas en R . . . . .	149
6.2.1	Pocket plot . . . . .	149
6.2.2	Variograma media recortada . . . . .	152
6.2.3	Resumen estadísticas de la validación cruzada . . . . .	154
6.2.4	Funciones rbf . . . . .	155
6.2.5	Mapa de predicciones . . . . .	156
<b>7</b>	<b>Conclusiones y futuras líneas de investigación</b>	<b>161</b>

7.1	Conclusiones . . . . .	161
7.2	Futuras líneas de investigación . . . . .	163
	<b>Referencias</b>	<b>165</b>
	<b>A Programación en R</b>	<b>179</b>
A.1	Funciones implementadas y utilizadas en el Capítulo 3 . . . . .	180
A.2	Funciones implementadas y utilizadas en los Capítulos 4 y 5 . . . . .	181
A.2.1	Predicción espacial basada en distancias con funciones de base radial . . . . .	181
A.2.2	Predicción espacio-temporal basada en distancias usando funciones de base radial . . . . .	185
A.3	Programación capítulo 3 . . . . .	187
A.4	Programación Capítulo 4 . . . . .	219
A.5	Programación Capítulo 5 . . . . .	233





# Lista de figuras

1.1	Forma general del variograma y covariograma de un proceso espacial homogéneo . . . . .	24
2.1	Relaciones entre los diferentes tipos de funciones de covarianza espacio-temporales . . . . .	49
3.1	Localización de los puntos de muestreo y regiones asociadas a la definición de la variable nominal . . . . .	83
3.2	RMSPE para los escenarios considerados en el Caso 1 . . . . .	86
3.3	$R^2$ para los escenarios considerados en el Caso 1 . . . . .	86
3.4	RMSPE para los escenarios considerados en el Caso 2 . . . . .	88
3.5	$R^2$ para los escenarios considerados en el Caso 2 . . . . .	88
3.6	Localizaciones de las estaciones meteorológicas en Croacia . . . . .	90
3.7	Mapas del variograma anisotrópico y modelos de variograma ajustados (azimut del semieje mayor es $135^\circ$ y azimut del semieje menor es de $45^\circ$ ) para los residuales de la temperatura media terrestre en los modelos clásico (dos paneles de izquierda) y DB (dos paneles de la derecha) . . . . .	91
3.8	Variograma experimental de media recortada para los residuos, ajustando un modelo de Matérn por WLS, OLS y REML . . . . .	92
3.9	Mapas de predicción de la temperatura media diaria terrestre en Croacia . . . . .	93

3.10	Mapas de predicción de las varianzas del error para la temperatura media diaria terrestre en Croacia . . . . .	93
3.11	Gráfica de círculo de contenido de calcio con las líneas que delimitan las sub-regiones (lugares de muestreo) . . . . .	95
3.12	Mapas de predicción del contenido de calcio en el suelo incluyendo sub-región . . . . .	97
3.13	Mapas de predicción de los errores estándar para el contenido de calcio en el suelo, incluyendo sub-región . . . . .	97
4.1	RMSPE para los escenarios espaciales simulados cuando $n_h = 8$	115
4.2	RMSPE para los escenarios espaciales simulados cuando $n_h = 32$	116
4.3	Localizaciones de muestreo y mapas de predicción bajo el método DB-SIRBF para el contenido de calcio en el suelo, incluyendo sub-región (tipo de suelo) . . . . .	119
5.1	Localización de los puntos de muestreo y regiones asociadas a la definición de la variable nominal . . . . .	137
5.2	RMSPE para los escenarios espacio-temporales simulados con 6 tiempos . .	141
5.3	RMSPE para los escenarios espacio-temporales simulados con 10 tiempos .	142
5.4	Localizaciones espaciales de las estaciones meteorológicas en Croacia y predictores estáticos topográficos: Modelo Digital de Elevación (DEM, en metros), la distancia topográfica ponderada desde la línea de costa (DSEA, en km) y el índice de humedad topográfica (TWI) . . . . .	144
5.5	Mapas de predicción de la temperatura promedio mensual de la tierra en Croacia bajo el método DBSTIRBF en enero, abril, julio y octubre (unidades de las coordenadas este y norte en 100.000 metros) . . . . .	146

6.1	Ubicación espacial de una muestra de cenizas de carbón (coal-ash), las unidades están en % en ubicaciones reorientadas (Cressie, 1993) . . . . .	150
6.2	POCKET-PLOT en dirección sur-norte: Claramente las filas 2, 6, y 8 son atípicas, esto sirve como verificación de que estas filas son potencialmente problemáticas. . . . .	151
6.3	Optimización de eta, en funciones de base radial . . . . .	156
6.4	Mapa de Croacia . . . . .	159



# Lista de tablas

1.1	Formas funcionales de algunos variogramas . . . . .	28
1.2	Formas funcionales de algunos covariogramas . . . . .	29
3.1	Escenarios simulados para los casos 1 y 2 . . . . .	84
3.2	Promedios de RMSPEs bajo los métodos UK y DBUK para los escenarios presentados en la Tabla 3.1 en el Caso 1 (sin omisión de variable) . . . . .	85
3.3	Promedios de RMSPEs bajo los métodos UK y DBUK para los escenarios presentados en la Tabla 3.1 en el Caso 2 (con una variable omitida) . . . . .	87
3.4	Comparación entre los métodos DB y clásico con los valores de los parámetros ajustados del variograma esférico utilizando máxima verosimilitud . . . . .	96
3.5	Comparación entre UK y DBUK para el contenido de calcio usando LOOCV . . . . .	98
4.1	Formas funcionales de algunas RBFs . . . . .	106
4.2	Escenarios considerados en los experimentos espaciales simulados	110
4.3	Escenarios espaciales simulados . . . . .	111
4.4	Promedios de RMSPEs bajo el método DBSIRBF en los escena- rios espaciales presentados en la Tabla 4.3 (casos nivel de ruido y densidad de diseño) . . . . .	112

4.5	Promedios de RMSPEs bajo el método DBSIRBF en los escenarios espaciales presentados en la Tabla 4.3 (casos varianza espacial y función de varianza) . . . . .	113
4.6	Comparación de algunos métodos DBSIRBFs para el contenido de calcio utilizando LOOCV . . . . .	117
5.1	Formas funcionales de algunas RBFs espacio-temporales . . . . .	131
5.2	Escenarios considerados en los experimentos simulados espacio-temporales . . . . .	136
5.3	Escenarios espacio-temporales simulados . . . . .	137
5.4	Promedios de RMSPEs bajo el método DBSTIRBF en los escenarios espacio-temporales presentados en la Tabla 5.3 (casos nivel de ruido y densidad de diseño) . . . . .	138
5.5	Promedios de RMSPEs bajo el método de DBSTIRBF de los escenarios espacio-temporales presentados en la Tabla 5.3 (casos variación espacio-temporal y función de varianza) . . . . .	139
5.6	Comparación de algunos métodos DBSTIRBF para las temperaturas promedios mensuales de 2008 en Croacia con LOOCV . . . . .	145
6.1	Algunas funciones del paquete <b>geosp</b> . . . . .	157

# Abreviaturas

<b>ASE:</b>	Average Standard Error (Error estándar promedio)
<b>CRS:</b>	Completely Regularized Spline (Spline completamente regularizada)
<b>DBSIRBF:</b>	Distance-Based Spatial Interpolation with Radial Basis Function (Interpolación espacial basada en distancias con funciones de base radial)
<b>DBSTIRBF:</b>	Distance-Based Spatio-Temporal Interpolation with Radial Basis Function (Interpolación espacio-temporal basada en distancias con funciones de base radial)
<b>DBUK:</b>	Distance-Based Universal Kriging (Kriging universal basado en distancias)
<b>DEM:</b>	Digital Elevation Models (Modelos digitales de elevación)
<b>DSEA:</b>	Distance (km) from the coast line (Distancia topográfica en kilómetros ponderada desde la línea a la costa)
<b>EXP:</b>	Exponential (Exponencial)
<b>GAU:</b>	Gaussian (Gaussiana)
<b>GLS:</b>	Generalized Least Squares (Mínimos cuadrados generalizados)
<b>IMQ:</b>	Inverse Multiquadratic (Multicuadrática inversa)
<b>LOOCV:</b>	Leave-One-Out Cross Validation (Validación cruzada dejando uno fuera)
<b>MPE:</b>	Mean Prediction Errors (Media de los errores de predicción)
<b>MSPE:</b>	Mean Standardized Prediction Errors (Media estandarizada de los errores de predicción)
<b>MQ:</b>	Multiquadratic (Multicuadrática)
<b>OLS:</b>	Ordinary Least Squares (Mínimos cuadrados ordinarios)
<b>ST:</b>	Spline with Tension (Spline con tensión)
<b>TPS:</b>	Thin Plate Spline (Spline capa delgada)
<b>TWI:</b>	Topographic Wetness Index (Índice de humedad topográfica)
<b>RMSPE:</b>	Root Mean Square Prediction Errors (Raíz media del cuadrado de los errores de predicción)
<b>RMSSPE:</b>	Root Mean Square Standardized Prediction Errors (Raíz media estandarizada del cuadrado del error de predicción)
<b>RBF:</b>	Radial Basis Function (Función base radial)
<b>UK:</b>	Universal Kriging (Kriging universal)
<b>WLS:</b>	Weighted Least Squares (Mínimos cuadrados ponderados)





# Introducción

La mayoría de los fenómenos naturales que se estudian se pueden describir mediante variables regionalizadas, tanto en el espacio como en el tiempo. Por ejemplo, considerando una superficie topográfica o una contaminación de las aguas subterráneas, se puede observar una alta variabilidad en distancias pequeñas. La variabilidad es el resultado de procesos naturales, por lo tanto es determinista. Pero como la mayoría de estos procesos son muy sensibles y las condiciones en las que tienen lugar no se conocen, basándose en leyes físicas y químicas no es posible describirlos por completo (Bárdossy 2001).

La teoría de variables regionalizadas, que es el tema de la presente investigación, se remonta a los años cincuenta, cuando en Sudáfrica D. Krige y sus colegas comenzaron a aplicar técnicas estadísticas para la estimación de reservas de mineral. En los años sesenta el matemático francés G. Matheron sentó las bases teóricas de los métodos anteriores. La geoestadística primero fue utilizada por la industria minera, en la cual, dado que los costes de las perforaciones eran altos, el análisis de los datos fue de suma importancia. El modelamiento de variables medidas en diferentes sitios de una región con continuidad espacial y que presentan alguna estructura de correlación espacial, ha sido desarrollada desde los años sesenta (Cressie 1993), con el desarrollo de los análisis geoestadísticos (Matheron 1962), incrementándose su uso en diferentes disciplinas científicas como la minería (Journel & Huijbregts 1978), geología (Samper & Carrera 1993), ecología (Robertson 1987), ciencias ambientales (Cressie & Majure 1995, Diggle et al. 1995, Paez & De Oliveira 2005), salud pública (Haining 2004), y climatología (Perčec Tadić 2010, Hengl et al. 2012, Yavuz & Erdoğan 2012). Los análisis geoestadísticos convencionales

contemplan una serie de pasos (Isaaks & Srisvastava 1989), que comienzan con el análisis estructural, el cual se realiza en el análisis del variograma (Samper & Carrera 1993), obteniendo en lo posible un modelo de variograma teórico (esférico, exponencial, gaussiano, circular o de Matern, entre otros que están disponibles), el cual es usado en la interpolación de la variable en los sitios no muestreados, para producir mapas que finalmente suelen ser empleados para análisis y toma de decisiones.

Muchos métodos tales como: suavizamiento kernel (Wand & Jones 1995), polinomios locales (Cleveland 1979, Fan & Gijbels 1996), wavelet (Donoho & Johnstone 1994), regresión splines (Wand 2000), y suavizamientos splines (Craven & Wahba 1979, Chen 2007), han sido propuestos para estimar y seleccionar curvas en modelos de regresión. Los resultados obtenidos por estos investigadores son útiles para la geoestadística ya que ofrecen la posibilidad de ser considerados en el modelamiento de las superficies a interpolar e inclusive en los ajustes de la correlación espacial, específicamente en el modelamiento del variograma.

El uso de modelos de correlación entre observaciones estimulan la necesidad de modelos de análisis geoestadístico. La información georeferenciada se recoge en muchas aplicaciones y no utilizar esta información puede obstruir las características importantes del mecanismo de generación de datos. En este sentido, técnicas como kriging simple y ordinario consideran una media constante de la variable regionalizada que es modelada, conocida y desconocida, respectivamente. Además, asumen con respecto a esta media unas condiciones de estacionariedad o cuasi-estacionariedad y la existencia de una varianza finita, estos mismos supuestos se requieren en la mayoría de los métodos kriging. Por otro lado, el método denominado kriging universal no asume una media constante y es con frecuencia usado en los procesos no estacionarios, por lo cual también recibe el nombre de kriging no estacionario (Wackernagel 2003). Por ejemplo, en el estudio de variables ambientales con frecuencia se utiliza los interpoladores geoestadísticos como en Le & Zidek (2006) y van de Kastelee et al. (2009), debido a que estas variables suelen ser no estacionarias, tal como lo indica el estudio de Braud (1990), en el cual se resume mensualmente la no

estacionariedad de la temperatura promedio de la superficie.

Si la deriva no es constante y es empleado algún método como kriging simple u ordinario, el estimador de la variable regionalizada no será insesgado y en extrapolaciones o en los límites de una región el estimador tenderá a subestimar o sobrestimar el verdadero valor de la variable estudiada. Por lo cual, es recomendable remover la deriva y obtener los residuos de la diferencia entre los valores muestreados y estimados por medio de la función asociada a la deriva, ya que así se garantiza que los residuales sean estacionarios o al menos cumplan la propiedad de ser estacionarios intrínsecamente. Con los residuos estacionarios intrínsecos se construye un modelo de semivarianza o covarianza, el cual es incorporado en el krigeado universal junto con la función asociada a la deriva para la generación de la estimación de la variable en el sitio de interés. Otra forma de emplear estos residuos es en el krigeado simple para la estimación del componente residual en el sitio de interés; al residual estimado se le adiciona el valor estimado de la deriva para obtener el valor final, esto último se suele denominar regresión kriging (Hengl 2009).

Por otro lado, muchos métodos de estadística y análisis de datos utilizan el concepto geométrico de distancia entre individuos o poblaciones. Las distancias, aparecen en muchos aspectos de la estadística: contraste de hipótesis, estimación, regresión, análisis discriminante, etc. (Arenas & Cuadras 2002). Cuadras & Arenas (1990) proponen el método de regresión múltiple basado en el análisis de distancias utilizando diferentes métricas para el trabajo con variables explicativas continuas y categóricas. Posteriormente, Cuadras et al. (1996) presentaron algunos resultados adicionales del modelo basado en distancias (DB) para la predicción de variables mixtas (continuas y categóricas) y exploran el problema de información faltante dando una solución utilizando DB. Uno de los trabajos más recientes es el de Esteve et al. (2009), quienes proponen un método donde incluyen términos polinomiales y de interacción en la regresión basada en distancias, bajo las propiedades de un producto de matrices semi-Hadamard o Khatri-Rao. Adicionalmente, en Cuadras (2009) se estudia la regresión multivariada basada en distancias Cuadras (2009b). En términos generales muchos métodos en la estadística se basan en el cálculo de

distancias geométricas, de por sí los métodos geoestadísticos se construyen con distancias, en particular distancias euclídeas espaciales, de aquí el interés de considerar también los métodos basados en distancias ya que tienen elementos en común, como lo es el cálculo de las distancias entre las observaciones, esto anidado a la información que aporta el variograma serán determinantes en la generación de pronósticos y permitirá mejorar el poder predictivo de los métodos kriging tradicionales.

Dichos resultados mencionados motivan la idea de trabajar en el modelamiento de la tendencia, a partir de los métodos desarrollados por Cuadras (1989), Cuadras & Arenas (1990) y Cuadras et al. (1996), ya que es una excelente alternativa para ayudar a mejorar las predicciones en el caso geoestadístico cuando se tienen variables explicativas asociadas a las coordenadas de los puntos (puestas en un polinómico de orden 1, 2 o 3), y covariables regionalizadas continuas, categóricas y binarias. La selección de variables explicativas se hace a partir de técnicas muy populares en el análisis de regresión tradicional (selección: forward F, backward B, y step-wise "B-F" ó "F-B"); recientemente se han propuesto otras técnicas (George & McCulloch 1993, Breiman 1995, Tibshirani 1996, Efron et al. 2004, Joseph et al. 2008). Sin embargo, aquí se presentan algunas alternativas a partir de la propuesta de Cuadras et al. (1996) para seleccionar las componentes principales o nuevas variables explicativas obtenidas a partir de la descomposición espectral de la matriz de covariables. Dado que los métodos basados en distancias en diferentes trabajos han mostrado ganancias importantes en los pronósticos con respecto a los métodos tradicionales, en esta tesis se elabora un método alternativo para el modelamiento de la tendencia en un modelo geoestadístico, ya que también el método basado en distancias es robusto ante los errores de especificación en la correlación de los parámetros.

Por lo tanto, en esta tesis se propone un método alterno de interpolación espacial con variables explicativas mixtas utilizando distancias entre individuos, tales como la distancia de Gower (Gower 1968); aunque, algunas otras distancias euclidianas se pueden usar. El método basado en distancias (Distance-Based, DB) se utiliza en los modelos geoestadísticos no sólo en la etapa de

estimación de la tendencia para su remoción, sino también en la etapa de estimación de la correlación espacial, cuando las variables explicativas son mixtas. En el caso de la regresión geoestadística, el método DB espacial propuesto se basa sobre los métodos desarrollados por (Cuadras & Arenas 1990) y (Cuadras et al. 1996). Esta estrategia es una excelente alternativa, ya que aprovecha al máximo la información obtenida debido a la relación entre las observaciones, la cual puede ser establecida a través del uso de la descomposición espectral, utilizando cualquier distancia euclídea. En consecuencia, este enfoque permite mejorar las predicciones ya que se puede elegir una mayor cantidad de coordenadas principales que de variables explicativas asociadas con la variable respuesta de interés en las localizaciones muestreadas.

Por otra parte, las funciones de base radial (RBF) tales como la multicuadrática (multiquadratic, MQ) o completamente regularizada spline (completely regularized spline, CRS) son útiles en la construcción de modelos digitales de elevación (Digital Elevation Models, DEM), como se muestra en (Mitášová & Hofierka 1993). Una variación de la función MQ se llama la inversa multicuadrática (inverse multiquadratic, IMQ), introducida por (Hardy & Gopfert 1975). En Späh (1969) se describe un método que permite evitar puntos de inflexión y contiene splines cúbicos como un caso especial, utilizando interpolación spline cúbica y exponencial (EXP). Más tarde, el spline capa delgada (thin plate spline, TPS) se introdujo en el diseño geométrico por (Duchon 1976), y la aproximación de Gauss (GAU) utilizada por (Schagen 1979) es una variante popular del TPS. Por último, (Mitáš & Mitášová 1988, Mitášová & Hofierka 1993, Mitášová & Mitáš 1993) desarrollan la formulación de la spline con tensión (spline with tensión, ST), e implementan un algoritmo de segmentación con un tamaño flexible de la superposición del vecindario.

En este punto, adicionalmente en esta tesis se propone el método de funciones de base radial espacial basado en distancias (distance-based spatial radial basis functions, DBSRBFs), el cual se aplica en el modelo geoestadístico para predecir la tendencia y estimar la estructura de covarianza cuando las variables explicativas son mixtas utilizando la distancia de Gower (1971).

En lo que respecta a la modelización espacio-temporal, se ha manifestado en los últimos años una gran demanda de modelos suficientemente realistas que describan la evolución de procesos medio ambientales en el espacio y tiempo, modelos que puedan capturar simultáneamente el comportamiento de ambas componentes. Si éstas fueran analizadas de forma separada, se corre el riesgo de obviar información importante, por ejemplo, de acuerdo a Gneiting (2003), para una adecuada predicción determinista hace falta conocer perfectamente el estado presente de la atmósfera y las leyes físicas involucradas en los procesos atmosféricos. Pero la realidad es que la incertidumbre juega un papel importante: tramas incompletas de observaciones, errores en las medidas y localizaciones, conocimiento incompleto de las leyes físicas, etc. Por tanto, está claro que probabilistas y estadísticos pueden jugar un papel importante en este contexto, puesto que los procedimientos estadísticos representan una adecuada alternativa para tratar convenientemente la incertidumbre.

La geoestadística espacio-temporal hace referencia al conjunto de técnicas geoestadísticas que analizan, describen y modelizan procesos espaciales con evolución temporal. Como es sabido, los procedimientos de interpolación basados en kriging dependen de la elección de la autocovarianza asociada al campo espacio-temporal. Por tanto, la perspectiva geoestadística se basa en la obtención de covarianzas espacio-temporales permisibles que analicen de forma adecuada las interacciones espacio-tiempo. En otras palabras, se necesitan modelos de covarianza espacio-temporal no separables asociados a campos aleatorios estacionarios y no estacionarios. Éste ha sido y es actualmente uno de los retos más importantes para la comunidad estadística que trabaja en este campo científico.

En cuanto a estas funciones de covarianza se vienen desarrollando extensiones al caso espacio-temporal, como el trabajo desarrollado por Martínez (2008), en donde se realizan simulaciones para comparar la habilidad predictiva de 4 clases distintas de modelos espacios temporales (dinámico de Huang & Cressie (1996), no separable tanto de Cressie & Huang (1999) como de Gneiting (2002), y modelo suma producto De Cesare et al. (2001*a*)), utilizando los valores AIC y BIC como criterios de selección. Además allí se propone un mo-

delo de función de covarianza denominado suma de productos generalizado, el cual genera modelos espacio-temporales a partir de combinaciones lineales de modelos espaciales y temporales. Esto motiva el interés en el trabajo de una propuesta que permita anidar modelos de covarianza espacio-temporales, que a la vez permita solucionar problemas de modelamiento de variables como la precipitación, la contaminación, entre muchas otras que se encuentran en disciplinas generalmente asociadas al medio ambiente.

En Lloyd (2010) se muestra como las funciones splines caso capa delgada son muy relevantes en la predicción espacial, a su vez se asocia al método kriging universal; los resultados allí son utilizados adicionándole la tendencia desde el enfoque de los métodos basados en distancias. Además, los splines son utilizados en el modelamiento de la función de covarianza y de semivarianza, tal como se indica en el trabajo desarrollado por García-Soidán et al. (2012). En este trabajo aplican la técnica de las series de Fourier para estimar la función de covarianza de un proceso aleatorio estacionario de segundo orden, se menciona el trabajo en la estimación de la covarianza no paramétrica empírica desarrollada por Cressie (1993) y el estimador de tipo kernel propuesto en Hall & Patil (1994).

En esta tesis, se considera también el problema de elegir un modelo basado en distancias que incorpora información, que se cree influencia la variable respuesta. Especialmente, en el análisis de datos espacio-temporal, a menudo trata con variables explicativas mixtas asociadas con la variable respuesta. Por lo tanto, se presenta un enfoque unificado que utiliza las RBFs en contextos espacio-temporales donde las variables explicativas son de naturaleza mixta, y por consiguiente, la distancia de Gower (1968) es empleada. Al igual que en las anteriores propuestas de esta tesis, el método de interpolación espacio-temporal basado en distancias usando funciones (distance-based spatio-temporal interpolation using radial basis functions, DBSTIRBFs) se aplica a los modelos métricos espacio-temporales para predecir la tendencia y estimar la estructura de covarianza cuando las variables explicativas son mixtas.

En todos los métodos de predicción presentados en esta tesis, las coordenadas principales obtenidas mediante el método de distancias se obtienen a par-



tir de las covariables asociadas con la variable de respuesta, y las coordenadas espaciales o espacio-temporales. La selección de las coordenadas principales se lleva a cabo usando los valores de la prueba- $t$  significativos estadísticamente y una caída significativa en la falta de predictibilidad, es decir, las coordenadas principales que están más asociadas con la variable respuesta. Además, para evaluar la exactitud del interpolador del método propuesto, se realizaron simulaciones incondicionales para cinco funciones de base radial en diferentes escenarios prácticos, y los resultados muestran que las RBFs utilizando el método DB tienen ventajas como la de trabajar con variables mixtas en el tendencia y el no requerir de la estimación de un variograma espacio-temporal, que normalmente requieren mucho tiempo computacional.

Este trabajo lo hemos dividido en seis capítulos de la siguiente forma:

**Capítulo 1.** Presenta brevemente conceptos básicos geoestadísticos para el análisis de datos espaciales, en cuanto a la dependencia espacial asociada al variograma o covariograma, y la predicción espacial generada con los métodos kriging, así como también la valoración de dichas predicciones por medio de la validación cruzada.

**Capítulo 2.** Describe los principales elementos y métodos utilizados en el análisis espacio-temporal, definiendo conceptos involucrados en la estimación de la estructura del variograma y covariograma espacio-temporal, presenta los modelos de covarianza espacio-temporales y muestra su uso en la predicción de procesos espacio-temporales. El capítulo termina con una corta explicación de la regresión basada en distancias y de las funciones de base radial en general.

**Capítulo 3.** En los dos primeros capítulos se introdujeron los conceptos que son la base para los métodos propuestos en esta investigación. Este capítulo contiene el modelo basado en distancias para la predicción espacial con tendencia, generado a partir de kriging universal, se describe el método propuesto, se realiza un estudio exhaustivo de simulación para comparar el método propuesto con respecto al tradicional kriging, y se desarrollan dos aplicaciones que ilustran la metodología propuesta.

**Capítulo 4.** Contiene el modelo propuesto basado en distancias para la predicción espacial usando funciones de base radial. Se desarrolla la propuesta

---

metodológica introduciendo la tendencia lineal local basada en distancias junto con las funciones de base radial, haciendo una aproximación a partir de la interpolación spline al método kriging para la predicción espacial. Un estudio de simulación intensivo y extensivo basado en algunos modelos splines se realiza. El capítulo termina con una aplicación que ilustra la metodología propuesta.

**Capítulo 5.** Se describe el modelo propuesto basado en distancias para la predicción espacio-temporal usando funciones de base radial. Al igual que en capítulo anterior se hace aproximación a partir de la interpolación spline al método kriging para la predicción espacio-temporal. Contiene un estudio de simulación que considera algunos escenarios bajo diferentes parámetros y funciones de base radial. Y finaliza, con una aplicación de la metodología propuesta para la temperatura media mensual terrestre en Croacia.

**Capítulo 6** En este capítulo, describimos parte del paquete **geospt** implementado en el programa (R Development Core Team (2012)), el cual puede ser usado para; optimización, predicción y validación cruzada en las funciones de base radial espaciales, generación de resumen de estadísticos a partir de validación cruzada para funciones de base radial y métodos kriging, y construcción del pocket plot para datos grillados.

**Capítulo 7** En este último capítulo se resumen las aportaciones del presente trabajo y enuncian algunas futuras líneas de investigación.



# Objetivos

## Objetivos Generales

Proponer innovaciones en la predicción espacio y espacio-temporal a partir de métodos geoestadísticos kriging y de funciones de base radial considerando métodos basados en distancias.

## Objetivos Específicos

- Por medio de las distancias entre las variables explicativas, incorporadas específicamente en la regresión basada en distancias, proponer una modificación al método kriging universal y a la interpolación con splines espaciales y espacio-temporales usando las funciones de base radial.
- Aplicar los métodos propuestos a casos reales de ciencias de la tierra, tales como; el modelado de la temperatura media terrestre espacio-temporal y el modelado de la variable edáfica ca20 (porcentaje de contenido de calcio en el suelo a una profundidad de 0 a 20 cm).
- Validar mediante simulaciones bajo ciertos escenarios, tanto para las variables (explicativas y explicada) como para los parámetros asociados a los métodos diseñados, los métodos propuestos con el fin de evaluar su funcionamiento.
- Crear una librería en el programa estadístico R con herramientas de análisis geoestadístico, fundamentalmente incorporando funciones de

base radial espaciales sin tendencia, y generar funciones asociadas a los métodos propuestos en esta investigación.

# Capítulo 1

## Conceptos básicos del análisis de datos geoestadísticos

### 1.1 Introducción

La estadística espacial ha venido tomando fuerza en diferentes áreas del conocimiento en los últimos años, y en éste caso la presente propuesta se genera a partir de una de las ramas que allí se desarrollan como lo es la geoestadística. Esta ciencia reúne métodos que permiten modelar las estructuras de relación espacial en funciones denominadas variogramas o covariogramas, y posteriormente, con la información que se extrae de tales funciones se realizan interpolaciones espaciales en los métodos denominados kriging.

El modelamiento de variables medidas en diferentes sitios de una región con continuidad espacial y que presentan alguna estructura de correlación espacial, ha sido desarrollada desde los años sesenta (Cressie 1993), con el desarrollo de los análisis geoestadísticos (Matheron 1962), incrementándose su uso en diferentes disciplinas científicas como la minería (Journel & Huijbregts 1978), geología (Samper & Carrera 1993), ecología (Robertson 1987), ciencias ambientales (Cressie & Majure 1995, Diggle et al. 1995, Paez & De Oliveira 2005), salud pública (Haining 2004), y climatología (Perčec Tadić 2010, Hengl et al. 2012, Yavuz & Erdoğan 2012). Los análisis geoestadísticos convencionales

contemplan una serie de pasos (Isaaks & Srisvastava 1989), que van desde el análisis estructural, el cual se realiza en el análisis del variograma (Samper & Carrera 1993), obteniendo en lo posible un modelo de variograma teórico (esférico, exponencial, gaussiano, circular o de Matérn, entre otros que están disponibles), el cual es usado en la interpolación de la variable en los sitios no muestreados.

En este capítulo se presentan algunos conceptos básicos en su mayoría tomados de Cressie (1993) y Martínez (2008), sobre análisis geoestadístico que son útiles para el desarrollo de las metodologías propuestas en cada uno de los siguientes capítulos. Por lo tanto, se describe brevemente los principales conceptos y resultados utilizados en el análisis de datos espaciales, muchos de los cuales se generalizarán con el fin de poder aplicarlos a procesos espacio-temporales. En estos últimos se definen los principales conceptos involucrados en su estudio y sus propiedades, así como los diferentes procedimientos de ajuste y predicción.

La Sección 1.2 introduce los conceptos más relevantes que se utilizaran en el análisis espacial, como son la estacionariedad, la isotropía, el covariograma o el variograma. En la Sección 1.3 se muestran los principales estimadores del variograma y covariograma. En la Sección 1.4 se presentan los principales modelos de variogramas y covariogramas isotrópicos. En las Secciones 1.5, 1.6 y 1.7 se profundiza en el ajuste de los modelos anteriores mediante las principales técnicas de estimación, se introducen las herramientas de predicción espacial más relevantes, y se muestra algunas herramientas de diagnóstico de los modelos ajustados, respectivamente.

## 1.2 Análisis geoestadístico tradicional

Cressie (1993) muestra una formulación general que permite la modelización de todas estas posibilidades. Sea  $\mathbf{s}$  una localización cualquiera del espacio Euclídeo  $d$ -dimensional  $\mathbb{R}^d$  (en general  $d = 2$ , aunque no necesariamente), suponga que se está interesado en analizar un determinado fenómeno de interés que toma un valor aleatorio  $Z(\mathbf{s})$  en cada localización  $\mathbf{s}$ . Si ahora se permite

que  $\mathbf{s}$  varíe sobre un determinado conjunto  $D \subseteq \mathbb{R}^d$ , se tendrá el proceso aleatorio  $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ , que es el objeto de estudio de la estadística espacial. La geoestadística estudiará aquellos fenómenos en los que el índice espacial  $\mathbf{s}$  varíe de forma continua sobre toda la región de estudio  $D$ . En este sentido, en esta tesis se supondrá que  $D$  es una determinada región fija y continua de estudio y que el índice espacial  $\mathbf{s}$  varía de forma continua en  $D$ , es decir, existe un número infinito de posibles localizaciones en las que se observa el proceso. El proceso objeto de estudio  $Z(\mathbf{s})$  podría representar, por ejemplo, la temperatura media diaria observada en una determinada localización  $\mathbf{s}$ .

### 1.2.1 Definiciones básicas

A lo largo de todo este capítulo se supondrá que, para cada localización  $\mathbf{s} \in D$ , existe la media y la varianza del proceso que se denotarán por

$$\mu(\mathbf{s}) = E(Z(\mathbf{s})) < \infty \quad \text{Var}(Z(\mathbf{s})) < \infty$$

**Definición 1.1.** *Se dice que el proceso  $Z(\mathbf{s})$  es Gaussiano si para cualquier conjunto de localizaciones  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \in D$ , el vector aleatorio  $\mathbf{Z}_s = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$  sigue una distribución normal multivariante.*

**Definición 1.2.** *Sea  $Z(\mathbf{s})$  un proceso estocástico de segundo orden. Se define su función de covarianza como*

$$C(\mathbf{s}_i, \mathbf{s}_j) = C(Z(\mathbf{s}_i), Z(\mathbf{s}_j)), \quad \forall \mathbf{s}_i, \mathbf{s}_j \in D$$

Generalmente en la práctica sólo se dispone de un conjunto  $\{z(\mathbf{s}_1), \dots, z(\mathbf{s}_n)\}$  de observaciones del proceso aleatorio  $\{Z(\mathbf{s}), \mathbf{s} \in D\}$  obtenidas sobre un conjunto de localizaciones  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ , que pueden distribuirse de forma regular sobre una rejilla o de forma irregular sobre la región de estudio  $D \subseteq \mathbb{R}^d$ . Por lo tanto, sólo se dispone de una única realización incompleta del proceso aleatorio que se quiere analizar, por lo que sería necesario asumir algún tipo de hipótesis simplificadora de la naturaleza del proceso que asegure cierta regularidad en los datos y permita hacer estimaciones e inferencias del modelo a partir de los datos observados. Esta condición es la de



estacionariedad, que permite que el proceso se repita a si mismo en el espacio, proporcionando la replicación necesaria para la estimación e inferencia del modelo. A continuación se verá los principales tipos de estacionariedad que generalmente se asume en los procesos a analizar.

**Definición 1.3.** *Se dice que el proceso  $Z(\mathbf{s})$  es estrictamente estacionario (o estacionario en sentido fuerte) si, para cualquier conjunto de localizaciones  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \in D$ , la función de distribución conjunta de las variables aleatorias  $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}$  permanece invariable ante una traslación. Sea  $F_{\mathbf{s}_1, \dots, \mathbf{s}_n}(z_1, \dots, z_n) = P(Z(\mathbf{s}_1) \leq z_1, \dots, Z(\mathbf{s}_n) \leq z_n)$  la función de distribución conjunta, entonces se cumple que*

$$F_{\mathbf{s}_1, \dots, \mathbf{s}_n}(z_1, \dots, z_n) = F_{\mathbf{s}_1 + \mathbf{h}, \dots, \mathbf{s}_n + \mathbf{h}}(z_1, \dots, z_n), \quad \forall \mathbf{h} \in \mathbb{R}^d$$

Esta condición es demasiado restrictiva para la mayoría de los fenómenos observados en la naturaleza, por lo que se necesita algún tipo de relajación de la misma, como la estacionariedad de segundo orden o la estacionariedad intrínseca.

**Definición 1.4.** *Se dice que un proceso espacial  $Z(\mathbf{s})$  es estacionario de segundo orden (o estacionario en sentido débil o simplemente estacionario) si*

1. *La función media existe y no depende de la localización, esto es,  $\mu(\mathbf{s}_i) = \mu, \forall \mathbf{s}_i \in D$ .*
2. *La función de covarianza existe y sólo depende de la distancia entre las localizaciones involucradas, esto es,  $C(\mathbf{s}_i, \mathbf{s}_j) = C(\mathbf{h}), \forall \mathbf{s}_i, \mathbf{s}_j \in D$ , siendo  $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$  el vector distancia entre dichas localizaciones. La función  $C(\cdot)$  recibe el nombre de covariograma (o autocovarianza).*

De la definición se deduce que si un proceso de segundo orden es estrictamente estacionario, entonces es estacionario de segundo orden. El recíproco es falso en general, aunque se cumple para los procesos gaussianos, que quedan completamente caracterizados por su media y su covariograma.

La estacionariedad de segundo orden implica que la varianza del proceso no depende de la localización, es decir, que

$$\text{Var}(Z(\mathbf{s})) = C(\mathbf{0}) = \sigma^2, \quad \forall \mathbf{s} \in D,$$

donde  $C(\mathbf{0})$  recibe el nombre de varianza a priori del proceso.

**Definición 1.5.** *Se dice que el proceso  $Z(\mathbf{s})$  es intrínsecamente estacionario si*

- i. *La función media existe y no depende de la localización, esto es,  $\mu(\mathbf{s}_i) = \mu, \forall \mathbf{s}_i \in D$ .*
- ii. *La varianza de la diferencia de dos variables aleatorias para dos localizaciones cualesquiera depende únicamente de la distancia entre las localizaciones involucradas, esto es,  $\text{Var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) = 2\gamma(\mathbf{h}), \forall \mathbf{s}_i, \mathbf{s}_j \in D$ , con  $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ . La función  $2\gamma(\cdot)$  recibe el nombre de variograma, mientras que  $\gamma(\cdot)$  se conoce como semivariograma.*

Esta condición es la menos restrictiva de las últimas tres definiciones dadas, ya que dado un proceso estacionario  $Z(\mathbf{s})$  de segundo orden con covariograma  $C(\cdot)$ , entonces

$$\begin{aligned} \text{Var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) &= \text{Var}(Z(\mathbf{s}_i)) + \text{Var}(Z(\mathbf{s}_j)) - 2\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \\ &= 2C(\mathbf{0}) - 2C(\mathbf{s}_i - \mathbf{s}_j) \end{aligned}$$

por lo que el proceso  $Z(\mathbf{s})$  es intrínsecamente estacionario con variograma

$$2\gamma(\mathbf{h}) = 2C(\mathbf{0}) - 2C(\mathbf{h}) \quad (1.1)$$

Para que un proceso intrínsecamente estacionario lo sea también de segundo orden, deberá tener un semivariograma acotado, esto es, con  $\lim_{\mathbf{h} \rightarrow \infty} \gamma(\mathbf{h}) = M < +\infty$ , en cuyo caso su covariograma existe y es igual a  $C(\mathbf{h}) = M - \gamma(\mathbf{h})$ .

**Definición 1.6.** *Se dice que el proceso  $Z(\mathbf{s})$  es isotrópico si la dependencia espacial del proceso entre dos localizaciones cualesquiera depende únicamente de la distancia existente entre ellas y no de su localización. En caso contrario se dice que el proceso es anisotrópico.*

**Definición 1.7.** *Se dice que el proceso  $Z(\mathbf{s})$  es homogéneo si es intrínsecamente estacionario e isotrópico.*

Si  $Z(\mathbf{s})$  es un proceso homogéneo, entonces su semivariograma es una función que, para cada par de localizaciones, depende únicamente de la longitud del vector distancia entre ellas, esto es,  $\gamma(\mathbf{h}) = \gamma(h), \forall \mathbf{h} \in \mathbb{R}^d$ , siendo  $h \equiv \|\mathbf{h}\|$ . En cambio si un proceso intrínsecamente estacionario  $Z(\cdot)$  es anisotrópico, la dependencia entre  $Z(\mathbf{s})$  y  $Z(\mathbf{s} + \mathbf{h})$  será función tanto de la magnitud como de la dirección de  $\mathbf{h}$ , por lo que el variograma no será únicamente una función de la distancia entre dos localizaciones espaciales. Las anisotropías están causadas por procesos subyacentes que se comportan de forma diferente en el espacio. Hay varias formas de trabajar con procesos anisotrópicos, considerándolos como generalizaciones más o menos directas de procesos isotrópicos. A continuación se presentan las más usuales.

**Definición 1.8.** *Se dice que el proceso  $Z(\mathbf{s})$  tiene anisotropía geométrica si su variograma es de la forma*

$$2\gamma(\mathbf{h}) = 2\gamma_0(\|A\mathbf{h}\|), \quad \mathbf{h} \in \mathbb{R}^d,$$

siendo  $\gamma_0$  un semivariograma isotrópico y  $A$  una matriz  $d \times d$  que representa una determinada transformación lineal en  $\mathbb{R}^d$ .

De otro lado, se tiene que dados  $Z_1(\cdot), \dots, Z_n(\cdot)$ ,  $n$  procesos intrínsecamente estacionarios independientes, entonces  $Z_1(\cdot) + \dots + Z_n(\cdot)$  es un proceso intrínsecamente estacionario con semivariograma dado por  $\gamma(\mathbf{h}) = \gamma_1(\mathbf{h}) + \dots + \gamma_n(\mathbf{h})$ , siendo  $\gamma_i(\mathbf{h})$  el semivariograma del proceso  $Z_i(\cdot)$ . Esta propiedad permite definir la siguiente generalización de la anisotropía geométrica.

**Definición 1.9.** *Se dice que el proceso  $Z(\mathbf{s})$  tiene anisotropía zonal si su variograma es de la forma*

$$2\gamma(\mathbf{h}) = 2 \sum_{i=1}^n \gamma_0(\|A_i \mathbf{h}\|)$$

siendo  $\gamma_0$  un semivariograma isotrópico y  $A_1, \dots, A_n$  matrices  $d \times d$ .

Otro tipo de tratamiento de la anisotropía es la de suponer que, dado el proceso original  $Z(\mathbf{s})$ , existe una función no lineal  $g(\mathbf{s})$ , de forma que el proceso  $Z(g(\mathbf{s}))$  es un proceso isotrópico estacionario. Esta idea permite analizar tanto

la anisotropía como la no estacionariedad, como se puede ver en Sampson & Guttorp (1992).

En ocasiones se trabaja con procesos en los que la hipótesis de estacionariedad no podría ser admitida, por lo que muchas de las técnicas de la geoestadística clásica no serán directamente aplicables. En los últimos años han surgido gran número de métodos para modelizar este tipo de procesos no estacionarios. Probablemente el más estudiado es el propuesto por Sampson & Guttorp (1992), que presenta un procedimiento de estimación no paramétrica para la estructura de covarianza espacial no estacionaria. Haas (1995) introduce una técnica de kriging de ventanas móviles para la estimación en procesos no estacionarios. Higdon et al. (1999) proponen una alternativa usando una representación de medias móviles de un proceso gaussiano. Nychka & Saltzman (1998) y Holland et al. (1999) desarrollan métodos que extienden la técnica de funciones ortogonales empíricas, muy utilizada por los meteorólogos. Otro modelo para procesos no estacionarios es el propuesto por Fuentes (2001), Fuentes (2002a), Fuentes (2002b) y desarrollado también en Fuentes & Smith (2001). En este modelo, se considera que el proceso es localmente un campo aleatorio estacionario e isotrópico, que se representará con un modelo cuyos parámetros variarían a lo largo de la región de estudio, lo que permite la realización de predicciones sobre el campo aleatorio no estacionario con una única realización del proceso.

### 1.2.2 El covariograma

Dado un proceso estacionario de segundo orden  $Z(\cdot)$ , se ha definido su covariograma como

$$C(\mathbf{h}) = \text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j))$$

con  $\mathbf{s}_i, \mathbf{s}_j \in D$  y  $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$  el vector distancia entre dichas localizaciones. De su definición se deduce fácilmente que  $C(\mathbf{h}) = C(-\mathbf{h})$ . Además, por la desigualdad de Cauchy-Schwartz se cumple que  $|C(\mathbf{h})| \leq C(\mathbf{0}), \forall \mathbf{h} \in \mathbb{R}^d$ .

La función de covarianza  $C(\cdot)$  de un proceso estacionario de segundo orden

debe ser definida positiva, esto es, debe cumplir

$$\sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0 \quad (1.2)$$

para cualquier número finito de localizaciones espaciales  $\{\mathbf{s}_i, i = 1, \dots, n\}$  y de números reales  $\{\varphi_i, i = 1, \dots, n\}$ . Esto es evidente de la definición de covariograma, ya que

$$\sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j C(\mathbf{s}_i - \mathbf{s}_j) = \text{Var} \left( \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) \right) \geq 0$$

### 1.2.3 El variograma

Se ha definido el variograma de un proceso intrínsecamente estacionario  $Z(\cdot)$  como la función

$$2\gamma(\mathbf{h}) = \text{Var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) \quad (1.3)$$

con  $\mathbf{s}_i, \mathbf{s}_j \in D$  y  $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$  el vector distancia entre dichas localizaciones. De (1.3) se deduce fácilmente que  $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$  y que  $\gamma(\mathbf{0}) = 0$ . Obsérvese que el variograma, al contrario del covariograma, no depende de la media del proceso, lo que como se verá tendrá implicaciones en la estimación de ambos.

Una condición necesaria que debe cumplir el variograma es que debe ser una función condicionalmente definida negativa, esto es,

$$\sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j 2\gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0 \quad (1.4)$$

para cualquier conjunto finito de localizaciones espaciales  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \in \mathbb{R}^d$  y para cualquier conjunto de números reales  $\{\varphi_1, \dots, \varphi_n\} \in \mathbb{R}$  con  $\sum_{i=1}^n \varphi_i = 0$ . Esto es evidente de su definición, ya que dados  $\{\varphi_1, \dots, \varphi_n\} \in \mathbb{R}$  con  $\sum_{i=1}^n \varphi_i = 0$ , entonces

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j 2\gamma(\mathbf{s}_i - \mathbf{s}_j) &= -2 \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \\ &= -\text{Var} \left( \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) \right) \leq 0 \end{aligned}$$

Otra condición que debe satisfacer un variograma (Matheron 1971) es que debe tener un ritmo de crecimiento inferior al de  $h^2$ , esto es

$$\lim_{h \rightarrow \infty} \frac{2\gamma(\mathbf{h})}{h^2} = 0$$

### 1.2.4 El correlograma

Sea  $Z(\cdot)$  un proceso estacionario de segundo orden con función de covarianza  $C(\cdot)$ . Se tiene que  $C(\mathbf{0}) = \text{Cov}(Z(\mathbf{s}), Z(\mathbf{s})) = \text{Var}(Z(\mathbf{s}))$ , por lo que  $C(\mathbf{0}) > 0$  a no ser que  $Z(\cdot)$  sea un proceso constante en  $D$ . Se define el correlograma (o función de autocorrelación) como

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

De la definición se desprende que  $\rho(\mathbf{h}) = \rho(-\mathbf{h})$  y que  $\rho(\mathbf{0}) = 1$ .

### 1.2.5 Forma general de estas funciones

El semivariograma representa un índice del cambio que una variable muestra con la distancia. Generalmente, el semivariograma crece con la distancia, ya que en la mayoría de procesos existen mayores similitudes en los valores observados en localizaciones próximas, que disminuyen al aumentar la distancia.

En ocasiones, este crecimiento del semivariograma con la distancia se estabiliza alrededor de un determinado valor  $c_s > 0$ , que es una cota superior de la función (esto es,  $c_s = \lim_{h \rightarrow \infty} \gamma(\mathbf{h})$ ). En este caso se dice que el variograma es acotado y el valor alrededor del cual se estabiliza recibe el nombre de meseta o varianza a-priori (*sill*, en inglés), que es igual por (1.1) a  $C(\mathbf{0})$ , siendo  $C(\cdot)$  el covariograma del proceso. Se llama rango (*range*, en inglés) al valor  $h_r$  en el que el semivariograma alcanza su meseta, esto es, la distancia para el que  $\gamma(h_r) = c_s$  y que representa el valor a partir del cual el covariograma se anula. Para algunos semivariogramas transitivos, la meseta  $c_s$  sólo se alcanza asintóticamente en el límite, por lo que estrictamente hablando el variograma tendrá rango infinito. En este caso se utilizará el término de rango efectivo,

que se define como la distancia en la que el semivariograma alcanza el 95% de su meseta.

Se sabe que el semivariograma es una función que debe cumplir que  $\gamma(\mathbf{0}) = 0$ , pero en la práctica suele ocurrir que  $\lim_{\mathbf{h} \rightarrow 0} \gamma(\mathbf{h}) = c_0 > 0$ , donde  $c_0$  recibe el nombre de pepita (*nugget*, en inglés). Esta discontinuidad en el origen puede estar causada por variaciones de pequeña escala (que sólo tienen sentido en procesos que no son  $L_2$ -continuos), o por errores de medida (es decir, que si se realizan varias observaciones en una misma localización, los valores observados fluctúan alrededor de un determinado valor, que es el valor real). En la práctica, sólo se habrán observado un conjunto de datos  $\{z(\mathbf{s}_i), i = 1, \dots, n\}$ , por lo que no se puede conocer nada del comportamiento del variograma a distancias menores de  $\min\{\|\mathbf{s}_i - \mathbf{s}_j\|, 1 \leq i < j \leq n\}$  y se suele determinar el valor de  $c_0$  extrapolando el comportamiento del variograma a distancias cercanas a cero. En este caso, se define la meseta parcial (partial sill, en inglés) como  $c_s - c_0$ .

Si el proceso  $Z(\cdot)$  es isotrópico, entonces  $2\gamma(\mathbf{h}) = 2\gamma(h)$ , es decir, el variograma depende únicamente de la distancia entre dos localizaciones y no de la dirección. En la Figura 1.1 se muestra la forma típica del variograma de un proceso homogéneo y de su covariograma asociado, donde se puede observar la interpretación de los parámetros introducidos anteriormente.

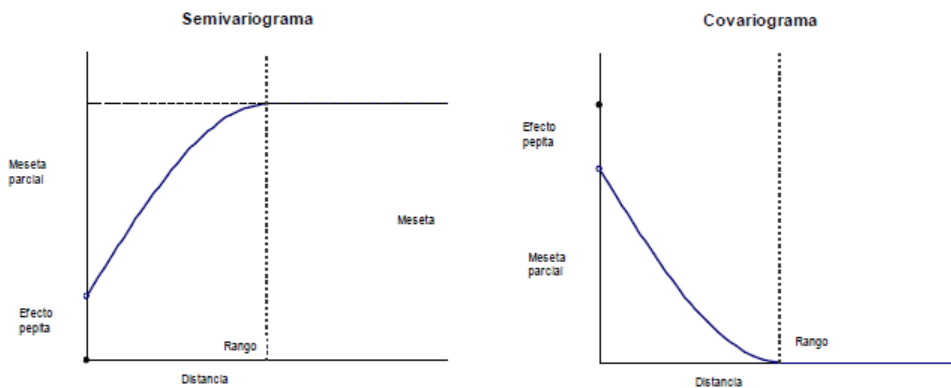


FIGURA 1.1: Forma general del variograma y covariograma de un proceso espacial homogéneo

## 1.3 Estimación del variograma y del covariograma

Dado un proceso espacial  $Z(\cdot)$  intrínsecamente estacionario, se va obtener una estimación del variograma  $2\gamma(\cdot)$  (y del covariograma  $C(\cdot)$  si el proceso es además estacionario de segundo orden) a partir de los valores observados sobre un conjunto de localizaciones  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ . Del conjunto de estimadores propuestos en la literatura para la estimación de estas medidas de variabilidad espacial, se vera a continuación el estimador clásico propuesto por Matheron (1962) o el estimador robusto de Cressie & Hawkins (1980).

### 1.3.1 Estimador clásico

La estimación del variograma más sencillo es la obtenida mediante el estimador del método de los momentos, que recibe el nombre de estimador clásico del variograma. Se tiene que, bajo la hipótesis de estacionariedad intrínseca y por tanto de media del proceso constante, se cumple que

$$2\gamma(\mathbf{h}) = \text{Var}(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})) = E[(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s}))^2]$$

Si los puntos de muestreo  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$  estuviesen localizados sobre una rejilla regular, el estimador del método de los momentos vendrá definido por

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \quad (1.5)$$

donde  $N(\mathbf{h})$  denota todos aquellos pares  $(\mathbf{s}_i, \mathbf{s}_j)$  para los que  $\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}$  y  $|N(\mathbf{h})|$  denota el cardinal de  $N(\mathbf{h})$ . Obsérvese que no es necesario estimar la media  $\mu$  del proceso.

Debido a que (1.5) es esencialmente una media muestral, tiene todas las desventajas asociadas comúnmente a este tipo de estimadores como la no robustez. Se trata de un estimador no paramétrico que es óptimo cuando se dispone de una malla regular de muestreo que sea representativa y la distribución es normal. No obstante, en la práctica el empleo de este estimador



produce en ocasiones variogramas experimentales erráticos, debido a desviaciones del caso ideal para la aplicación del mismo, como son distribuciones alejadas de la normalidad, heterocedasticidad, desviaciones en el muestreo o existencia de valores atípicos.

Para la covarianza, el estimador obtenido por el método de los momentos sería

$$\hat{C}(\mathbf{h}) = |N(\mathbf{h})| \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})} (Z(\mathbf{s}_i) - \hat{Z})(Z(\mathbf{s}_j) - \hat{Z})$$

donde  $\hat{Z} = \frac{1}{n} \sum_{i=1}^n Z(\mathbf{s}_i)$  es un estimador de la media  $\mu$  del proceso y  $N(\mathbf{h})$  se define como antes.

### 1.3.2 Estimador robusto

Como se comentó anteriormente, aunque el método de estimación clásico presenta la ventaja de su facilidad de cálculo, tiene algunos inconvenientes prácticos como que no es robusto frente valores extremos de  $Z(\mathbf{s})$ . Cressie & Hawkins (1980) presentan una variación de (1.5) de mayor robustez como estimador insesgado del variograma y que se define como

$$2\bar{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|(0.457 + 0.494/|N(\mathbf{h})|)} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} \quad (1.6)$$

Los coeficientes de la expresión (1.6) se introducen para asegurar la insesgadez del estimador propuesto. En Cressie (1993) se encuentra un análisis detallado de este estimador comparado con el anterior, así como otras variantes robustas para la estimación empírica del variograma.

## 1.4 Principales modelos de variogramas y covariogramas isotrópicos

En la sección anterior se vio algunos estimadores del variograma o covariograma de los procesos espaciales. El problema es que estas estimaciones no se

pueden utilizar directamente en la práctica geoestadística ya que no satisfacen en general la condición de ser condicionalmente definidas negativas que deben verificar los variogramas (o definidas positivas para los covariogramas). Su uso tendrá efectos no deseables, como la obtención de varianzas negativas en la predicción espacial mediante kriging. Es por ello que, en lugar de utilizar directamente las predicciones, se ajustará a las estimaciones obtenidas anteriormente uno de los modelos válidos de variograma o covariograma que se verán en esta sección.

En las Tablas 1.1 y 1.2 se presentan algunos de los principales modelos de variogramas isotrópicos más utilizados en la práctica geoestadística, junto con sus covariogramas. Como se ha visto, para obtener los correspondientes variogramas bastaría con multiplicar por 2 cada una de las funciones de semivariograma. Estos modelos, además de constituir una herramienta esencial del tratamiento de los datos geoestadísticos, servirán como base para la posterior construcción de modelos espacio-temporales. Como se ha dicho, todos los variogramas y covariogramas que se muestran en dichas tablas son isotrópicos, es decir, dependen de la distancia  $\mathbf{h}$  entre localizaciones únicamente por su módulo  $h = \|\mathbf{h}\|$ , ya que son el punto de arranque sobre los que se construyen modelos más complejos.

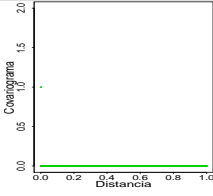
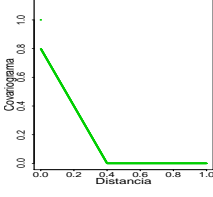
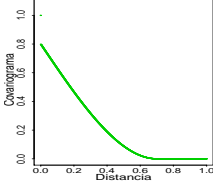
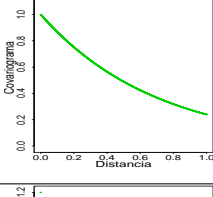
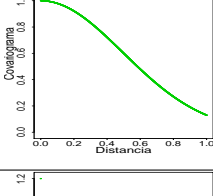
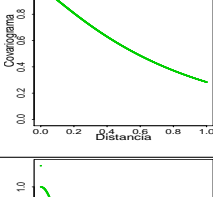
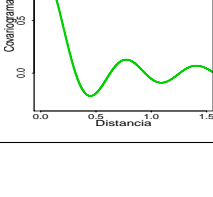
## 1.5 Estimación de los parámetros del variograma

Sea  $Z(\cdot)$  un proceso observado sobre un conjunto de localizaciones  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ . Sean  $\hat{\gamma}(\mathbf{h}_j)$  los valores estimados del semivariograma a partir de los datos aplicando alguno de los métodos que se han visto en la Sección 1.2. Aunque son muchas las buenas propiedades de estos estimadores, carecen de la propiedad de ser semidefinidos positivos, con lo que sería posible que algunas predicciones espaciales derivadas a partir de tales estimadores presenten varianzas negativas. La forma más común de evitar esta dificultad es reemplazando el semivariograma empírico por algún modelo paramétrico  $\gamma(\mathbf{h}, \boldsymbol{\vartheta})$  de los que se han presentado anteriormente que se aproxime a la dependencia espacial encontrada

TABLA 1.1: Formas funcionales de algunos variogramas

Modelo	Función de Semivariograma	Variograma
Efecto pepita	$\gamma(h) = \begin{cases} c_0 & \text{si } h > 0 \\ 0 & \text{si } h = 0 \end{cases}$	
Lineal con meseta	$\gamma(h) = \begin{cases} c_0 + c_s \left(\frac{h}{a_l}\right) & \text{si } 0 \leq h \leq a_l \\ c_0 + c_s & \text{si } h > a_l \end{cases}$	
Esférico	$\gamma(h) = \begin{cases} 0 & \text{si } h = 0 \\ c_0 + c_s \left(\frac{3}{2}\left(\frac{h}{a_s}\right) - \frac{1}{2}\left(\frac{h}{a_s}\right)^3\right) & \text{si } 0 \leq h \leq a_s \\ c_0 + c_s & \text{si } h > a_s \end{cases}$	
Exponencial	$\gamma(h) = \begin{cases} c_0 + c_s(1 - \exp(-3h/a_e)) & \text{si } h > 0 \\ 0 & \text{si } h = 0 \end{cases}$	
Gaussiano	$\gamma(h) = \begin{cases} c_0 + c_s(1 - \exp(-3h^2/a_g^2)) & \text{si } h > 0 \\ 0 & \text{si } h = 0 \end{cases}$	
Circular	$\gamma(h) = \begin{cases} 0 & \text{si } h = 0 \\ c_0 + c_s \left(1 - \frac{2}{\pi} \cos^{-1}\left(\frac{h}{a_c}\right) - \frac{2h}{\pi a_c} \sqrt{1 - \left(\frac{h}{a_c}\right)^2}\right) & \text{si } 0 \leq h \leq a_c \\ c_0 + c_s & \text{si } h > a_c \end{cases}$	
Efecto agujero	$\gamma(h) = \begin{cases} c_0 + c_s \left(1 - \frac{\sin(h)}{h}\right) & \text{si } h > 0 \\ 0 & \text{si } h = 0 \end{cases}$	

TABLA 1.2: Formas funcionales de algunos covariogramas

Modelo	Función de Covariograma	Covariograma
Efecto pepita	$C(h) = \begin{cases} 0 & \text{si } h > 0 \\ c_0 & \text{si } h = 0 \end{cases}$	
Lineal con meseta	$C(h) = \begin{cases} c_s(1 - \frac{h}{a_l}) & \text{si } 0 \leq h \leq a_l \\ 0 + c_s & \text{si } h > a_l \end{cases}$	
Esférico	$C(h) = \begin{cases} c_0 + c_s & \text{si } h = 0 \\ c_s(1 - \frac{3}{2}(\frac{h}{a_s}) + \frac{1}{2}(\frac{h}{a_s})^3) & \text{si } 0 \leq h \leq a_s \\ 0 & \text{si } h > a_s \end{cases}$	
Exponencial	$C(h) = \begin{cases} c_s(\exp(-3h/a_e)) & \text{si } h > 0 \\ c_0 + c_s & \text{si } h = 0 \end{cases}$	
Gaussiano	$C(h) = \begin{cases} c_s(\exp(-3h^2/a_g^2)) & \text{si } h > 0 \\ c_0 + c_s & \text{si } h = 0 \end{cases}$	
Circular	$C(h) = \begin{cases} c_0 + c_s & \text{si } h = 0 \\ c_1(\frac{2}{\pi} \cos^{-1}(\frac{h}{a_c}) + \frac{2h}{\pi a_c} \sqrt{(1 - \frac{h}{a_c})^2}) & \text{si } 0 \leq h \leq a_c \\ 0 & \text{si } h > a_c \end{cases}$	
Efecto agujero	$\gamma(h) = \begin{cases} c_s(\frac{\sin(h)}{h}) & \text{si } h > 0 \\ c_0 + c_s & \text{si } h = 0 \end{cases}$	

por el semivariograma empírico, y del que se sabe cumple la condición de ser semidefinido positivo. Obsérvese que, en general, no es necesario restringirse a modelos isotrópicos, aunque suelen ser los primeros que son considerados.

El objetivo será elegir de entre todos los semivariogramas posibles  $\{\gamma(\mathbf{h}, \boldsymbol{\vartheta}), \boldsymbol{\vartheta} \in \Upsilon\}$  aquél que mejor se ajuste a las observaciones realizadas, obteniendo con ello un modelo de semivariograma que más tarde será utilizado en el proceso de predicción espacial. En esta sección se verá los principales métodos de estimación.

### 1.5.1 Estimación por mínimos cuadrados

La estimación por mínimos cuadrados ordinarios (OLS) consiste en obtener el valor  $\hat{\boldsymbol{\vartheta}}$  que minimiza

$$\sum_{j=1}^n (\hat{\gamma}(\mathbf{h}_j) - \gamma(\mathbf{h}_j, \boldsymbol{\vartheta}))^2 = [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\vartheta})]' [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\vartheta})]$$

siendo  $\hat{\boldsymbol{\gamma}} = [\hat{\gamma}(\mathbf{h}_1), \dots, \hat{\gamma}(\mathbf{h}_n)]'$  y  $\boldsymbol{\gamma}(\boldsymbol{\vartheta}) = [\gamma(\mathbf{h}_1, \boldsymbol{\vartheta}), \dots, \gamma(\mathbf{h}_n, \boldsymbol{\vartheta})]'$ . Un problema que presenta este procedimiento es que en este caso las estimaciones están correladas y tienen varianzas diferentes.

Una solución es aplicar mínimos cuadrados generalizados (GLS), que consiste en minimizar

$$[\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\vartheta})]' V(\boldsymbol{\vartheta})^{-1} [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\vartheta})]$$

siendo  $V(\boldsymbol{\vartheta})$  la matriz de varianzas-covarianzas de  $\hat{\boldsymbol{\gamma}}$ , que depende del valor  $\boldsymbol{\vartheta}$  desconocido y cuyos elementos pueden ser además difíciles de estimar.

Un compromiso entre las dos anteriores es la estimación por mínimos cuadrados ponderados (WLS), que consiste en minimizar

$$\sum_{j=1}^n w_j [\hat{\gamma}(\mathbf{h}_j) - \gamma(\mathbf{h}_j, \boldsymbol{\vartheta})]^2 = [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\vartheta})]' W(\boldsymbol{\vartheta})^{-1} [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\vartheta})] \quad (1.7)$$

siendo  $W(\boldsymbol{\vartheta})$  una matriz diagonal cuyos elementos son las varianzas de  $\hat{\gamma}$ , las cuales pueden aproximarse bajo la hipótesis que el proceso es gaussiano y las estimaciones son incorreladas por  $2\gamma(\mathbf{h}_j, \boldsymbol{\vartheta})^2 / N(\mathbf{h}_j)$ , con  $N(\mathbf{h}_j)$  el número de

localizaciones a distancia  $\mathbf{h}_j$ . Por tanto, los pesos de (1.7) vendrán dados por  $w_j = N(\mathbf{h}_j)/(2\gamma(\mathbf{h}_j, \boldsymbol{\vartheta})^2)$ .

En general, los tres estimadores OLS, WLS y GLS aparecen en orden creciente de eficiencia pero decreciente en simplicidad, siendo la estimación por mínimos cuadrados ponderados la más utilizada en la práctica estadística debido a la facilidad de su implementación y a las ventajas computacionales que presenta. No obstante, presenta inconvenientes prácticos como, por ejemplo, depende de las estimaciones del semivariograma, son correladas, muy sensibles a la selección de las distancias y las regiones de tolerancia utilizadas para su cálculo.

### 1.5.2 Estimación mediante máxima verosimilitud

Supóngase que el proceso espacial analizado es un proceso gaussiano, el cual se puede escribir como

$$Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i) \quad i = 1, \dots, n \quad (1.8)$$

donde la media del proceso  $\mu(\mathbf{s}_i) = \theta_0 + v(\mathbf{s}_i)' \boldsymbol{\theta}$  es una función lineal de un conjunto de  $p$  regresores y  $\varepsilon(\mathbf{s}_i)$  representa el error espacial, con  $v(\mathbf{s}_i) = (v_1(\mathbf{s}_i), \dots, v_p(\mathbf{s}_i))'$  es un vector que contiene las variables explicativas asociadas a la localización espacial  $\mathbf{s}_i$ ,  $\theta_0$  es el parámetro desconocido asociado al intercepto, y  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)'$  es un vector de parámetros desconocidos. En este caso es bastante sencillo obtener la forma exacta de la verosimilitud y maximizarla numéricamente.

En forma matricial el modelo (1.8) se puede expresar como:

$$\mathbf{Z}_s = \mathbf{V} \tilde{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}_s \quad (1.9)$$

donde  $\mathbf{Z}_s = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$ ,  $\mathbf{V} = (\mathbf{1}, V_1, \dots, V_n)$  es la matriz de diseño de dimensión  $n \times (p+1)$  con el vector de intercepto  $\mathbf{1}$  de dimensión  $n \times 1$  y  $p$  variables explicativas  $V_j = (v_j(\mathbf{s}_1), \dots, v_j(\mathbf{s}_n))'$  de dimensión  $n \times 1$ , con  $j = 1, \dots, p$ . Además,  $\tilde{\boldsymbol{\theta}} = (\theta_0, \boldsymbol{\theta}')'$  el vector de dimensión  $(p+1) \times 1$  de parámetros desconocidos,  $\boldsymbol{\varepsilon}_s = (\varepsilon(\mathbf{s}_1), \dots, \varepsilon(\mathbf{s}_n))'$  y  $\Sigma_{\boldsymbol{\vartheta}}$  la matriz de varianzas-covarianzas de

las observaciones. La estimación mediante máxima verosimilitud consiste en obtener de forma simultánea los valores de  $\tilde{\boldsymbol{\theta}}$  y  $\boldsymbol{\vartheta}$  que maximizan la función de distribución conjunta normal multivariante, o lo que es lo mismo, que minimizan  $-2$  veces la log-verosimilitud

$$L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) = (\mathbf{Z}_s - \mathbf{V}\tilde{\boldsymbol{\theta}})' \Sigma_{\boldsymbol{\vartheta}}^{-1} (\mathbf{Z}_s - \mathbf{V}\tilde{\boldsymbol{\theta}}) + \log|\Sigma_{\boldsymbol{\vartheta}}| + n\log(2\pi) \quad (1.10)$$

Minimizar la ecuación anterior puede ser computacionalmente demasiado costoso si se tienen muchos parámetros en  $\tilde{\boldsymbol{\theta}}$ . Generalmente,  $\boldsymbol{\vartheta}$  sólo incluye tres parámetros (la pepita, la meseta parcial y el rango). Se puede lograr un ahorro computacional minimizando (1.10) en dos fases: en una primera etapa se supone que  $\tilde{\boldsymbol{\theta}}$  conocido, y por tanto  $\Sigma_{\boldsymbol{\vartheta}}$  también, con lo que la estimación de  $\tilde{\boldsymbol{\theta}}$  se obtendría mediante el estimador de mínimos cuadrados generalizado  $\hat{\tilde{\boldsymbol{\theta}}} = (\mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{V})^{-1} \mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{Z}_s$ . Ahora se puede sustituir este estimador en la expresión (1.10), obteniendo

$$L_{ML}(\boldsymbol{\vartheta}) = (\mathbf{Z}_s - \mathbf{V}\hat{\tilde{\boldsymbol{\theta}}})' \Sigma_{\boldsymbol{\vartheta}}^{-1} (\mathbf{Z}_s - \mathbf{V}\hat{\tilde{\boldsymbol{\theta}}}) + \log|\Sigma_{\boldsymbol{\vartheta}}| + n\log(2\pi) \quad (1.11)$$

Ahora la expresión (1.11) a minimizar únicamente depende de  $\boldsymbol{\vartheta}$ , lo que hace el proceso más sencillo y permite obtener  $\hat{\boldsymbol{\vartheta}}$ . Para obtener  $\hat{\tilde{\boldsymbol{\theta}}}$  bastaría con sustituir esta estimación en la expresión del estimador de mínimos cuadrados generalizado. Aunque la estimación por máxima verosimilitud es asintóticamente insesgada, presenta un sesgo considerable para muestras pequeñas porque los elementos de la diagonal de  $(\mathbf{V}'\Sigma_{\hat{\boldsymbol{\vartheta}}}^{-1}\mathbf{V})^{-1}$  son demasiado pequeños, lo que produce infraestimaciones de los parámetros  $\tilde{\boldsymbol{\theta}}$ .

La estimación por máxima verosimilitud restringida (REML) tiene muchas mejores propiedades de sesgo que la anterior. La función a minimizar en este caso viene dada por la expresión

$$\begin{aligned} L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) &= (\mathbf{Z}_s - \mathbf{V}\tilde{\boldsymbol{\theta}})' \Sigma_{\boldsymbol{\vartheta}}^{-1} (\mathbf{Z}_s - \mathbf{V}\tilde{\boldsymbol{\theta}}) + \log|\Sigma_{\boldsymbol{\vartheta}}| \\ &\quad + \log(|\mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{V}|) + (n-p)\log(2\pi) \end{aligned} \quad (1.12)$$

Como antes, se sustituye el parámetro desconocido  $\boldsymbol{\vartheta}$  por el estimador de mínimos cuadrados generalizado  $\hat{\boldsymbol{\vartheta}}$  obtenido a partir de unos valores iniciales

de  $\boldsymbol{\vartheta}$ , con lo que la función a minimizar sería

$$L(\boldsymbol{\vartheta}) = \left(\mathbf{Z}_s - \mathbf{V}\hat{\boldsymbol{\theta}}\right)' \Sigma_{\boldsymbol{\vartheta}}^{-1} \left(\mathbf{Z}_s - \mathbf{V}\hat{\boldsymbol{\theta}}\right) + \log|\Sigma_{\boldsymbol{\vartheta}}| \\ + \log(|\mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{V}|) + (n-p)\log(2\pi) \quad (1.13)$$

Después de minimizar esta expresión se obtiene una estimación  $\hat{\boldsymbol{\theta}}$ , que se sustituye en la expresión del estimador de mínimos cuadrados generalizado para obtener  $\hat{\boldsymbol{\theta}} = \left(\mathbf{V}'\Sigma_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{V}\right)^{-1} \mathbf{V}'\Sigma_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{Z}_s$ .

Cuando la matriz de diseño  $V$  contiene únicamente una columna de unos y  $\tilde{\boldsymbol{\theta}}$  un único parámetro que representa la media general del proceso, entonces el variograma empírico es insesgado respecto al variograma teórico, y el método de mínimos cuadrados funciona suficientemente bien (Cressie 1993). En cambio, cuando  $\tilde{\boldsymbol{\theta}}$  tiene múltiples parámetros se necesita ajustar un variograma a los residuos del modelo de regresión y las estimaciones dejan de ser insesgadas. Aunque los procedimientos de máxima verosimilitud se han desarrollado bajo la hipótesis de que los datos provienen de una distribución normal multivariante, desviaciones en la distribución del error del modelo no afectan demasiado a sus estimaciones con estos métodos por ser suficientemente robustos. Otra ventaja es que no se necesitan condiciones previas en el proceso de estimación, tales como distancias máximas y regiones de tolerancia. Además, de estas técnicas se pueden utilizar procedimientos de diagnóstico clásicos para los modelos ajustados (Faraway 2005) como el valor de menos dos veces la log-verosimilitud asociada ( $-2LL$ ), el criterio de información de Akaike (AIC, Akaike (1973) y Akaike (1974)) o el criterio de información bayesiano (BIC, Schwarz (1978)).

## 1.6 Predicción espacial

El objetivo de la predicción espacial será, a partir de las observaciones realizadas, obtener una estimación de  $g(Z(\mathbf{s}_0))$ , siendo  $g(Z(\cdot))$  alguna característica de interés del proceso  $Z(\cdot)$  y  $\mathbf{s}_0$  una región de interés en  $D$ , es decir sobre una determinada localización  $\mathbf{s}_0 \in D$ , con lo que  $g(Z(\cdot)) \equiv Z(\cdot)$ .



Aunque existen multitud de procedimientos de predicción espacial, esta sección se va a centrar en el kriging, que es el término genérico adoptado en geoestadística para dar nombre a una metodología de interpolación basada en una familia de algoritmos de regresión generalizada por mínimos cuadrados que utiliza las propiedades de segundo orden del proceso  $Z(\cdot)$ . Recibe este nombre en reconocimiento a los trabajos pioneros de Krige (1951).

Se parte del problema de la predicción del proceso original  $Z(\cdot)$  sobre una determinada localización  $\mathbf{s}_0 \in D$ , esto es, de la predicción puntual sin considerar error de medida y luego se generalizan los resultados obtenidos para problemas más generales. El predictor que el kriging utiliza para la predicción de  $Z(\mathbf{s}_0)$  es un predictor lineal de los datos

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) = \boldsymbol{\varphi}' \mathbf{Z}_s \quad (1.14)$$

donde  $\varphi_i$  es el peso asignado a cada uno de los datos que intervienen en el sumatorio y  $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_n)'$ . De todos los posibles predictores lineales, la técnica de kriging selecciona aquel predictor lineal que sea insesgado ( $E[Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)] = 0$ ) y además sea óptimo en el sentido que minimize la varianza del error cuadrático medio de la predicción ( $\min Var[Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)]$ ). Por este motivo el predictor resultante recibe el nombre de BLUP (Best Linear Unbiased Predictor).

Si denotamos por  $\sigma_K^2(\mathbf{s}_0)$  el error cuadrático medio de la predicción, bajo la condición de insesgadez vendrá dado por

$$\sigma_K^2(\mathbf{s}_0) = Var \left[ \left( Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0) \right) \right] = E \left( Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0) \right)^2 \quad (1.15)$$

A continuación se describen algunos de los tipos de kriging más utilizados en la práctica geoestadística, como son el kriging ordinario (supone que la media es constante pero desconocida) y el kriging universal (supone que la media es desconocida y no es una constante, aunque sí una función lineal de un conjunto de variables que dependen de la localización). En Cressie (1993) se encuentra un mayor desarrollo de cada uno de estos métodos y otros procedimientos de gran aplicación como el kriging indicador, el kriging lognormal, el kriging transgaussiano, el kriging robusto y el cokriging para el caso multinomial.

### 1.6.1 Kriging ordinario

Suponga que el proceso espacial  $Z(\cdot)$  es intrínsecamente estacionario con variograma  $2\gamma(\cdot)$ . Se tiene entonces que  $\varepsilon(\cdot)$  también es un proceso intrínsecamente estacionario con el mismo variograma que  $Z(\cdot)$ , ya que

$$\begin{aligned} 2\gamma_\varepsilon(\mathbf{h}) &= \text{Var}(\varepsilon(\mathbf{s}_1) - \varepsilon(\mathbf{s}_2)) = \text{Var}(Z(\mathbf{s}_1) - \mu(\mathbf{s}_1) - Z(\mathbf{s}_2) + \mu(\mathbf{s}_2)) \\ &= \text{Var}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)) = 2\gamma(\mathbf{h}) \end{aligned}$$

En este caso se considera que la tendencia del proceso  $\mu(\mathbf{s})$  es desconocida pero constante a lo largo de toda el área de estudio  $D$ , es decir,  $\mu(\mathbf{s}) = \mu, \forall \mathbf{s} \in D$ . La insesgadez del predictor (1.14) implica que los pesos  $\varphi_i$  deben sumar uno, ya que

$$\text{E}[Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)] = \text{E}[Z(\mathbf{s}_0)] - \text{E}\left[\sum_{i=1}^n \varphi_i Z(\mathbf{s}_i)\right] = \theta_0 - \sum_{i=1}^n \varphi_i \theta_0 = 0 \quad (1.16)$$

entonces  $\sum_{i=1}^n \varphi_i = 1$ . El objetivo es calcular los pesos  $\varphi_i$  de (1.14) que minimizan (1.15) bajo la restricción de insesgadez dada en (1.16). Si se define  $l$  el multiplicador de Lagrange asociado a la restricción de insesgadez, este problema es equivalente a minimizar la expresión

$$\text{E}\left(Z(\mathbf{s}_0) - \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i)\right)^2 - 2l\left(\sum_{i=1}^n \varphi_i - 1\right)$$

donde

$$\begin{aligned} \left(Z(\mathbf{s}_0) - \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i)\right)^2 &= \left(\sum_{i=1}^n \varphi_i Z(\mathbf{s}_0) - \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i)\right)^2 \\ &= \left(\sum_{i=1}^n \varphi_i (Z(\mathbf{s}_0) - Z(\mathbf{s}_i))\right)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j [Z(\mathbf{s}_0) - Z(\mathbf{s}_i)][Z(\mathbf{s}_0) - Z(\mathbf{s}_j)] \end{aligned}$$

Pero se tiene que

$$\begin{aligned} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2 &= [Z(\mathbf{s}_i) - Z(\mathbf{s}_0) + Z(\mathbf{s}_0) - Z(\mathbf{s}_j)]^2 = [Z(\mathbf{s}_i) - Z(\mathbf{s}_0)]^2 \\ &\quad + [Z(\mathbf{s}_j) - Z(\mathbf{s}_0)]^2 - 2[Z(\mathbf{s}_i) - Z(\mathbf{s}_0)][Z(\mathbf{s}_j) - Z(\mathbf{s}_0)] \end{aligned}$$

Por lo tanto, se tiene que

$$\begin{aligned} E \left( Z(\mathbf{s}_0) - \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) \right)^2 &= E \left[ \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j (Z(\mathbf{s}_0) - Z(\mathbf{s}_i))(Z(\mathbf{s}_0) - Z(\mathbf{s}_j)) \right] \\ &= 2 \sum_{i=1}^n \varphi_i E[Z(\mathbf{s}_0) - Z(\mathbf{s}_i)]^2 / 2 - \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j E[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2 / 2 \end{aligned}$$

Así, la expresión a minimizar sería

$$2 \sum_{i=1}^n \varphi_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \gamma(\mathbf{s}_i - \mathbf{s}_j) - 2l \left( \sum_{i=1}^n \varphi_i - 1 \right)$$

Si se deriva respecto a  $\varphi_i$  y  $l$ , y se iguala a cero, se tiene el siguiente sistema de ecuaciones lineales

$$\begin{aligned} - \sum_{j=1}^n \varphi_j \gamma(\mathbf{s}_i - \mathbf{s}_j) + \gamma(\mathbf{s}_0 - \mathbf{s}_i) &= l \quad i = 1, \dots, n \\ \sum_{i=1}^n \varphi_i &= 1 \end{aligned}$$

que puede expresarse de forma matricial como

$$\begin{pmatrix} \Gamma & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\varphi} \\ l \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma} \\ 1 \end{pmatrix} \quad (1.17)$$

siendo  $\Gamma$  la matriz  $n \times n$  cuyo elemento  $(i, j)$ -ésimo es  $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ ,  $\boldsymbol{\gamma}$  el vector de dimensión  $n$  cuyo elemento  $i$ -ésimo es  $\gamma(\mathbf{s}_0 - \mathbf{s}_i)$ ,  $\mathbf{1}$  el vector unidad de dimensión  $n$  y  $\boldsymbol{\varphi}$  el vector con los coeficientes del predictor del kriging ordinario.

Resolviendo (1.17), se llega a que

$$\hat{\boldsymbol{\varphi}}'_{KO} = \left( \boldsymbol{\gamma} + \mathbf{1} \frac{(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}} \right)' \boldsymbol{\Gamma}^{-1} \quad (1.18)$$

$$\hat{l}_{KO} = - \frac{1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}} \quad (1.19)$$

En este caso, el error cuadrático medio de la predicción que se ha minimizado (y que también se conoce como varianza del kriging) viene dado por

la siguiente expresión

$$\begin{aligned}\hat{\sigma}_{KO}^2(\mathbf{s}_0) &= \boldsymbol{\gamma}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} - \frac{(1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})^2}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}} \\ &= 2 \sum_{i=1}^n \varphi_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \gamma(\mathbf{s}_i - \mathbf{s}_j)\end{aligned}\quad (1.20)$$

Todas estas expresiones pueden escribirse en términos de la función de covarianza en el caso de tener un proceso estacionario de segundo orden. Supóngase que se tiene un proceso  $Z(\cdot)$  estacionario de segundo orden de media cero y covariograma  $C(\cdot)$ . En este caso

$$\begin{aligned}\left[ Z(\mathbf{s}_0) - \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) \right]^2 &= \left[ Z(\mathbf{s}_0) - \theta_0 + \theta_0 - \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) \right]^2 \\ &= [Z(\mathbf{s}_0) - \theta_0]^2 + \left[ \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) - \theta_0 \right]^2 - 2[Z(\mathbf{s}_0) - \theta_0] \left[ \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) - \theta_0 \right] \\ &= [Z(\mathbf{s}_0) - \theta_0]^2 + \left[ \sum_{i=1}^n \varphi_i (Z(\mathbf{s}_i) - \theta_0) \right]^2 - 2[Z(\mathbf{s}_0) - \theta_0] \left[ \sum_{i=1}^n \varphi_i (Z(\mathbf{s}_i) - \theta_0) \right] \\ &= [Z(\mathbf{s}_0) - \theta_0]^2 + \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j [Z(\mathbf{s}_i) - \theta_0][Z(\mathbf{s}_j) - \theta_0] \\ &\quad - 2 \sum_{i=1}^n \varphi_i [Z(\mathbf{s}_0) - \theta_0][Z(\mathbf{s}_i) - \theta_0]\end{aligned}$$

Tomando esperanzas en los dos lados de la expresión anterior se tiene que

$$E \left[ Z(\mathbf{s}_0) - \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) \right]^2 = C(\mathbf{0}) + \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j C(\mathbf{s}_i - \mathbf{s}_j) - 2 \sum_{i=1}^n \varphi_i C(\mathbf{s}_0 - \mathbf{s}_i)$$

Luego la expresión a minimizar es

$$C(\mathbf{0}) + \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j C(\mathbf{s}_i - \mathbf{s}_j) - 2 \sum_{i=1}^n \varphi_i C(\mathbf{s}_0 - \mathbf{s}_i) - 2l \left( \sum_{i=1}^n \varphi_i - 1 \right)$$

Si se deriva como antes respecto a  $\varphi_i$  y  $l$  y se iguala a cero, se tiene un sistema de ecuaciones lineales que puede expresarse en forma matricial como

$$\begin{pmatrix} \Sigma_{\varphi} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\varphi} \\ l \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ 1 \end{pmatrix}$$

siendo  $\Sigma_{\vartheta}$  la matriz  $n \times n$  cuyo elemento  $(i, j)$ -ésimo es  $C(\mathbf{s}_i - \mathbf{s}_j)$  y  $\mathbf{c}$  el vector de dimensión  $n$  cuyo elemento  $i$ -ésimo es  $C(\mathbf{s}_0 - \mathbf{s}_i)$ . Para el predictor del kriging ordinario, dado en (1.14), se tiene que

$$\hat{\boldsymbol{\varphi}}'_{KO} = \left( \mathbf{c} + \mathbf{1} \frac{(1 - \mathbf{1}'\Sigma_{\vartheta}^{-1}\mathbf{c})}{\mathbf{1}'\Sigma_{\vartheta}^{-1}\mathbf{1}} \right)' \Sigma_{\vartheta}^{-1} \quad (1.21)$$

$$\hat{l}_{KO} = \frac{1 - \mathbf{1}'\Sigma_{\vartheta}^{-1}\mathbf{c}}{\mathbf{1}'\Sigma_{\vartheta}^{-1}\mathbf{1}} \quad (1.22)$$

La varianza del kriging vendría dada por

$$\hat{\sigma}_{KO}^2(\mathbf{s}_0) = C(\mathbf{0}) - \boldsymbol{\varphi}'\mathbf{c} + \frac{1 - \mathbf{1}'\Sigma_{\vartheta}^{-1}\mathbf{c}}{\mathbf{1}'\Sigma_{\vartheta}^{-1}\mathbf{1}} \quad (1.23)$$

## 1.6.2 Kriging universal

El kriging universal generaliza el kriging ordinario, permitiendo que el valor medio del proceso no sea constante sino una combinación lineal de funciones conocidas o covariables ligadas a las mismas localizaciones. De esta forma, el kriging universal incorpora términos de regresión y correlación espacial.

La hipótesis de partida es que el proceso objeto de estudio  $Z(\mathbf{s})$  puede descomponerse como (1.10). Si la matriz de diseño  $\mathbf{V}$  contiene únicamente funciones polinomiales de las coordenadas espaciales de la localización  $\mathbf{s}$ , entonces el método de predicción recibe el nombre de kriging universal con tendencia interna.

Como se hizo en el kriging ordinario, se quiere predecir el valor de  $Z(\mathbf{s}_0)$  a partir del conjunto de observaciones  $\mathbf{Z}_s$ . Para ello, se utilizara un predictor lineal de la forma (1.14), donde al igual que en kriging ordinario los coeficientes  $\varphi_i$  se obtienen al imponer que el estimador resultante sea insesgado y de mínima varianza. En este caso, la insesgidez de (1.14) implica que  $\boldsymbol{\varphi}'\mathbf{V} = \mathbf{v}'(\mathbf{s}_0)$ , donde  $\mathbf{v}(\mathbf{s}_0) = (1, v_1(\mathbf{s}_0), \dots, v_p(\mathbf{s}_0))'$  ya que en este caso

$$\text{E} \left[ \hat{Z}(\mathbf{s}_0) \right] = \text{E}[\boldsymbol{\varphi}'\mathbf{Z}_s] = \boldsymbol{\varphi}'\text{E}[\mathbf{Z}_s] = \boldsymbol{\varphi}'\mathbf{V}\tilde{\boldsymbol{\theta}} = \mathbf{v}'(\mathbf{s}_0)\tilde{\boldsymbol{\theta}} = \text{E}[Z(\mathbf{s}_0)] \quad (1.24)$$

Obsérvese que el kriging universal es una generalización del kriging ordinario, ya que para  $\mathbf{V} = \mathbf{1}$  y  $\tilde{\boldsymbol{\theta}} = \theta_0$  se tiene el modelo  $\mathbf{Z}_s = \theta_0\mathbf{1} + \boldsymbol{\varepsilon}$  que se asumió en aquel, y además, la condición (1.24) se reduce a (1.17).

Procediendo como en el caso del kriging ordinario, se llega al siguiente sistema de ecuaciones lineales

$$\begin{pmatrix} \Gamma & \mathbf{V} \\ \mathbf{V}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\varphi} \\ \mathbf{l} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma} \\ \mathbf{v}(\mathbf{s}_0) \end{pmatrix}$$

siendo  $\Gamma$  una matriz  $n \times n$  cuyo elemento  $(i, j)$ -ésimo es  $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ ,  $\boldsymbol{\gamma}$  el vector de dimensión  $n$  cuyo elemento  $i$ -ésimo es  $\gamma(\mathbf{s}_0 - \mathbf{s}_i)$  y  $\mathbf{l}$  un vector de  $p + 1$  multiplicadores de Lagrange.

Resolviendo este sistema de ecuaciones, se tiene que

$$\begin{aligned} \hat{\boldsymbol{\varphi}}'_{UK} &= [\boldsymbol{\gamma} + \mathbf{V}(\mathbf{V}'\Gamma^{-1}\mathbf{V})^{-1}(\mathbf{v}(\mathbf{s}_0) - \mathbf{V}'\Gamma^{-1}\boldsymbol{\gamma})]'\Gamma^{-1} \\ \hat{\mathbf{l}}_{UK} &= -(\mathbf{V}'\Gamma^{-1}\mathbf{V})^{-1}[\mathbf{v}(\mathbf{s}_0) - \mathbf{V}'\Gamma^{-1}\boldsymbol{\gamma}] \end{aligned} \quad (1.25)$$

El error cuadrático de la predicción sería, en este caso

$$\hat{\sigma}_{UK}^2(\mathbf{s}_0) = \boldsymbol{\gamma}'\Gamma^{-1} - [\mathbf{v}(\mathbf{s}_0) - \mathbf{V}'\Gamma^{-1}\boldsymbol{\gamma}]'(\mathbf{V}'\Gamma^{-1}\mathbf{V})^{-1}[\mathbf{v}(\mathbf{s}_0) - \mathbf{V}'\Gamma^{-1}\boldsymbol{\gamma}] \quad (1.26)$$

Se puede expresar también todas estas ecuaciones en función de la función de covarianza del proceso  $C(\cdot)$ . En este caso

$$\begin{aligned} \hat{\boldsymbol{\varphi}}'_{UK} &= [\mathbf{c} + \mathbf{V}(\mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{V})^{-1}(\mathbf{v}(\mathbf{s}_0) - \mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{c})]'\Sigma_{\boldsymbol{\vartheta}}^{-1} \\ \hat{\mathbf{l}}_{UK} &= -(\mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{V})^{-1}[\mathbf{v}(\mathbf{s}_0) - \mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{c}] \end{aligned} \quad (1.27)$$

siendo  $\Sigma_{\boldsymbol{\vartheta}}$  la matriz  $n \times n$  cuyo elemento  $(i, j)$ -ésimo es  $C(\mathbf{s}_i - \mathbf{s}_j)$  y  $\mathbf{c}$  el vector de dimensión  $n$  cuyo elemento  $i$ -ésimo es  $C(\mathbf{s}_0 - \mathbf{s}_i)$ .

Respecto a la estimación óptima de los parámetros de la media  $\tilde{\boldsymbol{\theta}}$ , como los datos  $\mathbf{Z}_s$  satisfacen un modelo lineal general con  $E(\mathbf{Z}_s) = \mathbf{V}\tilde{\boldsymbol{\theta}}$  y  $\text{Var}(\mathbf{Z}_s) = \Sigma_{\boldsymbol{\vartheta}}$ , puede obtenerse por mínimos cuadrados generalizados como

$$\hat{\boldsymbol{\theta}}_{gls} = (\mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{V})^{-1}\mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{Z}_s$$

con  $\text{Var}(\hat{\boldsymbol{\theta}}_{gls}) = (\mathbf{V}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{V})^{-1}$ . A partir de estas expresiones, se puede obtener intervalos de confianza para estos parámetros o combinaciones lineales de los mismos. Se deduce que aun cuando para estimar los parámetros de la media de forma óptima es necesario conocer las covarianzas involucradas en  $\Sigma_{\boldsymbol{\vartheta}}$ , para la predicción óptima de  $Z(\mathbf{s}_0)$  únicamente se ha de conocer los variogramas implicados en  $\Gamma$ .

## 1.7 Diagnóstico mediante validación cruzada

La validación cruzada es una técnica que permite evaluar la capacidad predictiva del modelo seleccionado. Supóngase que se ha observado el valor del proceso  $Z(\mathbf{s})$  sobre un conjunto de localizaciones  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ . Sean  $\hat{Z}(\mathbf{s}_i)$  y  $\hat{\sigma}(\mathbf{s}_i)$  el valor predicho y el error típico de la predicción para la localización  $\mathbf{s}_i$  obtenida a partir de las observaciones  $\{\mathbf{s}_1, \dots, \mathbf{s}_{i-1}, \mathbf{s}_{i+1}, \dots, \mathbf{s}_n\}$ .

Esta técnica se justifica debido a que los métodos de interpolación kriging son exactos, es decir los valores de los pronósticos coinciden con los valores observados para los puntos muestreados y da una idea de qué tan buenos son los pronósticos, por lo cual brinda información acerca de cuál modelo provee predicciones más exactas.

Como se observa el método consiste en excluir la observación de uno de los  $n$  puntos muestrales (por lo general asociados a un vecindario), y con los  $n - 1$  valores restantes y el modelo de semivarianza escogido, predecir vía kriging el valor de la variable en estudio en la ubicación del punto que se excluyó. Si el modelo de semivarianza o covarianza elegido describe bien la estructura de autocorrelación espacial, entonces la diferencia entre el valor observado y el valor predicho debe ser pequeña y se podrá producir el mapa. Este procedimiento se realiza en forma secuencial con cada uno de los puntos muestrales y así se obtiene un conjunto de  $n$  errores de predicción.

Sea  $\hat{Z}_{[i]}(\mathbf{s}_i)$  el valor predicho a partir de la validación cruzada, y sea  $\hat{\sigma}_{[i]}(\mathbf{s}_i)$  la predicción para la desviación estándar en la localización  $\mathbf{s}_i$ , con estos estadísticos se construyen los siguientes contrastes de bondad de ajuste:

- i. La media de los errores de predicción (MPE) debe ser igual a cero.

$$\text{MPE} = \frac{\sum_{i=1}^n \left( \hat{Z}_{[i]}(\mathbf{s}_i) - Z(\mathbf{s}_i) \right)}{n}$$

- ii. La raíz media del cuadrado de los errores de predicción (RMSPE).

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^n \left( \hat{Z}_{[i]}(\mathbf{s}_i) - Z(\mathbf{s}_i) \right)^2}{n}}$$

iii. Error estándar promedio (ASE).

$$\text{ASE} = \frac{\sum_{i=1}^n \hat{\sigma}_{[i]}(\mathbf{s}_i)}{n}$$

iv. La media estandarizada de los errores de predicción (MSPE) debe ser muy cercana a cero.

$$\text{MSPE} = \frac{\sum_{i=1}^n \left( \left( \hat{Z}_{[i]}(\mathbf{s}_i) - Z(\mathbf{s}_i) \right) / \hat{\sigma}_{[i]}(\mathbf{s}_i) \right)^2}{n}$$

v. La raíz media estandarizada del cuadrado del error de predicción (RMSSPE) debe ser muy cercana a uno 1.

$$\text{RMSSPE} = \sqrt{\frac{\sum_{i=1}^n \left( \left( \hat{Z}_{[i]}(\mathbf{s}_i) - Z(\mathbf{s}_i) \right) / \hat{\sigma}_{[i]}(\mathbf{s}_i) \right)^2}{n}}$$

vi. El coeficiente de determinación esta dado por:

$$R^2 = 1 - \frac{\sum_{i=1}^n \left( \hat{Z}_{[i]}(\mathbf{s}_i) - Z(\mathbf{s}_i) \right)^2}{\sum_{i=1}^n \left( Z(\mathbf{s}_i) - \bar{Z} \right)^2}$$

$$\text{donde } \bar{Z} = \frac{\sum_{j=1}^n Z(\mathbf{s}_j)}{n}$$

Este último se encuentra en Bivand et al. (2008). Las demás funciones se encuentran en Johnston et al. (2001). Luego se selecciona el método de interpolación que mejores resultados estadísticos arroje, así; el RMSPE se espera sea muy cercano al ASE, en la medida en que esto suceda el modelo presentará un buen ajuste. El criterio al que se da mayor peso es al v., seguido del criterio iv.

Para un modelo que provea predicciones exactas, el MPE debería ser cercano a cero, el RMSPE y el ASE deberían ser tan pequeños como sea posible (lo cual es recomendable cuando se comparan modelos), y el RMSSPE debería ser cercana a 1. El término error de predicción es usado para las diferencias entre la predicción y el valor observado actual. Para un modelo que provea predicciones exactas, el MPE debería ser cercano a 0 si las predicciones son



insegadas, el RMSSPE debería ser cercano a 1 si los errores estándar son exactos y el RMSPE debería ser pequeño si las predicciones son cercanas a los valores observados.

El proceso anterior de eliminar un caso y ajustar con los restantes es un caso particular del método Jackknife. Si se dispone de suficientes datos espacialmente distribuidos de forma homogénea del proceso espacial  $Z(\mathbf{s})$ , se puede entonces dividir los datos en dos subconjuntos, uno de modelización y otro de validación, sobre el que se analizara las diferencias entre los valores observados y las predicciones obtenidas mediante el modelo ajustado con el primer subconjunto.

## Capítulo 2

# Conceptos básicos del análisis espacio-temporal, de distancias y funciones de bases radial

### 2.1 Introducción

En los estudios geoestadísticos se analizan procesos de la naturaleza que se desarrollan en el espacio, pero para los que suele ser habitual disponer de un determinado seguimiento temporal. Por ejemplo, si estamos estudiando la temperatura terrestre en una determinada ciudad, dispondremos de un conjunto de estaciones de control distribuidas espacialmente por la ciudad en las que se mide la temperatura, y para las que sería habitual disponer de un seguimiento temporal de mediciones en cada una de ellas. Como se verá en el capítulo 5, en la modelización y predicción de un determinado fenómeno se obtienen importantes beneficios si, en lugar de considerar únicamente su distribución espacial para un determinado instante temporal de interés (proceso meramente espacial) o la evolución temporal del proceso sobre una localización determinada (proceso meramente temporal), se considera la evolución conjunta del proceso en el espacio-tiempo.

En los últimos años esta área ha tomado gran relevancia y viene ex-

pandiéndose significativamente, en particular en aplicaciones climatológicas y en ciencias ambientales (Kyriakidis & Journel 1999, Kolovos et al. 2004, Le & Zidek 2006, Finkenstädt et al. 2006, Gaetan & Guyon 2010, Cressie & Wikle 2011).

La utilización de modelos estadísticos en el estudio de fenómenos espacio-temporales lleva asociado un conjunto de problemas inherentes a la naturaleza del problema analizado. En particular, se necesita manipular grandes conjuntos de datos obtenidos con una elevada resolución espacial y periodicidad temporal, lo que hará indispensable la utilización de métodos computacionales intensivos para su manejo. También será necesario disponer de procedimientos estadísticos suficientemente flexibles que permitan el ajuste de las diferentes situaciones analizadas.

La geoestadística espacio-temporal provee un cuadro probabilístico para el análisis y predicción de los datos, el cual es construido a partir de la dependencia espacio temporal entre las observaciones (Kyriakidis & Journel 1999). El análisis puede ser enfocado en la interpolación espacial sobre específicos instantes del tiempo. En este caso el objetivo puede centrarse en la comparación de diferentes mapas pronóstico de variables tales como la precipitación, contaminación, nivel de radiación solar, entre otras. Más aun el análisis puede estar enfocado sobre el modelamiento múltiple de series de tiempo donde cada una de las asociaciones espaciales está asociada a una determinada serie de tiempo. Actualmente, la teoría geoestadística de predicción de superficies espaciales, incluye la dimensión asociada al tiempo.

Por otro lado, muchos métodos de estadística y análisis de datos utilizan el concepto geométrico de distancia entre individuos o poblaciones. Las distancias, aparecen en muchos aspectos de la estadística: contraste de hipótesis, estimación, regresión, análisis discriminante, etc. (Arenas & Cuadras 2002). Cuadras & Arenas (1990) proponen el método de regresión múltiple basado en el análisis de distancias utilizando diferentes métricas para el trabajo con variables explicativas continuas y categóricas. De por si los métodos geoestadísticos y espacio-temporales se basan en el cálculo de distancias geométricas, en particular distancias euclídeas espaciales o espacio temporales, de aquí el interés

de considerar también los métodos basados en distancias ya que tienen elementos en común, como lo es el cálculo de las distancias entre las observaciones, esto anidado a la información que aporta el variograma serán determinantes en la generación de pronósticos y permitirá mejorar el poder predictivo de los métodos kriging tradicionales como se verán en los capítulos posteriores.

Adicionalmente, las funciones de base radial (RBF) tales como la multicuadrática (MQ) o completamente regularizada spline (CRS) son útiles en la construcción de modelos digitales de elevación (DEM), como se muestra en (Mitášová & Hofierka 1993). Una variación de la función MQ se llama la inversa multicuadrática (IMQ), introducido por (Hardy & Gopfert 1975). En particular, en Späh (1969) se describe un método que permite evitar puntos de inflexión y contiene splines cúbicos como un caso especial, utilizando interpolación spline cúbica y exponencial (EXP). Más tarde, el spline capa delgada (TPS) se introdujo en el diseño geométrico por (Duchon 1976), y la aproximación de Gauss (GAU) utilizada por (Schagen 1979) es una variante popular del TPS. Por último, (Mitáš & Mitášová 1988, Mitášová & Hofierka 1993, Mitášová & Mitáš 1993) desarrollan la formulación de la spline con tensión (spline with tensión, ST), e implementan un algoritmo de segmentación con un tamaño flexible de la superposición del vecindario. En este punto, en esta tesis se proponen métodos para interpolación espacial y espacio-temporal basada en distancias con funciones de base radial, los cuales se aplican en el modelo geoestadístico para predecir la tendencia y estimar la estructura de covarianza cuando las variables explicativas son mixtas utilizando la distancia de Gower (1971).

Por lo tanto, en este capítulo se presentan algunos conceptos básicos sobre análisis espacio-temporal, de distancias y funciones de base radial, que son útiles para el desarrollo de las metodologías propuestas en cada uno de los siguientes capítulos. Por lo tanto, en las Secciones 2.2 a 2.6 se describe brevemente los principales conceptos, se hace estimación del variograma, se presentan los principales modelos de covarianza considerados en esta tesis, y se presentan los principales resultados de modelización y predicción en el análisis de datos espacio-temporales. En la Sección 2.7 se presentan algunos conceptos

básicos sobre distancias Euclidianas muy útiles para cuando se tienen variables explicativas continuas, categóricas, binarias, e inclusive una mezcla de todas las anteriores. Finalmente, en la Sección 2.8 se concluye este capítulo con una pequeña revisión sobre funciones de base radial espacial que son extendibles a funciones de base radial espacio-temporales y se utilizan en los siguientes capítulos de este trabajo.

## 2.2 Geoestadística espacio-temporal

El objetivo de esta sección es introducir aquellos conceptos más relevantes en el análisis de procesos estocásticos espacio-temporales. Muchos de los conceptos tratados serán generalizaciones del caso espacial que se han visto en el Capítulo 1.

Sea  $\{Z(\mathbf{s}, t), \mathbf{s} \in D \subseteq \mathbb{R}^d, t \in T\}$  un proceso estocástico espacio-temporal observado en  $n$  coordenadas espacio-tiempo  $(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)$ , donde el índice espacial varía en  $D \in \mathbb{R}^d$  y el índice temporal en  $T \in \mathbb{R}$  o  $\mathbb{Z}$ .

Como en la práctica generalmente se dispone de una única realización del proceso a estudiar, cualquier inferencia de las leyes que lo rigen requerirá de la admisión de determinadas leyes simplificadoras relacionadas con la regularidad del proceso, como son la estacionariedad, la separabilidad, la simetría completa, que se verán más adelante. Bajo estas hipótesis de regularidad, se podrán analizar juntas todas las observaciones separadas por un vector de distancia espacio-temporal  $(\mathbf{h}, u)$  obteniendo con ello la replicación necesaria para el análisis.

**Definición 2.1.** *El proceso espacio-temporal  $Z(\mathbf{s}, t)$  tiene función de covarianza espacialmente estacionaria si, para cualquier par de localizaciones  $(\mathbf{s}_i, t_i)$  y  $(\mathbf{s}_j, t_j)$  en  $\mathbb{R}^d \times \mathbb{R}$ , la covarianza  $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$  depende únicamente de la distancia entre las localizaciones espaciales  $\mathbf{s}_i - \mathbf{s}_j$  y de los tiempos  $t_i - t_j$ ,  $i, j = 1, \dots, n$ .*

**Definición 2.2.** *El proceso espacio-temporal  $Z(\mathbf{s}, t)$  tiene función de covarianza temporalmente estacionaria si, para cualquier par de localizaciones*

$(\mathbf{s}_i, t_i)$  y  $(\mathbf{s}_j, t_j)$  en  $\mathbb{R}^d \times \mathbb{R}$ , la covarianza  $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$  depende únicamente de los tiempos  $t_i - t_j$  y de localizaciones espaciales  $\mathbf{s}_i - \mathbf{s}_j$ ,  $i, j = 1, \dots, n$ .

**Definición 2.3.** Si el proceso espacio-temporal  $Z(\mathbf{s}, t)$  tiene función de covarianza estacionaria tanto espacial como temporal, entonces decimos que tiene función de covarianza estacionaria. En ese caso, la función de covarianza puede expresarse como

$$C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = C(\mathbf{h}, u)$$

siendo  $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$  y  $u = t_i - t_j$  las distancias espacial y temporal, respectivamente.

Si un proceso tiene una función de covarianza estacionaria, entonces su varianza no depende de la localización, ya que

$$\text{Var}(Z(\mathbf{s}, t)) = C(\mathbf{0}, 0) = \sigma^2, \quad \forall (\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}$$

donde  $C(\mathbf{0}, 0) \geq 0$  recibe el nombre de varianza a priori del proceso.

**Definición 2.4.** Decimos que un proceso espacio-temporal  $Z(\mathbf{s}, t)$  es estacionario de segundo orden (o estacionario en sentido débil, o simplemente estacionario) si tiene media constante y función de covarianza estacionaria.

**Definición 2.5.** Se dice que el proceso espacio-temporal  $Z(\mathbf{s}, t)$  es un proceso Gaussiano si para cualquier conjunto de localizaciones espacio-temporales  $(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)$ , el vector aleatorio  $\mathbf{Z}_{st} = (Z(\mathbf{s}_1, t_1), \dots, Z(\mathbf{s}_n, t_n))'$  sigue una distribución normal multivariante.

**Definición 2.6.** Un proceso espacio-temporal  $Z(\mathbf{s}, t)$  tiene función de covarianza separable si existe una función de covarianza puramente espacial  $C_s(\cdot)$  y una función de covarianza puramente temporal  $C_t(\cdot)$  tales que (Gneiting et al. 2005)

$$C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = C_s(\mathbf{s}_i, \mathbf{s}_j)C_t(t_i, t_j), \quad i, j = 1, \dots, n$$

para cualquier par de localizaciones  $(\mathbf{s}_i, t_i)$  y  $(\mathbf{s}_j, t_j) \in \mathbb{R}^d \times \mathbb{R}$ . Si esta descomposición no es posible, se dice que la función de covarianza es no separable.

**Definición 2.7.** *Un proceso espacio-temporal  $Z(\mathbf{s}, t)$  tiene función de covarianza completamente simétrica si*

$$C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = C((\mathbf{s}_i, t_j), (\mathbf{s}_j, t_i)) \quad i, j = 1, \dots, n$$

para cualquier par de localizaciones  $(\mathbf{s}_i, t_i)$  y  $(\mathbf{s}_j, t_j) \in \mathbb{R}^d \times \mathbb{R}$ .

**Definición 2.8.** *Un proceso espacio-temporal  $Z(\mathbf{s}, t)$  tiene función de covarianza con soporte compacto si, para cualquier par de localizaciones  $(\mathbf{s}_i, t_i)$  y  $(\mathbf{s}_j, t_j) \in \mathbb{R}^d \times \mathbb{R}$ , la covarianza  $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$  tiende a cero cuando la distancia espacial o temporal es suficientemente grande.*

El esquema de la Figura 2.1, tomada de (Gneiting et al. 2005), representa la relación existente entre las covarianzas separables, completamente simétricas, estacionarias y de soporte compacto, dentro del conjunto general de funciones de covarianza espacio-temporales (estacionarias o no estacionarias). Como vemos, una función de covarianza separable puede ser estacionaria o no estacionaria, y lo mismo ocurre con las funciones de covarianza completamente simétricas. También se observa que las estructuras de covarianza que no son completamente simétricas son no separables.

**Definición 2.9.** *Un proceso espacio-temporal estacionario  $Z(\mathbf{s}, t)$  tiene función de covarianza espacialmente isotrópica si*

$$C(\mathbf{h}, u) = C(\|\mathbf{h}\|, u), \quad \forall(\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}$$

**Definición 2.10.** *Un proceso espacio-temporal estacionario  $Z(\mathbf{s}, t)$  tiene función de covarianza temporalmente isotrópica (o simétrica) si*

$$C(\mathbf{h}, u) = C(\mathbf{h}, |u|), \quad \forall(\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}$$

Obsérvese que si la función de covarianza de un proceso estacionario es espacial y temporalmente isotrópica, entonces es completamente simétrica.

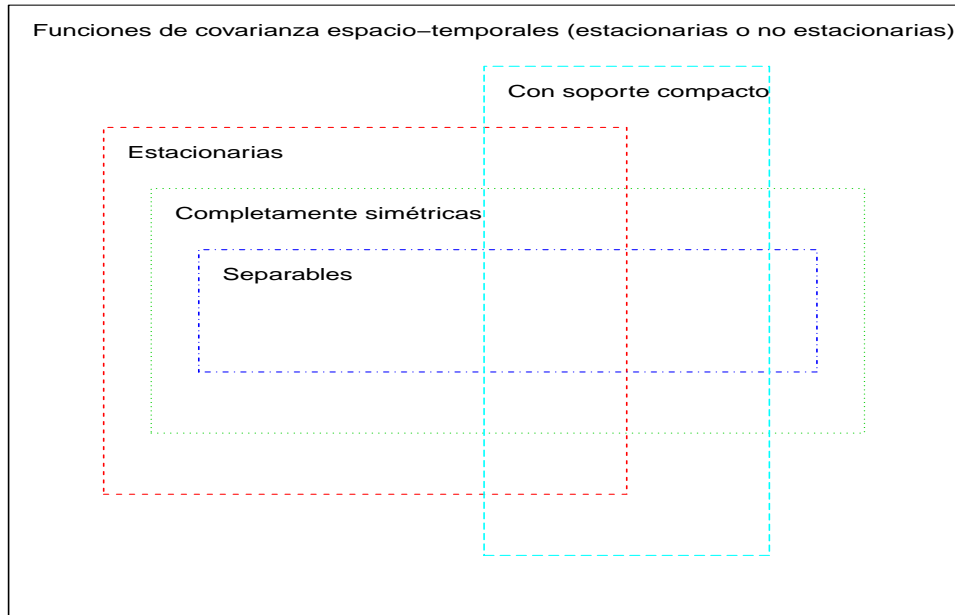


FIGURA 2.1: Relaciones entre los diferentes tipos de funciones de covarianza espacio-temporales

Como ocurría con los procesos espaciales, en ocasiones se modeliza la estructura de segundo orden de un proceso espacio-temporal utilizando variogramas en lugar de funciones de covarianzas. Se define el variograma espacio-temporal como la función

$$2\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = \text{Var}(Z(\mathbf{s}_i, t_i) - Z(\mathbf{s}_j, t_j)), \quad i, j = 1, \dots, n$$

mientras que la mitad de esta cantidad recibe el nombre de semivariograma. En el caso en que el proceso tenga media constante, entonces

$$2\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = \text{E}[Z(\mathbf{s}_i, t_i) - Z(\mathbf{s}_j, t_j)]^2$$

Siempre que sea posible definir la función de covarianza y el variograma, se relacionarán mediante la siguiente expresión

$$\begin{aligned} 2\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) &= \text{Var}(Z(\mathbf{s}_i, t_i)) + \text{Var}(Z(\mathbf{s}_j, t_j)) - 2C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) \\ &= 2C(\mathbf{0}, 0) - 2C(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) \end{aligned}$$



por lo que el proceso  $Z(\mathbf{s}, t)$  sería intrínsecamente estacionario con semivariograma

$$\gamma(\mathbf{h}, u) = C(\mathbf{0}, 0) - C(\mathbf{h}, u)$$

Otra función muy utilizada en la modelización de la dependencia espacio-temporal de los procesos estacionarios es la función de correlación.

**Definición 2.11.** *Sea  $Z(\mathbf{s}, t)$  un proceso estacionario de segundo orden con varianza a priori  $\sigma^2 = C(\mathbf{0}, 0) > 0$ , se define la función de correlación del proceso como*

$$\rho(\mathbf{h}, u) = \frac{C(\mathbf{h}, u)}{C(\mathbf{0}, 0)}$$

Es evidente que si  $\rho(\mathbf{h}, u)$  es una función de correlación sobre  $\mathbb{R}^d \times \mathbb{R}$ , entonces sus marginales  $\rho(\mathbf{0}, u)$  y  $\rho(\mathbf{h}, 0)$  serán, respectivamente, funciones de correlación espacial sobre  $\mathbb{R}^d$  y temporal sobre  $\mathbb{R}$ .

Una condición necesaria y suficiente para que una función  $C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$  de valores reales definida sobre  $\mathbb{R}^d \times \mathbb{R}$  sea una función de covarianza es que sea simétrica y definida positiva, esto es,

$$\sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j C((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) \geq 0 \quad (2.1)$$

para cualquier  $n \in \mathbb{N}$ , y para cualesquiera  $(\mathbf{s}_i, t_i) \in \mathbb{R}^d \times \mathbb{R}$  y  $\varphi_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ .

Análogamente, una condición necesaria y suficiente para que una función de valores reales  $\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j))$  no negativa definida sobre  $\mathbb{R}^d \times \mathbb{R}$  sea un semivariograma es que sea una función simétrica  $\gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) = \gamma((\mathbf{s}_j, t_j), (\mathbf{s}_i, t_i))$  y que sea condicionalmente definida negativa, esto es,

$$\sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \gamma((\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)) \leq 0$$

para cualquier  $n \in \mathbb{N}$ , y cualesquiera  $(\mathbf{s}_i, t_i) \in \mathbb{R}^d \times \mathbb{R}$  y  $\varphi_i \in \mathbb{R}$ ,  $i = 1, \dots, n$  con  $\sum_{i=1}^n \varphi_i = 0$ .

## 2.3 Estimación del variograma y del covariograma

Para obtener una estimación empírica de la función de covarianza o del variograma de un proceso espacio-temporal se puede generalizar de forma natural los procedimientos vistos en la Sección 1.3 para procesos espaciales. Sea  $Z(\cdot, \cdot)$  un proceso intrínsecamente estacionario observado sobre un conjunto de  $n$  localizaciones espacio-temporales  $\{(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)\}$ , se puede obtener una estimación de su variograma  $2\gamma(\cdot, \cdot)$  (o de su función de covarianza  $C(\cdot, \cdot)$  si el proceso es además estacionario de segundo orden) a partir de los valores observados utilizando el estimador clásico propuesto por Matheron (1971) o el estimador robusto de Cressie & Hawkins (1980).

El estimador clásico obtenido aplicando el método de los momentos, para el variograma del proceso, viene dado por

$$2\hat{\gamma}(\mathbf{h}, u) = \frac{1}{|N(\mathbf{h}, u)|} \sum_{N(\mathbf{h}, u)} (Z(\mathbf{s}_i, t_i) - Z(\mathbf{s}_j, t_j)) \quad (2.2)$$

donde  $N(\mathbf{h}, u) = \{(\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j) : \mathbf{s}_i - \mathbf{s}_j \in T(\mathbf{h}), t_i - t_j \in T(u)\}$  siendo  $T(\mathbf{h})$  una región de tolerancia en  $\mathbb{R}^d$  alrededor de  $\mathbf{h}$  y  $T(u)$  una región de tolerancia en  $\mathbb{R}$  alrededor de  $u$ , y  $|N(\mathbf{h}, u)|$  el número de elementos distintos en  $N(\mathbf{h}, u)$ .

Para la covarianza, el estimador obtenido por el método de los momentos sería

$$\hat{C}(\mathbf{h}, u) = |N(\mathbf{h}, u)| \sum_{N(\mathbf{h}, u)} \left( Z(\mathbf{s}_i, t_i) \hat{Z} \right) \left( Z(\mathbf{s}_j, t_j) - \hat{Z} \right)$$

donde  $\hat{Z} = \frac{1}{n} \sum_{i=1}^n Z(\mathbf{s}_i, t_i)$  es un estimador de la media del proceso.

Aunque el método de estimación clásico presenta la ventaja de su facilidad de cálculo, tiene algunos inconvenientes prácticos, como la falta de robustez frente a valores extremos de  $Z(\mathbf{s})$ . Para evitar este problema, Cressie & Hawkins (1980) definen el estimador del variograma siguiente

$$2\hat{\gamma}(\mathbf{h}, u) = \left( \frac{1}{|N(\mathbf{h}, u)|} \sum_{N(\mathbf{h}, u)} |Z(\mathbf{s}_i, t_i) - Z(\mathbf{s}_j, t_j)|^{1/2} \right)^4 \left( 0.457 + \frac{0.494}{|N(\mathbf{h}, u)|} \right)^{-1}$$

Como ocurría en el caso espacial, estos estimadores de la función de covarianza o del variograma del proceso no satisfacen en general la condición de ser definidos positivos o condicionalmente definidas negativas, respectivamente. Por ello, en la práctica se selecciona un modelo paramétrico o no paramétrico de funciones de covarianza o de variogramas que dan lugar a modelos válidos.

## 2.4 Modelos de covarianza espacio-temporales

Algunos modelos de covarianza espacio-temporales que se pueden emplear aquí se definen a continuación.

### 2.4.1 Modelo métrico

Sea  $C(\cdot)$  una función definida positiva en  $\mathbb{R}^d \times \mathbb{R}$ , el modelo métrico esta como (Dimitrakopoulos & Luo 1994)

$$C(\mathbf{h}, u) = C(a^2\|\mathbf{h}\|^2 + b^2|u|^2)$$

donde  $a, b \in \mathbb{R}$  son constantes que definen la métrica espacio-temporal. Este modelo supone la misma estructura de la dependencia en el espacio y el tiempo, y sólo permite cambios en el rango de las dos funciones de covarianza. Se pueden encontrar aplicaciones de este modelo en Armstrong & Hubert (1993) y Snepvangers & Huisman (2003), estos últimos autores lo aplican para modelar el contenido de agua del suelo en un pastizal de 0.36 hectáreas en los Países Bajos.

### 2.4.2 Modelo producto

Se basa en la consideración de las funciones de covarianza espacio-temporales separables estableciendo la dependencia en las dos (Rodriguez-Iturbe & Mejia 1974, De Cesare et al. 1997). Este modelo es

$$C(\mathbf{h}, u) = C_s(\mathbf{h})C_t(u)$$

donde  $C_s$  es una función definida-positiva en  $\mathbb{R}^d$  y  $C_t$  es una función definida-positiva en  $\mathbb{R}$ . Algunos modelos de covarianza espaciales y temporales se encuentran disponibles en Cressie (1993). Si se reescribe en términos de variogramas se tiene

$$\gamma(\mathbf{h}, u) = C_t(0)\gamma_s(\mathbf{h}) + C_s(\mathbf{0})\gamma_t(u) - \gamma_s(\mathbf{h})\gamma_t(u)$$

donde  $\gamma(\mathbf{h}, u)$  es el variograma espacio-temporal,  $\gamma_t(u)$  es el variograma temporal,  $\gamma_s(\mathbf{h})$  es el variograma espacial,  $C_s(\mathbf{0})$  es la meseta de  $\gamma_s$  y  $C_t(0)$  es la meseta de  $\gamma_t$ . La principal ventaja de escribir el producto en términos de modelo de variograma es que, aunque la suma de dos variogramas es generalmente semidefinida, y el producto de dos de esas funciones no será del mismo tipo, cuando la suma y el producto se combinan se puede obtener un modelo válido (Myers et al. 2002).

### 2.4.3 Modelo suma

Este modelo de covarianza estacionaria fue introducido por Rouhani & Hall (1989) y consiste en considerar la covarianza espacio-temporal del proceso como la suma de las covarianzas espaciales y temporales. Este modelo esta dado por

$$C(\mathbf{h}, u) = C_s(\mathbf{h}) + C_t(u)$$

donde  $C_s(\mathbf{h})$  es la función de covarianza espacial definida en  $\mathbb{R}^d$  y  $C_t(u)$  la función de covarianza temporal definida en  $\mathbb{R}$ . El principal problema que se presenta es que la suma de los modelos de covarianza espacial y temporal, no es generalmente una función definida positiva, tan sólo semidefinida positiva. Esto puede causar que la matriz de coeficientes de las ecuaciones de kriging no sea invertible para determinadas configuraciones de algunos datos espacio-temporales (Myers & Journel 1990, Rouhani & Myers 1990).

### 2.4.4 Modelo producto-suma

De Cesare et al. (2001b) introducen este nuevo modelo de covarianza espacio-temporal estacionario, el cual es una generalización del modelo producto, que

se obtiene mediante la combinación de sumas y productos de la covarianza puramente espacial y temporal del proceso, entonces

$$C(\mathbf{h}, u) = k_1 C_s(\mathbf{h}) C_t(u) + k_2 C_s(\mathbf{h}) + k_3 C_t(u)$$

es una covarianza espacio-temporal para algún  $k_1 > 0$  y  $k_2 \geq 0$ ,  $k_3 \geq 0$ . En términos del variograma es

$$\gamma(\mathbf{h}, u) = [k_2 + k_1 C_t(0)] \gamma_s(\mathbf{h}) + [k_3 + k_1 C_s(\mathbf{0})] \gamma_t(u) + k_1 \gamma_s(\delta_s) \gamma_t(\delta_t)$$

donde  $\gamma_s(\mathbf{h})$  y  $\gamma_t(u)$  son las funciones de variograma correspondientes, y  $C_s(\mathbf{h})$  y  $C_t(u)$  son las correspondientes funciones de covarianza, aquí  $C(\mathbf{0}, 0)$  es la meseta de  $\gamma(\mathbf{h}, t)$ ,  $C_s(\mathbf{0})$  es la meseta de  $\gamma_s(\mathbf{h})$  y  $C_t(0)$  es la meseta de  $\gamma_t(u)$ . Por definición, se sabe que  $\gamma(\mathbf{0}, 0) = \gamma_s(\mathbf{0}) = \gamma_t(0) = 0$ . En este caso, los variogramas marginales son

$$\gamma(\mathbf{h}, 0) = [k_2 + k_1 C_t(0)] \gamma_s(\mathbf{h}) \quad \text{y} \quad \gamma(\mathbf{0}, u) = [k_3 + k_1 C_s(\mathbf{0})] \gamma_t(u)$$

es decir, son iguales a los variogramas puramente espaciales y temporales, respectivamente, excepto por una constante de proporcionalidad.

### 2.4.5 Modelo Cressie-Huang

Un procedimiento para construir modelos paramétricos a partir de funciones de covarianza estacionarias no separables espacio-temporales, y por lo tanto que incluye la posible interacción espacio-tiempo del proceso, se ha presentado en Cressie & Huang (1999). Si  $C(\mathbf{h}, u)$  es integrable se puede expresar la función de covarianza en la forma

$$C(\mathbf{h}, u) = \int_{\mathbb{R}^k} e^{i\mathbf{h}'\boldsymbol{\omega}} \rho(\boldsymbol{\omega}, u) k(\boldsymbol{\omega}) d\boldsymbol{\omega},$$

donde  $k(\boldsymbol{\omega})$  es la densidad espectral de un proceso espacial y  $\rho(\boldsymbol{\omega}; \cdot)$  una función de autocorrelación temporal, la cual asume los siguientes requisitos:

- i. Para cada  $\boldsymbol{\omega} \in \mathbb{R}^d$ ,  $\rho(\boldsymbol{\omega}, \cdot)$  es una función de autocorrelación continua,  $\int \rho(\boldsymbol{\omega}, u) du < \infty$  y  $k(\boldsymbol{\omega}) > 0$ ,
- ii. la función positiva  $k(\boldsymbol{\omega})$  satisface:

$$\int k(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty.$$

## 2.5 Modelización de procesos espacio-temporales

En la modelización práctica, el investigador debe tomar la importante decisión de cuál es el modelo espacio-temporal que mejor ajusta sus datos empíricos. En Kyriakidis & Journel (1999), Banerjee et al. (2004) y Chen (2007) se puede encontrar una revisión completa y exhaustiva de las diferentes técnicas de modelización de procesos espacio-temporales.

El proceso espacio-temporal  $Z(\mathbf{s}_i, t_i)$  suele descomponerse como

$$Z(\mathbf{s}_i, t_i) = \mu(\mathbf{s}_i, t_i) + \varepsilon(\mathbf{s}_i, t_i) \quad (\mathbf{s}_i, t_i) \in \mathbb{R}^d \times \mathbb{R} \quad (2.3)$$

con  $i = 1, \dots, n$  y donde  $\mu(\mathbf{s}_i, t_i) = \mathbb{E}[Z(\mathbf{s}_i, t_i)]$  y  $\varepsilon(\mathbf{s}_i, t_i)$  es un proceso estocástico de media cero y variograma  $2\gamma(\cdot, \cdot)$ . Este proceso caracteriza la dependencia espacio-temporal y modeliza las fluctuaciones espacio-temporales de  $Z(\mathbf{s}_i, t_i)$  alrededor de su media  $\mu(\mathbf{s}_i, t_i)$ .

Generalmente se considera una media determinista para el proceso que se descompone como

$$\mu(\mathbf{s}_i, t_i) = \sum_{k=0}^p \theta_k f_k(\mathbf{s}_i, t_i), \quad (\mathbf{s}_i, t_i) \in \mathbb{R}^d \times \mathbb{R} \quad (2.4)$$

con lo que  $\mu(\mathbf{s}_i, t_i)$  se expresa como función de  $p + 1$  funciones conocidas  $f_k(\mathbf{s}_i, t_i)$  incluyendo el intercepto  $f_0(\mathbf{s}_i, t_i) = 1$ , estas funciones son elegidas durante la modelización para ajustar la media observada de los datos, y donde  $\theta_k$  son  $p + 1$  coeficientes desconocidos. Por ejemplo, podemos considerar funciones periódicas que dependen de  $t$  para capturar variaciones periódicas sobre el eje temporal, y funciones polinomiales o continuas a trozos que dependen de  $\mathbf{s}$  para modelizar variaciones suaves o discontinuidades en el espacio. Los coeficientes de (2.4) pueden considerarse fijos, o dependientes del espacio y del tiempo,  $\theta_k = \theta_k(\mathbf{s}, t)$ .

## 2.6 Predicción de procesos espacio-temporales

Se supone a lo largo de toda esta sección que se ha observado el valor del proceso sobre un conjunto de  $n$  localizaciones espacio-temporales  $\{(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)\}$ . Generalmente el objetivo será la predicción del valor del proceso sobre una nueva localización  $(\mathbf{s}_0, t_0)$ ,  $Z(\mathbf{s}_0, t_0)$ , donde  $\mathbf{s}_0$  será una determinada localización espacial en  $\mathbb{R}^d$  y  $t_0$  será un tiempo de interés en  $\mathbb{R}$ . Se supone que el proceso  $Z(\cdot, \cdot)$  satisface la condición de regularidad,  $\text{Var}(Z(\mathbf{s}, t)) < \infty, \forall (\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}$ , con lo que tanto su media  $\mu(\mathbf{s}_i, t_i) = E(Z(\mathbf{s}_i, t_i))$  como su función de covarianza  $\text{Cov}(Z(\mathbf{s}_i, t_i), Z(\mathbf{s}_j, t_j))$  existen, para cualesquiera  $(\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j) \in \mathbb{R}^d \times \mathbb{R}$ . Además se supone que el proceso  $Z(\mathbf{s}_i, t_i)$  admite la descomposición (2.4), es decir, puede expresarse como suma de una tendencia  $\mu(\mathbf{s}_i, t)$  que expresa la media del proceso y un proceso estocástico estacionario  $\varepsilon(\mathbf{s}, t)$  de media cero que captura la variabilidad espacio-temporal respecto a dicha media.

En esta sección se considera el predictor  $\hat{Z}(\mathbf{s}_0, t_0)$  definido como combinación lineal de las observaciones de forma que sea insesgado y minimice el error cuadrático medio de la predicción. Como se vio en el caso espacial, existen diferentes modalidades de kriging dependiendo de las hipótesis adoptada sobre la media  $\mu(\mathbf{s}_i, t_i)$  del proceso. Así, en el kriging ordinario se supone que la media es desconocida pero constante y en el kriging universal se supone que la media es desconocida pero puede expresarse como función lineal de un conjunto de variables que dependen de la localización espacio-temporal.

### 2.6.1 Kriging ordinario espacio-temporal

Se supone que la media del proceso  $\mu(\mathbf{s}, t) = \mu$  es una constante desconocida. En este caso, el proceso estacionario  $Z(\mathbf{s}, t)$  tiene función de covarianza  $C(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) = C(Z(\mathbf{s}_i, t_i), Z(\mathbf{s}_j, t_j)) = C(\varepsilon(\mathbf{s}_i, t_i), \varepsilon(\mathbf{s}_j, t_j)) = C_\varepsilon(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j)$ .

En este caso, la condición de insesgades  $E[Z(\mathbf{s}_0, t_0) - \hat{Z}(\mathbf{s}_0, t_0)] = 0$  del predictor impone que los pesos  $\varphi_i$  deben sumar uno ( $\sum_{i=1}^n \varphi_i = 1$ ), y el predictor

asociado a este kriging vendrá dado por

$$\hat{Z}(\mathbf{s}_0, t_0) = \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i, t_i) \quad (2.5)$$

El objetivo es encontrar aquellos pesos  $\varphi_i$  que minimizan el error cuadrático medio  $\sigma_{OK}^2$  asociado a la predicción, y dado por

$$\sigma_{OK}^2(\mathbf{s}_0, t_0) = \text{Var} \left[ Z(\mathbf{s}_0, t_0) - \hat{Z}(\mathbf{s}_0, t_0) \right] = \text{E} \left[ Z(\mathbf{s}_0, t_0) - \hat{Z}(\mathbf{s}_0, t_0) \right]^2 \quad (2.6)$$

bajo la restricción dada por la condición de insesgadez. Aplicando el mismo procedimiento explicado de la Sección 1.6 para el caso espacial, se llega a que el predictor buscado se obtiene considerando

$$\hat{\boldsymbol{\varphi}}' = \left( \mathbf{c} + \mathbf{1} \frac{1 - \mathbf{1}' \Sigma_{\boldsymbol{\vartheta}}^{-1} \mathbf{c}}{\mathbf{1}' \Sigma_{\boldsymbol{\vartheta}}^{-1} \mathbf{1}} \right)' \sigma^{-1} \quad (2.7)$$

siendo  $\Sigma_{\boldsymbol{\vartheta}}$  una matriz  $n \times n$  cuyo  $(i, j)$ -ésimo elemento es  $C(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j)$  y  $\mathbf{c}$  el vector de dimensión  $n$  cuyo  $i$ -ésimo elemento es  $C(\mathbf{s}_0 - \mathbf{s}_i, t_0 - t_i)$ . La varianza del error de predicción estará dada por

$$\hat{\sigma}_{OK}^2(\mathbf{s}_0, t_0) = C(\mathbf{0}, 0) - \hat{\boldsymbol{\varphi}}' \mathbf{c} + \frac{1 - \mathbf{1}' \Sigma_{\boldsymbol{\vartheta}}^{-1} \mathbf{c}}{\mathbf{1}' \Sigma_{\boldsymbol{\vartheta}}^{-1} \mathbf{1}} \quad (2.8)$$

Como en el caso espacial, todas estas expresiones pueden escribirse en términos del variograma del proceso si éste es además intrínsecamente estacionario. Sea  $2\gamma(\mathbf{h}, u)$  el variograma del proceso, los pesos  $\varphi_i$  a considerar estarán dados por

$$\hat{\boldsymbol{\varphi}}' = \left( \gamma + \mathbf{1} \frac{1 - \mathbf{1}' \Gamma^{-1} \mathbf{c}}{\mathbf{1}' \Gamma^{-1} \mathbf{1}} \right)' \Gamma^{-1}$$

siendo  $\Gamma$  la matriz  $n \times n$  cuyo  $(i, j)$ -ésimo elemento es  $\gamma(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j)$  y  $\gamma$  el vector de dimensión  $n$  cuyo  $i$ -ésimo elemento es  $\gamma(\mathbf{s}_0 - \mathbf{s}_i, t_0 - t_i)$ . Si el objetivo es únicamente la predicción del proceso en el punto  $(\mathbf{s}_0, t_0)$ , mediante el procedimiento descrito anteriormente no será necesario estimar la media del proceso. En este caso, el error cuadrático medio de la predicción que se ha minimizado estará dado por la siguiente expresión

$$\hat{\sigma}_{OK}^2(\mathbf{s}_0, t_0) = \gamma' \Gamma^{-1} \gamma - \frac{(\mathbf{1}' \Gamma^{-1} \gamma - 1)^2}{\mathbf{1}' \Gamma^{-1} \mathbf{1}}$$



## 2.6.2 Kriging Universal espacio-temporal

Se supone que la media del proceso  $\mu(\mathbf{s}, t)$ , aunque desconocida, es una combinación lineal de funciones conocidas o covariables ligadas a las localizaciones espacio-temporales. Esto es

$$\mu(\mathbf{s}_i, t_i) = \sum_{k=0}^p \theta_k v_k(\mathbf{s}_i, t_i) = \mathbf{v}'(\mathbf{s}_i, t_i) \tilde{\boldsymbol{\theta}} \quad (2.9)$$

donde  $\mathbf{v}(\mathbf{s}_i, t_i)$  es el vector formado por los valores de las  $p + 1$  variables explicativas incluyendo el intercepto, las cuales son consideradas en la  $i$ -ésima localización espacio-temporal, con  $v_0(\mathbf{s}, t) = 1$ , y  $\tilde{\boldsymbol{\theta}}$  es un vector de  $p + 1$  parámetros desconocidos. Se denotara por  $V$  la matriz  $n \times (p + 1)$  cuyo  $(i, j)$ -ésimo elemento es  $v_j(\mathbf{s}_i, t_i)$ ,  $i = 1, \dots, n$ ,  $j = 0, 1, \dots, p$ .

Como en los casos anteriores, se quiere predecir el valor de  $Z(\mathbf{s}_0, t_0)$  a partir del conjunto de observaciones  $\mathbf{Z}$  utilizando el predictor lineal (2.5), escogiendo los pesos  $\varphi_i$  de forma que sea insesgado y minimice el error cuadrático medio asociado. La condición de insesgaredad establecida anteriormente, implica que  $\boldsymbol{\varphi}'V = \mathbf{v}(\mathbf{s}_0, t_0)'$ . El predictor asociado a este tipo de kriging viene dado por

$$\hat{Z}(\mathbf{s}_0, t_0) = \hat{\boldsymbol{\varphi}}'V$$

donde

$$\hat{\boldsymbol{\varphi}}' = \left[ \mathbf{c} + V (V'\Sigma_{\boldsymbol{\vartheta}}^{-1}V')^{-1} (\mathbf{v}(\mathbf{s}_0, t_0) - V'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{c}) \right]' \Sigma_{\boldsymbol{\vartheta}}^{-1} \quad (2.10)$$

siendo  $\Sigma_{\boldsymbol{\vartheta}}$  la matriz  $n \times n$  cuyo  $(i, j)$ -ésimo elemento es  $C(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j)$  y  $\mathbf{c}$  es el vector de dimensión  $n$  cuyo  $i$ -ésimo elemento es  $C(\mathbf{s}_0 - \mathbf{s}_i, t_0 - t_i)$ . El error cuadrático medio asociado a este predictor viene dado por

$$\hat{\sigma}_{UK}^2(\mathbf{s}_0, t_0) = C(\mathbf{0}, 0) - \mathbf{c}'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{c} + [\mathbf{v}(\mathbf{s}_0, t_0) - V'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{c}]' \times (V'\Sigma_{\boldsymbol{\vartheta}}^{-1}V)^{-1} [\mathbf{v}(\mathbf{s}_0, t_0) - V'\Sigma_{\boldsymbol{\vartheta}}^{-1}\mathbf{c}] \quad (2.11)$$

Como en el caso espacial, todas estas expresiones pueden escribirse en términos del variograma del proceso si este es además intrínsecamente estacionario. Sea  $2\gamma(\mathbf{h}, u)$  el variograma del proceso, los pesos  $\varphi_i$  a considerar estarán dados por

$$\hat{\boldsymbol{\varphi}}' = \left[ \gamma + V (V'\Gamma^{-1}V)^{-1} (\mathbf{v}(\mathbf{s}_0, t_0) - V'\Gamma^{-1}\gamma) \right]' \Gamma^{-1}$$

siendo  $\Gamma$  la matriz  $n \times n$  cuyo  $(i, j)$ -ésimo elemento es  $\gamma(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j)$  y  $\gamma$  el vector de dimensión  $n$  cuyo  $i$ -ésimo elemento es  $\gamma(\mathbf{s}_0 - \mathbf{s}_i, t_0 - t_i)$ . El error cuadrático medio de la predicción esta dado por la siguiente expresión

$$\hat{\sigma}_{UK}^2(\mathbf{s}_0, t_0) = \gamma' \Gamma^{-1} \gamma - [\mathbf{v}(\mathbf{s}_0, t_0) - V' \Gamma^{-1} \gamma]' (V' \Gamma^{-1} V)^{-1} [\mathbf{v}(\mathbf{s}_0, t_0) - V' \Gamma^{-1} \gamma]$$

Respecto a la estimación óptima de los parámetros de la media  $\tilde{\boldsymbol{\theta}}$ , como los datos  $\mathbf{Z}$  satisfacen un modelo lineal general con  $E(\mathbf{Z}) = V \tilde{\boldsymbol{\theta}}$  y  $\text{Var}(\mathbf{Z}) = \Sigma_{\boldsymbol{\theta}}$ , puede obtenerse por mínimos cuadrados generalizados como

$$\hat{\boldsymbol{\theta}}_{gls} = (V' \Sigma_{\boldsymbol{\theta}}^{-1} V)^{-1} V' \Sigma_{\boldsymbol{\theta}}^{-1} \mathbf{Z}$$

donde  $\text{Var}(\hat{\boldsymbol{\theta}}_{gls}) = (V' \Sigma_{\boldsymbol{\theta}}^{-1} V)^{-1}$ . A partir de estas expresiones, se puede obtener intervalos de confianza para estos parámetros o combinaciones lineales de los mismos. Se deduce que aún cuando para estimar los parámetros de la media de forma óptima es necesario conocer las covarianzas involucradas en  $\Sigma_{\boldsymbol{\theta}}$ , para la predicción óptima de  $Z(\mathbf{s}_0, t_0)$  únicamente se ha de conocer los variogramas implicados en  $\Gamma$ .

## 2.7 Regresión basada en distancias

En esta sección se presentan los principales conceptos de distancias y de regresión basada en distancias propuestas por Cuadras (1989), Cuadras & Arenas (1990) y Arenas & Cuadras (2002). Las distancias, aparecen en muchos aspectos de la estadística: contraste de hipótesis, estimación, regresión, análisis discriminante, etc. (Arenas & Cuadras 2002). Cuadras & Arenas (1990) proponen el método de regresión múltiple basado en el análisis de distancias utilizando diferentes métricas para el trabajo con variables explicativas continuas y categóricas. Un resumen de dichas propuestas es presentado a continuación.

### 2.7.1 Distancia y similaridad

**Definición 2.12.** Una distancia  $\delta$  sobre un conjunto (finito o no)  $\Omega$  es una aplicación que a cada par de individuos  $(w_i, w_j) \in \Omega \times \Omega$ , le hace corresponder

un número real  $\delta(w_i, w_j) = \delta_{ij}$ , que cumple con las siguientes propiedades básicas:

- i.  $\delta_{ij} \geq 0$ .
- ii.  $\delta_{ii} = 0$ .
- iii.  $\delta_{ij} = \delta_{ji}$ .
- iv.  $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$ .

Este último denominado desigualdad triangular, si se cumple se dice que la distancia es métrica.

**Definición 2.13.** Si  $\Omega$  es un conjunto finito, que se indica por  $\Omega = \{w_1, w_2, \dots, w_n\}$ , las distancias  $\delta_{ij}$  se expresan mediante la matriz simétrica  $\Delta$ , llamada matriz de distancias sobre  $\Omega$ .

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{pmatrix}$$

con  $\delta_{ii} = 0$ ,  $\delta_{ij} = \delta_{ji}$ . Se llama preordenación de  $\Omega$  asociada a  $\Delta$ , a la ordenación de menor a mayor de los  $q = n \times (n + 1)/2$  pares de distancias no nulas:

$$\delta_{i_1 j_1} \leq \delta_{i_2 j_2} \leq \cdots \leq \delta_{i_q j_q}$$

es decir, la ordenación de los pares  $(w_i, w_j)$  de  $\Omega$  de acuerdo con su proximidad.

Una matriz de distancias  $\Delta$  puede ser transformada de diversos modos. Por ejemplo:

$$\tilde{\delta}_{ij} = \begin{cases} 0 & \text{si } i = j \\ \delta_{ij} + c & \text{si } i \neq j \end{cases} \quad (2.12)$$

La transformación (2.12), consiste en sumar una constante fuera de la diagonal de  $\Delta$ , se llama aditiva. Esta transformación es útil para conseguir que la nueva distancia cumpla propiedades que la distancia original no posee, conservando la preordenación, es decir, la relación de proximidad entre los individuos.

**Definición 2.14.** Una similaridad  $m$  en un conjunto  $\Omega$ , es una aplicación que asigna a cada par  $(w_i, w_j) \in \Omega \times \Omega$  un número real  $m_{ij} = m(i, j)$ , que cumple:

i.  $0 \leq m_{ij} \leq m_{ii} = 1$ .

ii.  $m_{ij} = m_{ji}$ .

Cuando  $\Omega$  es un conjunto finito, entonces la matriz

$$M = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nn} \end{pmatrix}$$

se denomina matriz de similaridades sobre  $\Omega$ .

Es inmediato pasar de similaridad a distancia y recíprocamente. Las dos transformaciones básicas son:

$$\delta_{ij} = 1 - m_{ij} \quad (2.13)$$

y

$$\delta_{ij} = \sqrt{1 - m_{ij}} \quad (2.14)$$

En general una matriz de similaridades puede tener en su diagonal elementos  $s_{ij} \neq 1$ . La transformación que permite pasar de similaridad a distancia es entonces:

$$\delta_{ij} = \sqrt{m_{ii} + m_{jj} - 2m_{ij}} \quad (2.15)$$

Por diversas razones (2.14) es preferible a (2.13). Pero en general, (2.15) es la transformación más apropiada (Cuadras & Arenas 1990, Mardia et al. 2002).

En el caso de contar con variables binarias se pueden obtener similaridades y distancias realizando el siguiente procedimiento: sean  $p$  variables binarias  $V_1, V_2, \dots, V_p$ , donde cada  $V_j$  ( $j = 1, \dots, p$ ) toma los valores 0 ó 1 según la presencia de una cierta característica. Entonces son bien conocidos los siguientes coeficientes de similaridad entre cada par de individuos  $w_i, w_j$ .

$$\begin{aligned} m_{ij} &= \frac{c_1 + c_4}{c_1 + c_2 + c_3 + c_4} && (\text{Sokal - Michener}) \\ m_{ij} &= \frac{c_1}{c_1 + c_2 + c_3} && (\text{Jaccard}) \end{aligned} \quad (2.16)$$

donde  $c_1, c_2, c_3, c_4$  son las frecuencias de  $(1, 1), (1, 0), (0, 1)$  y  $(0, 0)$ , respectivamente. Note que  $p = c_1 + c_2 + c_3 + c_4$ . Estas similitudes pueden ser transformadas en distancias utilizando (2.13) o (2.14).

De otro lado cuando las variables son mixtas: continuas, binarias o cualitativas, entonces es adecuado utilizar la similitud de Gower (1968) y Gower (1971):

$$m_{ij} = \frac{\sum_{l=1}^{p_1} \left( 1 - \frac{|v_{il} - v_{jl}|}{G_l} \right) + c_1 + \omega}{p_1 + (p_2 - c_4) + p_3} \quad (2.17)$$

donde  $G_l$  es el rango de la  $l$ -ésima variable cuantitativa,  $p_1$  es el número de variables cuantitativas,  $c_1$  y  $c_4$  corresponden al número de coincidencias y no coincidencias para las  $p_2$  variables binarias, respectivamente, y  $\omega$  es el número de coincidencias para las  $p_3$  variables cualitativas. Este coeficiente admite la posibilidad de tratar datos faltantes y se reduce al coeficiente de Jaccard cuando  $p_1 = p_2 = 0$ . Además, en este caso, se utiliza la distancia (2.14) elevada al cuadrado, es decir  $\delta_{ij}^2 = 1 - m_{ij}$ .

Una vez se utiliza alguna de las anteriores distancias según los intereses del investigador, se realiza la descomposición espectral con la finalidad de realizar el modelo de regresión basado en distancias. En este sentido, sea  $A_{n \times n} = (a_{ij})$  la matriz con elementos  $a_{ij} = -\delta_{ij}^2/2$ , y sea  $B = HAH$  donde  $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}'$  es la matriz centrada, con  $I$  una matriz identidad  $n \times n$  y  $\mathbf{1}$  un vector de unos  $n \times 1$ . Además,  $B$  es una matriz semidefinida positiva (Mardia et al. 2002) de rango  $k$ , entonces la matriz  $X$  de coordenadas principales se puede obtener a partir de la siguiente descomposición espectral

$$B = HAH = U\Lambda U' = XX' \quad (2.18)$$

donde  $\Lambda$  es una matriz diagonal conformada por los valores propios de  $A$ ,  $X = U\Lambda^{1/2}$  es una matriz  $n \times n$  de rango  $k \leq n - 1$ , y  $U$  contiene la coordenadas estandarizadas. La matriz  $B$  proporciona las coordenadas euclídeas del conjunto  $\Omega = \{w_1, w_2, \dots, w_n\}$ . Cada fila  $x_i$  de  $X$  contiene las coordenadas, llamadas coordenadas principales del individuo  $i$ .

## 2.7.2 Modelo de regresión basado en distancias

Suponiendo que se tienen  $p$  variables  $V_1, V_2, \dots, V_p$  observables de tipo continuo, binario o categórico o incluso los tres tipos a la vez, en cuyo caso se dirá que los datos son mixtos. Sea  $\delta(w_i, w_j)$  una distancia adecuada entre pares  $w_i$  y  $w_j$  de individuos. Si los datos son binarios  $\delta(w_i, w_j)$  se puede basar en (2.16) y si son mixtos en el coeficiente de similaridad de Gower (2.17). A partir de  $\delta(w_i, w_j)$  se puede obtener la matriz  $n \times n$  de distancias  $\Delta$  y aplicando la descomposición espectral (2.18), se obtiene la matriz  $X$ , de coordenadas principales, que reproducen las distancias originales. El modelo que se propone en Cuadras (2007), es entonces

$$\mathbf{Z} = \beta_0 \mathbf{1} + X\beta + e \quad (2.19)$$

donde  $\mathbf{1}$  es el vector de unos de  $n \times 1$ , mientras que  $\mathbf{Z}_{(n \times 1)}$  es un vector conocido con  $n$  observaciones de una variable respuesta cuantitativa,  $X_{(n \times k)}$  es conocida de  $\text{rang}(X) = \text{rang}(B) = k$ ,  $\beta_{(k \times 1)}$  es un vector desconocido de parámetros y  $e$  es un vector aleatorio. Obsérvese que como  $B\mathbf{1} = 0$ , y tanto  $\mathbf{1}$  como las columnas  $X_1, X_2, \dots, X_k$  de  $X$ , son vectores propios de  $B$ .

El modelo (2.19), se puede también escribir

$$\mathbf{Z} = \beta_0 \mathbf{1} + \sum_{i=1}^k \beta_i X_i + e \quad (2.20)$$

donde  $X_1, X_2, \dots, X_k$  juegan el papel de variables predictoras.

Como valor de  $k$ , se puede tomar  $k$  el número inicial de variables observables explicativas. Una buena selección de las columnas  $X_1, \dots, X_k$  de  $X$  consiste en escogerlas por orden de correlación con  $\mathbf{Z}$ , es decir,

$$r(\mathbf{Z}, X_1) > r(\mathbf{Z}, X_2) > \dots > r(\mathbf{Z}, X_k) \quad (2.21)$$

Otra selección obvia consiste en ordenarlas de acuerdo con la variabilidad explicada por las variables predictoras (columnas de  $X$ ):  $\lambda_1 > \dots > \lambda_k$ , es decir seleccionar los  $k$  primeros ejes principales. Pero si resultara que la variable  $X_{k+1}$  tiene una correlación  $r_{k+1} = r(\mathbf{Z}, X_{k+1})$ , relativamente alta, se

podría haber perdido una variable predictiva importante (véase Cuadras & Fortiana (1993) para una discusión de este problema).

Cuando  $n$  es muy grande, la selección de coordenadas puede volverse en un cálculo muy arduo o imposible. Un procedimiento que requiere solo calcular los primeros  $k$  vectores propios adecuados, es el siguiente. De acuerdo a Cuadras et al. (1996), se define la secuencia:

$$c(0) = 0, \quad c(i) = \frac{\sum_{j=1}^i r_j^2 \lambda_j}{\sum_{j=1}^k r_j^2 \lambda_j} \quad i = 1, \dots, k \quad (2.22)$$

donde  $r_j = \text{Corr}(\mathbf{Z}, X_j)$ . Cada  $c(i)$  mide la predictibilidad de las primeras  $i$  dimensiones, ponderadas por los correspondientes valores propios. El valor inicial  $c(0) = 0$ , puede ser interpretado como la falta de predictibilidad de  $\mathbf{1}$ , el vector constante de unos, que es también un valor propio de  $B$ .

La selección debe ser realizada representando gráficamente los puntos

$$(i, 1 - c(i)) \quad i = 0, 1, \dots, k^* < k$$

donde  $k^*$  es tal que  $1 - c(i)$  esté muy próximo a 0. Esto es, el corte en  $k^*$  es tal que, a la derecha de  $k^*$  el gráfico está muy próximo al eje horizontal, indicando que las dimensiones superiores no deben ser tenidas en cuenta. La dimensión principal  $1 \leq i \leq k^*$  debe ser seleccionada si se aprecia una caída entre el punto  $(i - 1, 1 - c(i - 1))$  y el  $(i, 1 - c(i - 1))$ . Entonces la dimensión  $i$  es aceptada o rechazada según si  $r_i^2$  o  $\lambda_i$  sean grandes o pequeños.

## 2.8 Funciones de base radial

En esta sección se realiza un breve estudio teórico sobre las funciones de base radial aplicadas al problema de interpolación espacial.

**Definición 2.15.** Una función  $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$  se llama radial, si existe una función invariante  $\Phi(s) = [0, \infty)$  tal que:

$$\Phi(\mathbf{s}) = \phi(\delta), \quad \text{donde } \delta = \|\mathbf{s}\|$$

y  $\|\cdot\|$  es la norma Euclidiana en  $\mathbb{R}^d$ . Esto significa que el valor de la función  $\Phi$  en el punto  $\mathbf{s} \in \mathbb{R}^d$  únicamente depende de la norma de  $\mathbf{s}$ .

A continuación se presentan algunas funciones de base radial espaciales que serán utilizadas en los siguientes capítulos, y que son adaptadas directamente a funciones de base radial espacio-temporales.

### 2.8.1 Multicuadrática (MQ)

Hardy (1990) llamo multicuadrático al método porque considero que la característica principal es la de ser una superposición de superficies cuádricas.

**Definición 2.16.** Dado un conjunto de  $n$  puntos distintos  $\{x_i\}_{i=1}^n \in \mathbb{R}^d$  y sus correspondientes valores escalares  $\{f_i\}_{i=1}^n \in \mathbb{R}$ , el interpolador multicuadrático de los datos tiene la siguiente forma:

$$p(\mathbf{s}_i) = \sum_{j=1}^n \mathbf{b}_j \sqrt{\eta_i^2 + \delta_i^2}, \quad j = 1, \dots, n$$

donde los coeficientes  $\mathbf{b}_j$  se determinan mediante la imposición de las condiciones de interpolación  $p(\mathbf{s}_i) = f_i$ , para  $i = 1, \dots, n$ , y  $\eta_j$  es un parámetro de suavizado. De aquí se obtiene el siguiente sistema de ecuaciones lineales y simétricas

$$\Phi \mathbf{b} = \mathbf{f}$$

donde  $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_n)'$  y las entradas de  $\Phi$  están dadas por  $\phi_{ij} = \sqrt{\eta^2 + \delta_{ij}^2}$ .

Particularmente, dado un conjunto de medidas dispersas provenientes de un conjunto de puntos en una superficie topográfica (por ejemplo, medidas de la elevación de una montaña), se construye de forma satisfactoria una función continua que representa la superficie. Por satisfactoria se entiende una función que consigue realizar un ajuste exacto de los datos y proporciona una buena aproximación de las características de la superficie, es decir, localización de las cumbres, llanuras y desembocaduras. Parte de la motivación para construir esta función fue crear un método automático que permitiese generar los mapas del contorno de una superficie cartográfica (Ortega-Pérez 2009).



Además de crear de forma automática y de manera objetiva los mapas del contorno, se obtuvo una motivación matemática para construir una función continua que representase una superficie topográfica, ya que teniendo esta función continua se podrían utilizar métodos analíticos y geométricos que permitiesen determinar volúmenes de tierra, cumbres, llanuras, distancias a lo largo de una curva (Hardy 1971, Hardy 1990).

### 2.8.2 Multicuadrática inversa (IM)

Una variación de la función multicuadrática fue introducida por (Hardy & Gopfert 1975), y esta dada por

$$\phi(\delta) = 1/\sqrt{\eta^2 + \delta^2}$$

donde  $\eta \neq 0$  es un parámetro de suavizado de libre escogencia. Franke (1982) encuentra que esta función de base radial puede proporcionar excelentes aproximaciones, incluso cuando el número de centros (vecinos más cercanos) es pequeño.

### 2.8.3 Spline con tensión (ST)

Esta función esta dada por la expresión

$$\phi(\delta) = \ln(\eta \cdot \delta/2) + K_0(\eta \cdot \delta) + C_E$$

donde  $K_0(x)$  es la función modificada de Bessel (Abramowitz & Stegun 1965, pág. 374) y  $C_E = -\int_0^\infty (\ln(x)/e^x)dx = 0.5772161$  es la constante de Euler (Abramowitz & Stegun 1965, pág. 255).

### 2.8.4 Spline capa delgada (TPS)

Este spline fue introducido en el diseño geométrico por Duchon (1976). El nombre, spline capa delgada, se refiere a una analogía física que implica la flexión de una hoja delgada de metal. Más tarde Thiébaux & Pedder (1987)

describió la TPS como un spline cúbico de 2 dimensiones (superficie). En el caso de un espacio Euclidiano con  $d = 2$ , la TPS tendrá la siguiente forma:

$$\phi(\delta) = \begin{cases} (\eta \cdot \delta)^2 \log(\eta \cdot \delta) & \text{si } \delta \neq 0 \text{ y } \eta > 0 \\ 0 & \text{si } \delta = 0 \end{cases}$$

Franke (1982) desarrolla un programa informático para la solución del problema de interpolación de datos dispersos. El algoritmo se basa en una suma ponderada de splines capa delgada definidos localmente y se obtiene una función de interpolación que es diferenciable.

### 2.8.5 Completamente regularizada spline (CRS)

Una variante de la TPS que usa la función base spline regularizada, se denomina CRS y viene dada por

$$\phi(\delta) = -\sum_{k=1}^{\infty} \frac{(-1)^k (\eta \cdot \delta)^{2k}}{k!k} = \ln(\eta \cdot \delta/2)^2 + E_1(\eta \cdot \delta/2)^2 + C_E$$

donde  $E_1(\cdot)$  es la función integral exponencial (Abramowitz & Stegun 1965, pág. 227) y  $C_E$  es la constante de Euler definida anteriormente.

### 2.8.6 Gaussiana (GAU)

Una de las funciones de base radial más popular, junto con la TPS, es la Gaussiana. Schagen (1979) fue el primero en usar la Gaussiana como función de base radial. Esta función esta dada por

$$\phi(\delta) = e^{-\eta\delta^2}$$

donde  $\eta \neq 0$  es el parámetro de suavizado de libre elección.

Franke (1982) encuentra que es muy sensible a la elección del parámetro  $\eta$ , como se podría esperar. En particular, parece que los usuarios de esta función se dejan seducir por su suavidad y rápida descomposición. Además, la matriz de interpolación Gaussiana es definida positiva si los centros son distintos, y también es adecuado su uso en técnicas iterativas (Baxter 1992).



# Capítulo 3

## Modelo basado en distancias para la predicción espacial en presencia de tendencia

### 3.1 Introducción

La información espacial se recoge en muchas aplicaciones de disciplinas tales como; la minería, la hidrogeología, ecología, ciencias de la tierra y el medio ambiente. Estas aplicaciones consideran técnicas como kriging simple, kriging ordinario y kriging universal (UK). Estos métodos de interpolación se han utilizado en la literatura para meta-modelos en geoestadística por Sacks et al. (1989), Cressie (1993), Jin et al. (2001), Santner et al. (2003), Wackernagel (2003), Le & Zidek (2006), Joseph et al. (2008) y van de Kasstele et al. (2009), entre otros. En este capítulo, se considera el problema de seleccionar un modelo basado en distancias que incorpora la información asociada a una variable respuesta  $Z_s$ . Además en estos casos, a menudo se tiene que lidiar con variables explicativas de diferente naturaleza que se asocian con la variable respuesta: variables categóricas y binarias tales como el tipo de suelo o roca, y variables continuas (por ejemplo, las mismas coordenadas espaciales y algunas covariables ambientales).

Por lo tanto, se propone en este capítulo un nuevo método de interpolación espacial con variables explicativas mixtas utilizando distancias entre individuos, tales como la distancia de Gower (1968); aunque, alguna otra distancia Euclidiana se puede utilizar. El método basado en distancias (Distance-Based, DB) se utiliza en los modelos geoestadísticos propuestos no sólo en la etapa de estimación de la tendencia para su remoción, sino también en la etapa de la estimación de la correlación de espacial.

En el caso de la regresión geoestadística, el método DB espacial propuesto se basa en los métodos desarrollados por Cuadras & Arenas (1990) y Cuadras et al. (1996), quienes presentan algunos resultados del modelo DB para la predicción con variables mixtas y exploran el problema de información faltante, estableciendo una solución a través de DB. Esta estrategia es una excelente alternativa, ya que aprovecha al máximo la información obtenida debido a la relación entre las observaciones, la cual puede ser establecida a través del uso de la descomposición espectral, utilizando cualquier distancia euclídea. En consecuencia, este enfoque permite mejorar las predicciones del modelo seleccionado para un número de coordenadas principales incluidas asociado a las localizaciones muestreadas.

En esta investigación, las coordenadas principales obtenidas mediante el método basado en distancias se consiguen a partir de las covariables asociadas con la variable de respuesta, y las coordenadas espaciales de los puntos definidos en una forma polinómica de orden 1, 2 o 3. La selección de las coordenadas principales se lleva a cabo usando los valores mas altos de la prueba estadística  $t$  (coordenadas principales significativas al 5%), y el gráfico asociado a la falta de predictibilidad (una caída significativa, véase la Sección 3.2), es decir, las variables explicativas (coordenadas principales) que están más asociadas con la variable respuesta. Aunque, las técnicas populares como el análisis de regresión tradicional (forward F, backward B, y step-wise “BF” or “FB”) se puede utilizar ya que las coordenadas principales no están correlacionadas. Recientemente, otras técnicas han sido propuestas por George & McCulloch (1993), Breiman (1995), Tibshirani (1996), Efron et al. (2004), Joseph et al. (2008), Emery & Silva (2009) y Emery & Cornejo (2010).

Además, la alternativa espacial propuesta basada en los métodos DB para el modelado de la tendencia puede ser también robusta ante errores de especificación en los parámetros de correlación (Cuadras et al. 1996). En este sentido en el capítulo, se realizan simulaciones incondicionales para validar la eficacia del método propuesto bajo diferentes condiciones y los resultados muestran una ganancia significativa en comparación con el tradicional modelo de krigado universal. También se consideran dos aplicaciones: la temperatura media diaria de la Tierra en Croacia (Hengl 2009) y la concentración de calcio medido a una profundidad de 0-20 cm en Brasil (Capeche et al. 1997).

En este capítulo se presenta el método espacial propuesto basado en distancias. En la Sección 3.2 se desarrolla la propuesta metodológica, el variograma para la tendencia DB, se muestra la propiedad de insesgadez del predictor y se presenta el krigado universal construido a partir de la tendencia basada en distancias. En la Sección 3.3 se presenta un estudio de simulación basado en modelos de campos aleatorios Gaussianos. Por último, en la Sección 3.4 se desarrollan dos aplicaciones que ilustran la metodología propuesta en este capítulo.

## 3.2 Modelo basado en distancias con tendencia

Supóngase que se está interesado en relacionar una variable respuesta continua con variables georeferenciadas explicativas medidas en cada sitio de muestreo, estas variables pueden ser del tipo: latitud y longitud, binarias, categóricas y continuas. Sea  $\mathbf{s} \in \mathbb{R}^d$  una ubicación en un espacio Euclidiano  $d$ -dimensional, y supóngase que  $Z(\mathbf{s})$  es un vector aleatorio en cada ubicación espacial  $\mathbf{s}$ . Si  $\mathbf{s}$  varía en un conjunto  $D \subseteq \mathbb{R}^d$  (por lo general  $d = 2$ , pero no necesariamente), se tiene que un proceso estocástico  $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ , el cual es objeto de estudio en el contexto de la geoestadística (Cressie 1993). También se asume que  $D$  es una región fija y continua, y el índice espacial  $\mathbf{s}$  varía de forma continua en  $D$ , es decir, hay un número infinito de posibles lugares donde se observa el proceso.

Además, se supone que el proceso estocástico sigue un modelo de un campo aleatorio Gaussiano dado por (Cressie 1993)

$$Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \quad i = 1, \dots, n \quad (3.1)$$

donde  $Z(\mathbf{s}_i)$  es la variable regionalizada dada por la suma de una función determinística asociada a la tendencia  $\mu(\mathbf{s}_i)$  y  $\varepsilon(\mathbf{s}_i)$  una componente estocástica estacionaria con media cero y variograma  $2\gamma(\cdot)$ . La tendencia espacial está formada por las variables categóricas, continuas y binarias, y se modela como

$$\mu(\mathbf{s}_i) = \theta_0 + v'(\mathbf{s}_i)\boldsymbol{\theta} \quad (3.2)$$

donde  $v(\mathbf{s}_i) = (v_1(\mathbf{s}_i), \dots, v_p(\mathbf{s}_i))'$  es un vector que contiene variables explicativas asociadas a la localización espacial  $\mathbf{s}_i$ ,  $\theta_0$  es el parámetro desconocido asociado al intercepto y  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)'$  es un vector de parámetros desconocidos.

En forma matricial el modelo (3.1) se puede expresar como:

$$\mathbf{Z}_s = \mathbf{1}\theta_0 + V\boldsymbol{\theta} + \boldsymbol{\varepsilon}_s \quad (3.3)$$

donde  $\mathbf{Z}_s = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$ ,  $\mathbf{1}$  es un vector de dimensión  $n \times 1$  asociado al intercepto,  $V = (V_1, \dots, V_n)$  es la matriz de diseño de dimensión  $n \times p$  con  $p$  variables explicativas  $V_j = (v_j(\mathbf{s}_1), \dots, v_j(\mathbf{s}_n))'$  de dimensión  $n \times 1$ ,  $j = 1, \dots, p$ . Además,  $\boldsymbol{\varepsilon}_s = (\varepsilon(\mathbf{s}_1), \dots, \varepsilon(\mathbf{s}_n))'$  y  $\Sigma_{\boldsymbol{\theta}}$  la matriz de varianzas-covarianzas de las observaciones.

Ahora, la idea es hacer una transformación de las variables explicativas utilizando el método basado en distancias. Para ello se definen las medidas de similitud o distancia Euclidiana presentadas en la Sección 2.7.1, que dependen de las características de las variables explicativas.

Si el vector  $v(\mathbf{s}_i)$  dado en (3.2) está formado por variables binarias, categóricas y continuas, entonces la similitud de acuerdo a Gower (1971) se puede definir para variables mixtas como la expresión presentada en (2.17). En el caso que las variables explicativas sean binarias o categóricas, como se mencionó en la Sección 2.7.1, la similitud se puede expresar mediante las expresiones presentadas en (2.16). Por medio de la transformación

$$\delta_{ij} = \sqrt{1 - m_{ij}}$$

es posible obtener las distancias Euclidianas. Si todas las variables explicativas en (3.2) son continuas, la distancia al cuadrado se define como

$$\delta_{ij}^2 = (v(\mathbf{s}_i) - v(\mathbf{s}_j))'(v(\mathbf{s}_i) - v(\mathbf{s}_j))$$

o alternativamente por la distancia absoluta  $\delta_{ij}^2 = \sum_{l=1}^p |v_l(\mathbf{s}_i) - v_l(\mathbf{s}_j)|$ . Entonces, en el caso de sólo disponer información de las coordenadas espaciales  $(w_x, w_y)$ , las distancias espaciales estarán dadas por  $\delta_{ij} = \sqrt{(w_{x_i} - w_{x_j})^2 + (w_{y_i} - w_{y_j})^2}$ . Expresiones para la similaridad de Gower como la dada en la ecuación (2.17) serán útiles en la medida de disponer de información asociada con las variables mixtas, no sólo para los puntos muestreados sino también para los no muestreados, lo cual restringe su uso en las zonas no muestreadas.

Una vez seleccionada alguna de las distancias presentadas anteriormente, se define la matriz  $A_{n \times n} = (a_{ij})$  con elementos  $a_{ij} = -\delta_{ij}^2/2$  y  $B = HAH$  donde  $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}'$  es la matriz centrada, con  $\mathbf{1}$  un vector de unos de tamaño  $n \times 1$ . Se sabe que  $B$  es una matriz semidefinida positiva (Mardia et al. 2002) de rango  $k$ , por lo que la matriz  $X$  de coordenadas principales se obtiene a partir de la descomposición espectral como

$$B = HAH = U\Lambda U' = XX'$$

donde  $\Lambda$  es una matriz diagonal que contiene los valores propios de  $A$ ,  $X = U\Lambda^{1/2}$  es una matriz  $n \times n$  de rango  $k \leq n - 1$ , y  $U$  contiene las coordenadas estandarizadas.

Luego, el modelo presentado en (3.2) se convierte en

$$\mathbf{Z}_s = \mathbf{1}\beta_0 + X\boldsymbol{\beta} + \boldsymbol{\epsilon}_s \quad (3.4)$$

donde  $X = (X_1, \dots, X_k)$ ,  $\beta_0$  y  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$  son los parámetros desconocidos. Notemos que  $\mathbf{1}, X_1, \dots, X_k$ , son vectores propios de  $B$  con valores propios  $0, \lambda_1, \dots, \lambda_k$ , respectivamente, y  $X_i'X_i = \lambda_i$ ,  $X_i'X_j = 0$  ( $i \neq j$ ), y  $X_i'\mathbf{1} = 0$ ,  $i, j = 1, \dots, k$ .

Para evitar el problema de tener un coeficiente de determinación  $R^2 \simeq 1$  cuando el rango de  $X$  es de  $k = n - 1$ , es necesario tener en cuenta sólo



los vectores propios más correlacionados de  $B$  dados por  $(X_1, \dots, X_k)$  con la variable regionalizada  $\mathbf{Z}_s$ , es decir, las covariables más significativamente correlacionadas con  $\mathbf{Z}_s$ . Con el fin de decidir si una coordenada principal debe ser incluida o eliminada del modelo, las coordenadas principales se arreglan en orden decreciente de acuerdo a su correlación en valor absoluto con  $\mathbf{Z}_s$ , es decir haciendo

$$r(\mathbf{Z}_s, X_1) > \dots > r(\mathbf{Z}_s, X_k) > r(\mathbf{Z}_s, X_{k+1}) > \dots > r(\mathbf{Z}_s, X_{n-1}) \quad (3.5)$$

donde  $r^2(\mathbf{Z}_s, X_i) = \frac{\mathbf{Z}'_s X_i X'_i \mathbf{Z}_s}{\lambda_i \sum_{j=1}^n [Z(\mathbf{s}_j) - \bar{Z}]^2}$ ,  $i = 1, \dots, n-1$ , con  $\bar{Z} = \sum_{j=1}^n Z(\mathbf{s}_j)/n$ .

Más aún, una coordenada principal  $X_i$  debería ser eliminada si la hipótesis nula  $\beta_i = 0$  no se rechaza. Una prueba de hipótesis se puede basar en (Cuadras et al. 1996)

$$t_i = \frac{\hat{\beta}_i}{\|\mathbf{Z}_s - \hat{\beta}_0 \mathbf{1} - X \hat{\boldsymbol{\beta}}\|^2} \sqrt{\lambda_i (n - k - 1)}, \quad i = 1, \dots, n-1 \quad (3.6)$$

donde  $\hat{\beta}_0 = \bar{Z}$ ,  $\hat{\boldsymbol{\beta}} = \Lambda^{-1} X' \mathbf{Z}_s$  y  $\hat{\beta}_i$  es la  $i$ -ésima componente de  $\hat{\boldsymbol{\beta}}$ . Así,  $t_i$  sigue una distribución  $t$ -student con  $(n - k - 1)$  grados de libertad.

Otra posibilidad es a través de la variabilidad explicada dada por las variables predictoras, la cual es establecida por los valores propios más grandes  $\lambda_1 > \dots > \lambda_k > \lambda_{k+1} > \dots > \lambda_{n-1}$  escogiendo las  $k$  primeras coordenadas principales. Sin embargo, una coordenada principal con valor propio pequeño puede estar correlacionada con la variable respuesta, entonces esta dimensión estará correlacionada con el “ruido” en lugar de la variabilidad principal de los datos (Cuadras 1993).

Otra buena alternativa para la selección de coordenadas principales se realiza de una manera similar a la selección del número de variables en regresión multivariada mediante la estadística llamada  $C_p$ -Mallows. Es decir, se construye un gráfico donde se representan los puntos  $(i, 1 - c(i))$   $i = 0, 1, \dots, n-1$ , y luego se determina el punto en donde hay una caída significativa de la falta

de predictibilidad dada por  $1 - c(i)$ , la predictibilidad  $c(i)$  está dada por

$$c(0) = 0, \quad c(i) = \frac{\sum_{j=1}^i r_j^2 \lambda_j}{\sum_{j=1}^{n-1} r_j^2 \lambda_j} \quad i = 1, \dots, n-1 \quad (3.7)$$

donde  $r_j = \text{Corr}(\mathbf{Z}_s, X_j)$  y  $\lambda_j$  es el  $j$ -ésimo valor propio asociado con  $X_j$ ,  $j = 1, 2, \dots, n-1$  (ver mayores detalles en (Cuadras et al. 1996)).

Finalmente, en cualquiera de los cuatro métodos presentados anteriormente, las  $X_{k+1}, \dots, X_{n-1}$  coordenadas principales deben ser removidas ya que son las menos relevantes.

### 3.2.1 Kriging universal basado en distancias (DBUK)

Hasta el momento, se ha descrito cómo las coordenadas principales calculadas a partir de las variables explicativas y utilizadas para estimar la componente de tendencia de la ecuación (3.1), se obtienen utilizando la descomposición espectral de la matriz de similitudes o distancias. Una vez que la tendencia se ha calculado utilizando el método DB, se puede eliminar la tendencia de los datos y obtener los residuales  $\varepsilon(\cdot)$  en el modelo (3.1). Por consiguiente, en la primera parte de esta subsección, se describe el ajuste del variograma de los residuales a partir de la tendencia estimada.

El variograma experimental  $\hat{\gamma}(h)$  es la herramienta clave para cualquier análisis geoestadístico porque describe las correlaciones espaciales de la variable regionalizada a diferentes distancias. Un estimador natural obtenido por el método de momentos, debido a Matheron (Cressie 1993), está dado por

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{N(h)} [\hat{\varepsilon}(\mathbf{s}_i) - \hat{\varepsilon}(\mathbf{s}_j)]^2$$

donde  $N(h)$  es el número de pares distintos de observaciones separadas a una distancia  $h$ ,  $\hat{\varepsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_i)$  y  $\hat{\varepsilon}(\mathbf{s}_j) = Z(\mathbf{s}_j) - \hat{Z}(\mathbf{s}_j)$  son los valores de los residuales en las localizaciones  $\mathbf{s}_i$  y  $\mathbf{s}_j$ , respectivamente, con  $\hat{Z}(\mathbf{s}_i)$  y  $\hat{Z}(\mathbf{s}_j)$  las predicciones en las localizaciones  $\mathbf{s}_i$  y  $\mathbf{s}_j$ , respectivamente, usando

el método DB. Este estimador es generalmente sesgado en presencia de información atípica, afectando la estimación, por lo cual es recomendable el uso de estimadores robustos como el de (Cressie & Hawkins 1980) o la media recortada (*trim.m*) establecida en (Bardossy et al. 1997) y dado por

$$\hat{\gamma}(h) = \frac{\left[ \text{trim.m} \left( |\hat{\varepsilon}(\mathbf{s}_i) - \hat{\varepsilon}(\mathbf{s}_j)|^{\frac{1}{2}} \right) \right]^4}{0.457 + 0.494/N(h)}$$

Una vez que el variograma experimental se encuentra, se ajusta el modelo del variograma. Hay varios métodos de estimación de parámetros, tales como mínimos cuadrados ordinarios (OLS), mínimos cuadrados ponderados (WLS), máxima verosimilitud (ML) y máxima verosimilitud restringida (REML). Los dos últimos requieren la normalidad, mientras que los dos primeros no. Para ilustrar la metodología y sin pérdida de generalidad, se presenta el método de máxima verosimilitud (ML). Bajo la consideración que  $Z(\mathbf{s}_i)$  en (3.1) o (3.4) es un proceso Gaussiano, dos veces el negativo de la log-verosimilitud esta dado por

$$L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\vartheta}) = \left( \mathbf{Z}_s - \mathbf{X}\tilde{\boldsymbol{\beta}} \right)' \Sigma_{\boldsymbol{\vartheta}}^{-1} \left( \mathbf{Z}_s - \mathbf{X}\tilde{\boldsymbol{\beta}} \right) + \log |\Sigma_{\boldsymbol{\vartheta}}| + n \log(2\pi) \quad (3.8)$$

donde  $\Sigma_{\boldsymbol{\vartheta}}$  es la matriz de covarianza del proceso  $Z(\mathbf{s}_i)$ ,  $\mathbf{X} = (\mathbf{1}, X) = (\mathbf{1}, X_1, \dots, X_k)$  y  $\tilde{\boldsymbol{\beta}} = (\beta_0, \boldsymbol{\beta}')' = (\beta_0, \beta_1, \dots, \beta_k)'$ . Los parámetros dados en  $\boldsymbol{\vartheta}$  usualmente incluyen la pepita ( $\tau^2$ ), rango ( $\phi$ ) y meseta parcial ( $\sigma^2$ ), y en el caso del modelo Matérn, éste incluye el parámetro de suavizamiento  $\kappa$ . Los valores  $\tilde{\boldsymbol{\beta}}$  y  $\boldsymbol{\vartheta}$  se obtienen maximizando iterativamente y simultáneamente la función de distribución normal multivariante, o, alternativamente, minimizando la expresión (3.8). La minimización se realiza en dos fases: en la primera etapa se supone que  $\boldsymbol{\vartheta}$  es conocido, y por lo tanto  $\Sigma_{\boldsymbol{\vartheta}}$  también, entonces la mejor estimación de la media de los parámetros  $\tilde{\boldsymbol{\beta}}$  se obtiene utilizando el método de mínimos cuadrados generalizados (generalized least squares, GLS):

$$\hat{\tilde{\boldsymbol{\beta}}} = \left( \mathbf{X}' \Sigma_{\boldsymbol{\vartheta}}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \Sigma_{\boldsymbol{\vartheta}}^{-1} \mathbf{Z}_s \quad (3.9)$$

En la segunda etapa, se reemplaza este valor en (3.8), para obtener

$$L(\boldsymbol{\vartheta}) = \left( \mathbf{Z}_s - \mathbf{X}\hat{\tilde{\boldsymbol{\beta}}} \right)' \Sigma_{\boldsymbol{\vartheta}}^{-1} \left( \mathbf{Z}_s - \mathbf{X}\hat{\tilde{\boldsymbol{\beta}}} \right) + \log |\Sigma_{\boldsymbol{\vartheta}}| + n \log(2\pi) \quad (3.10)$$

Así, la expresión (3.10) se minimiza sólo con respecto a  $\boldsymbol{\vartheta}$  para hacer el proceso más simple, y de esa expresión se encuentra  $\hat{\boldsymbol{\vartheta}}$ . Iterando varias veces las dos etapas anteriores, se encuentran las estimaciones  $\tilde{\boldsymbol{\beta}}$  y  $\boldsymbol{\vartheta}$ . Para iniciar el proceso iterativo, se puede obtener una primera estimación de los parámetros  $\boldsymbol{\vartheta}$  de  $\Sigma_{\boldsymbol{\vartheta}}$  del variograma experimental basada en el juicio del investigador.

Una vez estimados los parámetros asociados con la función de covarianza o semivarianza (si  $\Gamma_{\boldsymbol{\vartheta}}$  se considera en lugar de  $\Sigma_{\boldsymbol{\vartheta}}$ ) y ya que las distancias  $\delta_{ij}$  son Euclidianas, la transformación a un modelo de semivarianza  $\gamma_{ij}$  es también Euclidiana. Esta es la forma más sencilla de garantizar que la función del variograma obtenido a partir de estas distancias sea condicionalmente definida negativa (Armstrong & Diamond 1984), la cual se expresa como

$$-\sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \gamma_{ij} \geq 0, \quad \forall \varphi_r$$

donde  $\varphi_r \in \mathbb{R}$ ,  $r = 1, \dots, n$ .

Después de que lo anterior, se hacen las predicciones espaciales en nuevas localizaciones espaciales,  $\mathbf{s}_0$ , en donde se observan un conjunto de variables explicativas mixtas. Para conseguirlo, se utiliza el método de kriging universal con la finalidad de construir a partir de la tendencia basada en distancias las predicciones espaciales.

Por lo tanto, supóngase que un nuevo individuo ( $n+1$ ) es observado con sus respectivas variables explicativas mixtas, es decir  $v(\mathbf{s}_0) = (v_1(\mathbf{s}_0), \dots, v_p(\mathbf{s}_0))'$  es conocido. Entonces, las distancias entre el nuevo individuo y cada uno de los individuos involucrados en el modelo propuesto en (3.2) se pueden calcular como  $\delta_{0i} = \delta(v(\mathbf{s}_0), v(\mathbf{s}_i))$ ,  $i = 1, \dots, n$ . A partir de estas distancias, se puede hacer una predicción usando un resultado de Gower (1971) y Cuadras & Arenas (1990), que relaciona el vector  $\boldsymbol{\delta}_0 = (\delta_{01}^2, \dots, \delta_{0n}^2)'$  de cuadrados de las distancias con el vector  $x(\mathbf{s}_0) = (x_1(\mathbf{s}_0), \dots, x_k(\mathbf{s}_0))'$  de coordenadas principales asociado al nuevo individuo mediante la expresión

$$\delta_{0i}^2 = [x(\mathbf{s}_0) - x(\mathbf{s}_i)]' [x(\mathbf{s}_0) - x(\mathbf{s}_i)]$$

con  $i = 1, \dots, n$ . Luego, se encuentra que

$$x(\mathbf{s}_0) = \frac{1}{2} \Lambda^{-1} X'(b - \boldsymbol{\delta}_0)$$

donde  $b = (b_{11}, \dots, b_{nm})'$  y  $b_{ii} = x(\mathbf{s}_i)'x(\mathbf{s}_i)$ ,  $i = 1, \dots, n$ .

Ahora, el próximo objetivo es predecir el valor de  $Z(\mathbf{s}_0)$  basado en un conjunto de observaciones  $\mathbf{Z}_s$ . Para esto, el predictor UK esta dado por

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) = \boldsymbol{\varphi}' \mathbf{Z}_s$$

donde el vector de coeficientes  $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_n)'$  satisface que  $\sum_{i=1}^n \varphi_i = 1$ . Este predictor cumple la condición de insesgamiento, la cual esta dada por la siguiente expresión

$$\begin{aligned} E\left(\hat{Z}(\mathbf{s}_0)\right) &= E(\boldsymbol{\varphi}' \mathbf{Z}_s) = \boldsymbol{\varphi}' E(\mathbf{Z}_s) = \boldsymbol{\varphi}' \mathbf{X} \tilde{\boldsymbol{\beta}} = E(Z(\mathbf{s}_0)) \\ &= \mathbf{x}'(\mathbf{s}_0) \tilde{\boldsymbol{\beta}} = \boldsymbol{\mu}(\mathbf{s}_0) \end{aligned} \quad (3.11)$$

donde  $\mathbf{x}(\mathbf{s}_0) = (1, x'(\mathbf{s}_0))' = (1, x_1(\mathbf{s}_0), \dots, x_k(\mathbf{s}_0))'$  es un vector formado por 1 y las coordenadas principales del nuevo individuo  $x(\mathbf{s}_0)$ , con  $x(\mathbf{s}_0) = (x_1(\mathbf{s}_0), \dots, x_k(\mathbf{s}_0))'$ . A partir de (3.11) se tiene  $\boldsymbol{\varphi}' \mathbf{X} = \mathbf{x}'(\mathbf{s}_0)$ . La condición (3.11) garantiza que el predictor sea insesgado y de mínima varianza (Cressie 1993).

Por otro lado, el error cuadrático medio de la predicción,  $\text{Var}(\mathbf{s}_0)$ , bajo la condición de insesgamiento, está dado por

$$\text{Var}(\hat{\varepsilon}(\mathbf{s}_0)) = \text{Var}\left[\left(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)\right)\right] = \boldsymbol{\varphi}' \Sigma_{\boldsymbol{\vartheta}} \boldsymbol{\varphi} + \sigma^2 - 2\boldsymbol{\varphi}' \mathbf{c} \quad (3.12)$$

donde  $\Sigma_{\boldsymbol{\vartheta}}$  es una matriz  $n \times n$  cuyo  $(i, j)$ -ésimo término es  $C(\varepsilon(\mathbf{s}_i), \varepsilon(\mathbf{s}_j))$ ,  $\sigma^2 = C(\mathbf{0})$  y  $\mathbf{c}$  es un vector de dimensión  $n$  cuyo  $i$ -ésimo elemento es  $C(\varepsilon(\mathbf{s}_0), \varepsilon(\mathbf{s}_i))$ .

Note que ahora se debe estimar  $\boldsymbol{\varphi}$ , para ello, se reemplaza  $\Sigma_{\boldsymbol{\vartheta}}$  y  $\sigma^2$  en (3.12) por sus estimaciones obtenidas mediante el procedimiento de optimización presentado en (3.9) y (3.10). Por lo tanto,  $\boldsymbol{\varphi}$  se encuentra al minimizar la siguiente expresión

$$\mathcal{L}(\boldsymbol{\varphi}, \mathbf{l}) = \boldsymbol{\varphi}' \Sigma_{\hat{\boldsymbol{\vartheta}}} \boldsymbol{\varphi} + \hat{\sigma}^2 - 2\boldsymbol{\varphi}' \mathbf{c} + 2\mathbf{l}' (\mathbf{X}' \boldsymbol{\varphi} - \mathbf{x}'(\mathbf{s}_0))$$

donde  $\mathbf{l}$  es el vector de  $k + 1$  multiplicadores de Lagrange asociados con la restricción de insesgidez.

Después de la diferenciación con respecto a  $\boldsymbol{\varphi}$  y  $\boldsymbol{l}$ , igualando el resultado a cero y realizando algunos procesos algebraicos, se encuentra el siguiente sistema matricial

$$\begin{pmatrix} \Sigma_{\hat{\vartheta}} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\varphi} \\ \boldsymbol{l} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \mathbf{x}(s_0) \end{pmatrix}$$

Resolviendo el sistema, los coeficientes para  $\boldsymbol{\varphi}$  y  $\boldsymbol{l}$  están dados por

$$\begin{aligned} \hat{\boldsymbol{\varphi}}' &= \left[ \mathbf{c} + \mathbf{X} \left( \mathbf{X}' \Sigma_{\hat{\vartheta}}^{-1} \mathbf{X} \right)^{-1} \left( \mathbf{x}(s_0) - \mathbf{X}' \Sigma_{\hat{\vartheta}}^{-1} \mathbf{c} \right) \right]' \Sigma_{\hat{\vartheta}}^{-1} \\ \hat{\boldsymbol{l}} &= - \left( \mathbf{X}' \Sigma_{\hat{\vartheta}}^{-1} \mathbf{X} \right)^{-1} \left( \mathbf{x}(s_0) - \mathbf{X}' \Sigma_{\hat{\vartheta}}^{-1} \mathbf{c} \right) \end{aligned} \quad (3.13)$$

La estimación del cuadrado medio del error de predicción en términos de  $\hat{\boldsymbol{\varphi}}$  y  $\hat{\boldsymbol{l}}$  puede ser expresado como

$$\widehat{\text{Var}}(\hat{\varepsilon}(s_0)) = \hat{\boldsymbol{\varphi}}' \mathbf{c} - \mathbf{x}'(s_0) \hat{\boldsymbol{l}} + \hat{\sigma}^2 - 2\hat{\boldsymbol{\varphi}}' \mathbf{c} = \hat{\sigma}^2 - \left( \hat{\boldsymbol{l}}' \mathbf{x}(s_0) + \hat{\boldsymbol{\varphi}}' \mathbf{c} \right) \quad (3.14)$$

Reemplazando (3.13) dentro de (3.14), la estimación del cuadrado medio del error de predicción es

$$\begin{aligned} \widehat{\text{Var}}(\hat{\varepsilon}(s_0)) &= \hat{\sigma}^2 - \mathbf{c}' \Sigma_{\hat{\vartheta}}^{-1} \mathbf{c} - \left( \mathbf{x}(s_0) - \mathbf{X}' \Sigma_{\hat{\vartheta}}^{-1} \mathbf{X} \right)' \left( \mathbf{X}' \Sigma_{\hat{\vartheta}}^{-1} \mathbf{X} \right)^{-1} \\ &\quad \left( \mathbf{x}(s_0) - \mathbf{X}' \Sigma_{\hat{\vartheta}}^{-1} \mathbf{c} \right) \end{aligned} \quad (3.15)$$

Finalmente, se puede resumir el procedimiento presentado en esta sección en los siguientes pasos:

1. Obtener las coordenadas principales utilizando la descomposición espectral de la matriz de similitudes (o distancias) calculada a partir de las variables explicativas.
2. Seleccionar las coordenadas principales más correlacionadas o significativas con la variable regionalizada  $\mathbf{Z}_s$ . En este paso, se recomienda usar el criterio dado en (3.5) para hacer una primera selección con el fin de remover las coordenadas principales pobremente correlacionadas con la variable regionalizada, y luego, emplear los criterios (3.6) o (3.7) para seleccionar las coordenadas principales mas significativas utilizando la regresión DB.

3. Construir los residuales  $\hat{\varepsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_i)$  y calcular el variograma experimental.
4. Ajustar el modelo del variograma estimando  $\tilde{\beta}$  y  $\Sigma_{\tilde{\beta}}$ , el cual se obtiene utilizando iterativamente (3.9) y (3.10).
5. Hacer las predicciones en los puntos muestreados y no muestreados para generar el mapa de predicción usando el método DBUK (es decir, haciendo  $\hat{Z}(\mathbf{s}_0) = \hat{\varphi}'\mathbf{Z}_s$ ) y hacer el mapa de la predicción de las varianzas del error asociado empleando (3.15).

### 3.2.2 Medidas de evaluación

Se considera la raíz del cuadrado medio de los errores de predicción (RMSPE) para evaluar la exactitud en los métodos de interpolación UK y DBUK. El RMSPE es obtenido mediante validación cruzada “leave-one-out” (LOOCV), el cual puede utilizarse para comparar el rendimiento de algunos métodos de interpolación. Como se explicó en la Sección 1.7, LOOCV consiste en remover una observación de los  $n$  puntos muestrales (por lo general asociado a un vecindario), y luego con los  $n - 1$  valores restantes y el modelo de variograma seleccionado, se predice via kriging universal el valor de la variable de estudio en la localización que se removió. Este procedimiento se realiza en forma secuencial con cada uno de los puntos muestreados y así se obtiene un conjunto de  $n$  errores de predicción. Si el modelo de variograma elegido describe bien la estructura de autocorrelación espacial, entonces las diferencias entre los valores observados y predichos deberían ser pequeñas, y así, se podrá producir el mapa.

Este procedimiento se justifica debido a que los métodos de interpolación kriging son exactos, es decir, los valores de predicción coinciden con los valores observados para los puntos muestreados. De esta manera, el proceso de validación cruzada da una idea de qué tan buenas son las predicciones, por lo cual brinda información acerca de cuál modelo provee predicciones más exactas. Las expresiones tanto para el RMSPE como para el coeficiente de determinación

( $R^2$ ) se muestran a continuación:

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^n \left( \hat{Z}_{[i]}(\mathbf{s}_i) - Z(\mathbf{s}_i) \right)^2}{n}} \quad (3.16)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n \left( \hat{Z}_{[i]}(\mathbf{s}_i) - Z(\mathbf{s}_i) \right)^2}{\sum_{i=1}^n \left( Z(\mathbf{s}_i) - \bar{Z} \right)^2} \quad (3.17)$$

donde  $\hat{Z}_{[i]}(\mathbf{s}_i)$  es el valor predicho a partir de la validación cruzada y  $Z(\mathbf{s}_i)$  es el valor muestreado en la localización  $\mathbf{s}_i$ .

Una variación de la metodología previa, consiste en dividir la muestra en dos submuestras; la primera submuestra es empleada para el modelamiento del variograma y la otra submuestra es utilizada para validar el método kriging. Después de esto, las medidas de validación pueden ser construidas a partir de los valores observados y de las predicciones (Bivand et al. 2008). Si todo funciona bien el RMSPE debería ser tan pequeño como sea posible (cercano a cero) y el  $R^2$  debería ser cercano a 1

### 3.3 Estudio de simulación y discusión

Para evaluar el método propuesto, un estudio de simulación se realizó bajo ciertas condiciones. En estos escenarios, para el método clásico UK se consideran las variables explicativas mixtas en la forma tradicional y en el método DBUK se utilizan las coordenadas principales obtenidas a partir de las variables explicativas. En este último caso, las coordenadas principales son obtenidas utilizando el criterio dado en (3.5) para hacer la primera selección, dejando afuera las coordenadas principales con correlación cercana a cero. Luego se utiliza el criterio dado en (3.6) para seleccionar las coordenadas principales más significativas en la regresión DB. También, estas son empleadas para la estimación de los modelos de variograma y tendencia, es decir, es aplicada la metodología expuesta en la Subsección 3.2.1. El estudio de simulación considera dos casos que con frecuencia se presentan en la práctica: en el primer caso todas las variables explicativas son incluidas, mientras en el segundo caso una variable explicativa relevante es removida (u omitida). Este segundo caso es



motivado por algunas razones: (a) la imposibilidad de realizar las mediciones, (b) la pérdida de información en algunos lugares, (c) una inapropiada forma funcional y (d) el desconocimiento sobre las variables que deberían considerarse en relación con respecto al fenómeno modelado, lo que provoca la pérdida de una o más variables que pueden ser relevantes para la investigación, y por lo tanto, el modelo geoestadístico propuesto bajo estas condiciones no es el adecuado.

### 3.3.1 Caso 1: Tendencia basada en variables mixtas sin omisión de variables explicativas

La tendencia es construida considerando una variable aleatoria binomial,  $V_1 \sim B_i(n, p = 0.4)$ , con  $n = 50, 100, 150$ . También, una variable nominal asociada con tres regiones fijadas en un cuadrado de una unidad, tal como se muestra en la Figura 3.1. Puesto que hay tres regiones, sólo dos variables dummy se consideran ( $D_2, D_3$ ) para evitar problemas de singularidad. Además, se asume que el error  $\varepsilon(\mathbf{s}_i)$  es un proceso Gaussiano isotrópico con media cero y función de covarianza generada a partir de un modelo de variograma específico con un rango de valores para los parámetros: pepita ( $\tau^2$ ), rango ( $\phi$ ), meseta parcial ( $\sigma^2$ ), kappa ( $\kappa$ ) y tamaño muestral ( $n$ ). Cuatro modelos de variograma teóricos fueron usados: Exponencial (Exponential, EXP), Matérn (MAT), Gaussiano (Gaussian, GAU) y Esférico (Spherical, SPH). Los rangos de los escenarios simulados se presentan en la Tabla 3.1. Los parámetros de tendencia se fijaron como  $\beta_0 = 10, \beta_1 = -4, \beta_2 = 2$  y  $\beta_3 = -4$  asociados a las coordenadas espaciales  $w_{x_i}$  y  $w_{y_i}$ ,  $i = 1, \dots, n$ . Así, el proceso regionalizado simulado está dada por

$$Z(\mathbf{s}_i) = \beta_0 + \beta_1 V_{i1} + \beta_2 D_{i2} w_{x_i} + \beta_3 D_{i3} w_{y_i} + \varepsilon(\mathbf{s}_i) \quad (3.18)$$

Con el fin de examinar la versatilidad del método DBUK con respecto al método UK, 96 escenarios fueron simulados, y para cada uno de ellos el proceso se repitió 100 veces.

En cada uno de los escenarios y con las simulaciones realizadas, los modelos teóricos fueron ajustados a los variogramas experimentales mediante máxima

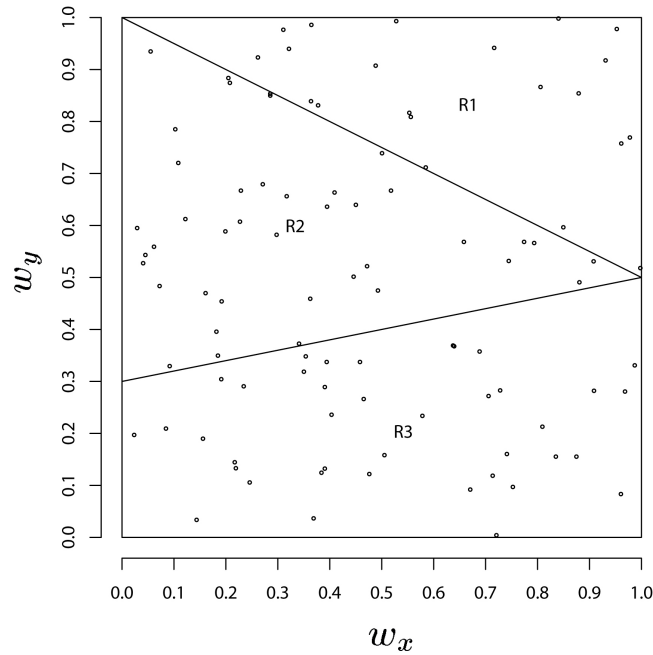


FIGURA 3.1: Localización de los puntos de muestreo y regiones asociadas a la definición de la variable nominal

verosimilitud y los estadísticos RMSPE (3.16) y  $R^2$  (3.17) fueron calculados por validación cruzada. Los resultados en términos de RMSPE se muestran en la Tabla 3.2. Por ejemplo, el escenario 48 corresponde a un modelo de variograma esférico con parámetros:  $\tau_1^2 = 0$ ,  $\sigma_1^2 = 2$ ,  $\phi_1 = 0.60$  y  $n = 150$ , este escenario reporta un valor de RMSPE= 2.713 para UK y RMSPE= 2.619 para DBUK, así el RMSPE fue mas bajo para el DBUK (método propuesto) que para el UK (método clásico). En general, el mismo comportamiento se observó en los otros escenarios de acuerdo a la lectura de los valores de RMSPE. Por lo tanto, hay un ganancia significativa en la reducción de los errores utilizando el método propuesto. La mayor ganancia proviene del modelo Matérn con  $\kappa = 1.5$  porque hay una reducción del 14% en promedio, mientras que la reducción para el modelo exponencial es del 13%, y para los modelos Gaussiano y esférico es tan sólo del 12%.

De acuerdo con la Tabla 3.2 y las Figuras 3.2 y 3.3, el método DBUK es mejor que el método UK porque bajo  $\tau_1^2 = 0$  y  $\tau_2^2 = 1$ , los promedios de RMSPE son 13.7% y 14.6%, respectivamente, mayores en UK que en DBUK.

TABLA 3.1: Escenarios simulados para los casos 1 y 2

Parámetros del modelo			$n$	Modelos de variograma			
$\tau^2$	$\sigma^2$	$\phi$		EXP	MAT ( $\kappa = 1.5$ )	GAU	SPH
0	1	0.15	50	1	4	7	10
			100	2	5	8	11
			150	3	6	9	12
		0.60	50	13	16	19	22
			100	14	17	20	23
			150	15	18	21	24
	2	0.15	50	25	28	31	34
			100	26	29	32	35
			150	27	30	33	36
		0.60	50	37	40	43	46
			100	38	41	44	47
			150	39	42	45	48
1	1	0.15	50	49	52	55	58
			100	50	53	56	59
			150	51	54	57	60
		0.60	50	61	64	67	70
			100	62	65	68	71
			150	63	66	69	72
	2	0.15	50	73	76	79	82
			100	74	77	80	83
			150	75	78	81	84
		0.60	50	85	88	91	94
			100	86	89	92	95
			150	87	90	93	96

Adicionalmente, el RMSPE promedio bajo UK disminuye cuando el tamaño muestral se incrementa, mientras que esto no siempre sucede con el RMSPE promedio en el DBUK. En general, los promedios de RMSPE en el método DBUK no cambian significativamente cuando se incrementa el tamaño muestral, y en la mayoría de los casos, son mas pequeños que los RMSPEs promedio bajo UK. Resultados similares se obtienen al analizar los  $R^2$ ; los promedios del  $R^2$  se incrementan al aumentar el tamaño de la muestra para UK, pero no hay un incremento significativo al variar los tamaños de muestra  $n$  usando el método DBUK, en ambos casos la variabilidad se reduce cuando se incrementa el tamaño de la muestra.

TABLA 3.2: Promedios de RMSPEs bajo los métodos UK y DBUK para los escenarios presentados en la Tabla 3.1 en el Caso 1 (sin omisión de variable)

Parámetros del modelo				Modelos de variograma							
				Exponencial		Matérn ( $\kappa = 1.5$ )		Gaussiano		Esférico	
$\tau^2$	$\sigma^2$	$\phi$	$n$	UK	DBUK	UK	DBUK	UK	DBUK	UK	DBUK
0	1	0.15	50	3.19	2.64	3.19	2.64	3.06	2.62	3.35	2.79
			100	2.90	2.59	2.89	2.60	2.75	2.53	3.06	2.76
			150	2.75	2.61	2.75	2.61	2.60	2.50	2.91	2.74
		0.60	50	2.96	2.38	3.03	2.35	2.82	2.25	2.96	2.51
			100	2.68	2.39	2.72	2.39	2.61	2.33	2.72	2.47
			150	2.57	2.43	2.59	2.44	2.53	2.37	2.60	2.49
	2	0.15	50	3.55	2.98	3.55	2.97	3.31	2.91	3.85	3.23
			100	3.17	2.87	3.16	2.88	2.89	2.73	3.46	3.16
			150	2.99	2.85	2.99	2.86	2.69	2.63	3.27	3.10
		0.60	50	3.09	2.52	3.18	2.47	2.83	2.28	3.15	2.71
			100	2.78	2.50	2.79	2.49	2.60	2.36	2.87	2.63
			150	2.65	2.52	2.66	2.53	2.53	2.39	2.71	2.62
1	1	0.15	50	3.89	3.22	3.88	3.21	3.79	3.22	4.00	3.31
			100	3.61	3.21	3.60	3.21	3.51	3.18	3.75	3.35
			150	3.48	3.23	3.46	3.20	3.37	3.17	3.61	3.35
		0.60	50	3.71	3.01	3.74	2.97	3.60	2.93	3.70	3.11
			100	3.44	3.04	3.47	3.02	3.38	2.97	3.47	3.12
			150	3.33	3.07	3.36	3.07	3.29	3.01	3.36	3.13
	2	0.15	50	4.20	3.53	4.19	3.53	4.02	3.50	4.43	3.72
			100	3.85	3.46	3.84	3.46	3.66	3.37	4.09	3.70
			150	3.69	3.44	3.70	3.45	3.47	3.31	3.93	3.67
		0.60	50	3.82	3.14	3.88	3.09	3.59	2.94	3.88	3.33
			100	3.52	3.13	3.57	3.12	3.37	3.00	3.59	3.27
			150	3.40	3.15	3.43	3.15	3.28	3.04	3.46	3.25

### 3.3.2 Caso 2: Tendencia como en el caso 1, pero omitiendo una variable explicativa

Se considera el mismo conjunto de variables y parámetros definidos en el Caso 1, presentados en la ecuación (3.18), pero se omite la variable  $V_1$ . Por lo tanto, el proceso regionalizado simulado es ahora

$$Z(\mathbf{s}_i) = \beta_0 + \beta_2 D_{i2} w_{x_i} + \beta_3 D_{i3} w_{y_i} + \varepsilon(\mathbf{s}_i) \quad (3.19)$$

Se consideran los escenarios descritos en la Tabla 3.1 y los resultados se presentan en la Tabla 3.3. Los valores en RMSPE fueron en general más bajos que en el Caso 1. Como en el Caso 1, los promedios de RMSPEs fueron

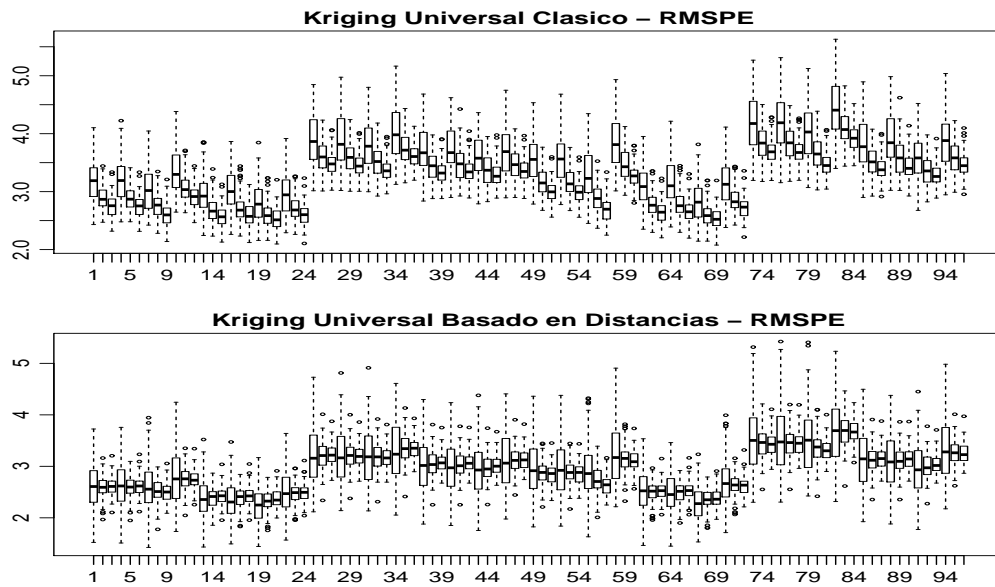
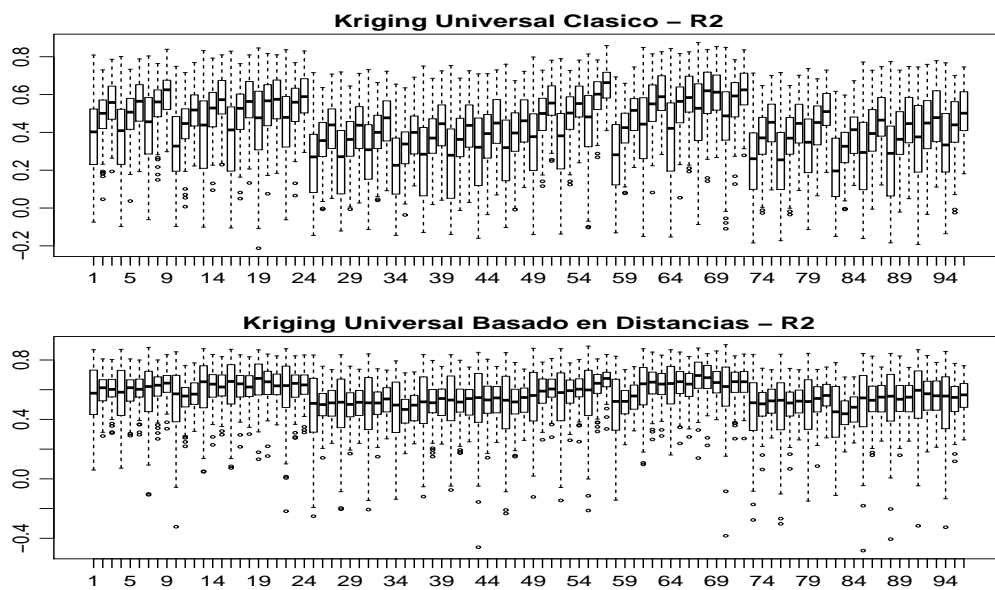


FIGURA 3.2: RMSPE para los escenarios considerados en el Caso 1

FIGURA 3.3:  $R^2$  para los escenarios considerados en el Caso 1

menores en el método DBUK que en el método UK, aunque se encontraron valores notoriamente más grandes en modelo de Matérn. Para todos los modelos de variograma, los promedios de RMSPEs en el método propuesto fueron

aproximadamente un 10% más bajo que en el método tradicional.

En general, las varianzas de los promedios de RMSPEs y  $R^2$  son mas bajas cuando  $N$  se incrementa, siendo muy similar en los tamaños de muestra  $n = 100$  y  $n = 150$  (ver Figuras 3.4 y 3.5). Además, la variabilidad es mayor cuando se utilizan los modelos de Matérn.

TABLA 3.3: Promedios de RMSPEs bajo los métodos UK y DBUK para los escenarios presentados en la Tabla 3.1 en el Caso 2 (con una variable omitida)

Parámetros del modelo			Modelos de variograma								
			Exponencial		Matérn ( $\kappa = 1.5$ )		Gaussiano		Esférico		
$\tau^2$	$\sigma^2$	$\phi$	$n$	UK	DBUK	UK	DBUK	UK	DBUK	UK	DBUK
0	1	0.15	50	1.04	0.91	2.34	1.99	0.90	0.85	1.20	1.01
			100	0.88	0.82	1.98	1.87	0.74	0.70	1.02	0.93
			150	0.81	0.78	1.82	1.79	0.66	0.62	0.94	0.89
		0.60	50	0.83	0.75	2.06	1.65	0.73	0.69	0.85	0.81
			100	0.71	0.66	1.68	1.64	0.68	0.63	0.74	0.70
			150	0.65	0.63	1.55	1.59	0.64	0.65	0.68	0.66
	2	0.15	50	1.28	1.12	2.79	2.39	1.04	1.01	1.52	1.28
			100	1.07	1.01	2.34	2.21	0.81	0.79	1.29	1.19
			150	0.98	0.96	2.14	2.11	0.69	0.67	1.18	1.13
		0.60	50	0.93	0.84	2.26	1.83	0.74	0.71	0.99	0.93
			100	0.78	0.73	1.81	1.77	0.68	0.63	0.84	0.81
			150	0.72	0.69	1.67	1.74	0.66	0.64	0.77	0.75
1	1	0.15	50	1.55	1.30	3.27	2.74	1.47	1.30	1.63	1.35
			100	1.42	1.28	2.96	2.67	1.38	1.26	1.51	1.36
			150	1.39	1.31	2.86	2.70	1.33	1.26	1.47	1.38
		0.60	50	1.45	1.18	3.10	2.45	1.44	1.13	1.44	1.24
			100	1.32	1.18	2.80	2.46	1.42	1.16	1.33	1.22
			150	1.30	1.22	2.69	2.51	1.42	1.19	1.31	1.25
	2	0.15	50	1.72	1.46	3.61	3.08	1.58	1.45	1.89	1.58
			100	1.55	1.42	3.23	2.95	1.44	1.35	1.71	1.56
			150	1.50	1.43	3.09	2.95	1.36	1.31	1.65	1.56
		0.60	50	1.50	1.25	3.26	2.58	1.43	1.15	1.52	1.34
			100	1.36	1.24	2.90	2.56	1.40	1.17	1.40	1.30
			150	1.34	1.26	2.78	2.61	1.39	1.20	1.37	1.31

## 3.4 Aplicaciones

En las dos aplicaciones, se ajustan dos modelos para la media de  $Z(\mathbf{s})$ : el primero clásico UK usando (3.2) con las variables explicativas mixtas origi-

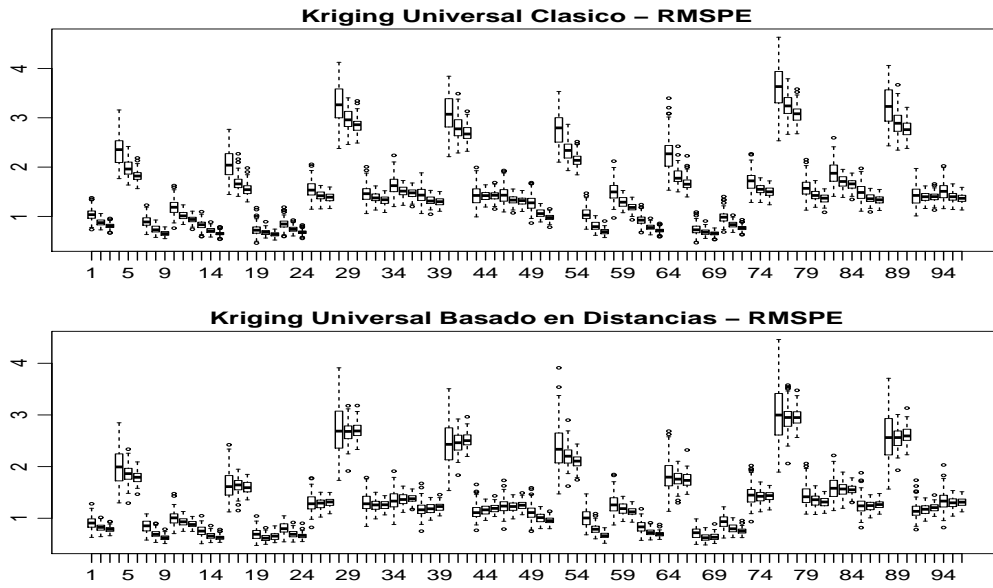
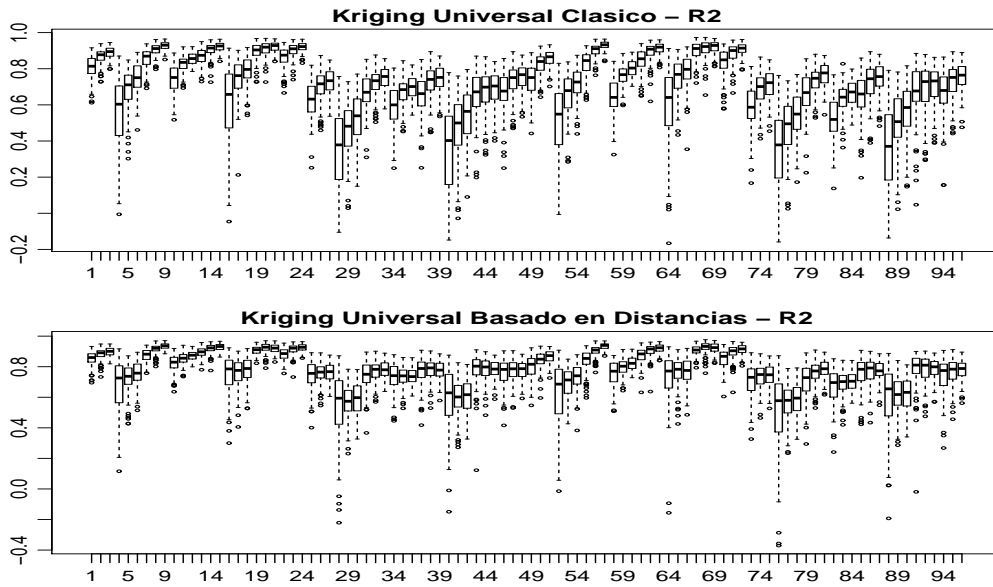


FIGURA 3.4: RMSPE para los escenarios considerados en el Caso 2

FIGURA 3.5:  $R^2$  para los escenarios considerados en el Caso 2

nales, y el segundo, DBUK usando (3.4) con las coordenadas principales generadas a partir de las variables mixtas. En el último método, tal como se hizo en el proceso de simulación, se emplea el criterio (3.5) para hacer la primera

selección, dejando afuera las coordenadas principales con correlación cercana a cero. Luego, se utiliza el criterio (3.6) para seleccionar las coordenadas principales más significativas en la regresión DB. Después, en ambos métodos, los residuales obtenidos de (3.2) y (3.4) son empleados para obtener los variogramas experimentales (clásico, robusto, mediana y media recortada) y sus correspondientes variogramas teóricos. Este procedimiento se realiza en diferentes direcciones para evaluar la isotropía, y en cada caso, se selecciona un modelo de variograma teórico (esférico, exponencial, Gaussiano y Matérn) compatible con el variograma experimental. Después de esto, con el fin de obtener los parámetros  $\hat{\boldsymbol{\vartheta}}$ , la estimación por OLS, WLS, ML y REML es realizada. Por lo tanto, se tiene un modelo de variograma,  $\hat{\gamma}(h; \hat{\boldsymbol{\vartheta}})$ , para la interpolación en los métodos UK y DBUK. Finalmente, en cada uno de los métodos, se hacen las predicciones en las localizaciones muestreadas y no muestreadas para la generación de los mapas de la variable analizada  $Z(\mathbf{s})$ .

### 3.4.1 Temperatura media diaria en Croacia

En 153 estaciones meteorológicas, la temperatura media diaria en Croacia fue medida el 1 de diciembre de 2008. Esta información es tomada de <http://spatial-analyst.net/book/HRclim2008> y esta fue proporcionada por Melita Perčec Tadić, de la Organización Meteorológica y de Servicios Hidrológicos Croatas (Hengl 2009). Croacia es un país relativamente pequeño, pero cuenta con varias regiones de clima diferente que son el resultado de su posición específica en el mar Adriático y de la topografía muy diversa que va desde las llanuras en el este, a través de una zona central montañosa que separa el territorio continental de la parte marítima del país. La región de estudio se caracteriza por una amplia gama de características topográficas y climáticas, lo que permite evaluar correctamente la metodología propuesta con respecto a la tradicional, ya que las temperaturas promedio de la tierra en tal región se ven fuertemente influenciadas por la topografía.

Las mediciones de temperatura se recogen automáticamente en 159 estaciones meteorológicas, pero dado que hay datos perdidos para el 1 de diciembre de 2008 se dispone de información sólo en 153 estaciones. En la mayoría de



las estaciones meteorológicas, la temperatura se mide tres veces al día, a las 7 am, 1 pm y 9 pm (Hengl et al. 2012). La media de la temperatura diaria ( $\Delta T$  en un día) se calcula como un promedio ponderado (Hiebl et al. 2009), de acuerdo a la siguiente expresión

$$\Delta T = \frac{T_{(7am)} + T_{(1pm)} + 2 \cdot T_{(9pm)}}{4}$$

La distribución espacial de las estaciones no es optima (Zaninovic et al. 2008), hay un cierto submuestreo a mayor altitud y en áreas con menor densidad de población; por razones prácticas, a las zonas de mayor densidad de población se les dio prioridad. Por lo tanto, se podría esperar que la precisión de la cartografía será menor a mayor altitud y en las tierras altas (Hengl 2009, Perčec Tadić 2010). Las coordenadas geográficas (latitud y longitud) fueron transformadas a un sistema de coordenadas cartesianas ( $w_x, w_y$ ). La ubicación de las 153 estaciones meteorológicas se muestra en la Figura 3.6. Las coordenadas principales fueron calculadas a partir de la descomposición espectral generada por las coordenadas espaciales,  $w_x$  y  $w_y$ . Las primeras dos coordenadas principales están altamente correlacionadas con la temperatura media de la tierra en Croacia, con una correlación al cuadrado superior a 0.18.

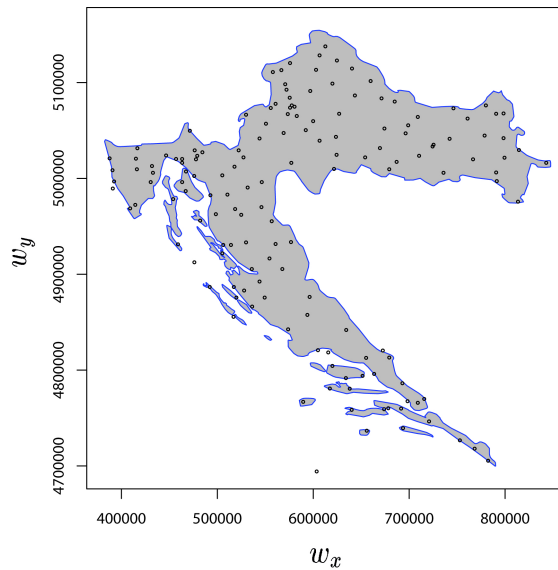


FIGURA 3.6: Localizaciones de las estaciones meteorológicas en Croacia

Con las dos coordenadas principales, se realizó una regresión lineal de primer orden teniendo en cuenta como variable respuesta la temperatura media de la tierra. Al mismo tiempo, en el modelo clásico, se realizó una regresión utilizando las coordenadas espaciales  $w_x$  y  $w_y$ , para lo cual se consideraron modelos lineales de orden uno y dos; en este caso, el modelo de orden 2 mostró un mejor ajuste. Posteriormente, se construyó un mapa del variograma ajustado a partir de los residuos obtenidos teniendo en cuenta el mejor modelo para los métodos clásicos y DB. Además, los variogramas ajustados obtenidos en las direcciones de 45 y 135 grados se muestran en la Figura 3.7; en ambos mapas de variograma de los residuales se observa un comportamiento anisotrópico.

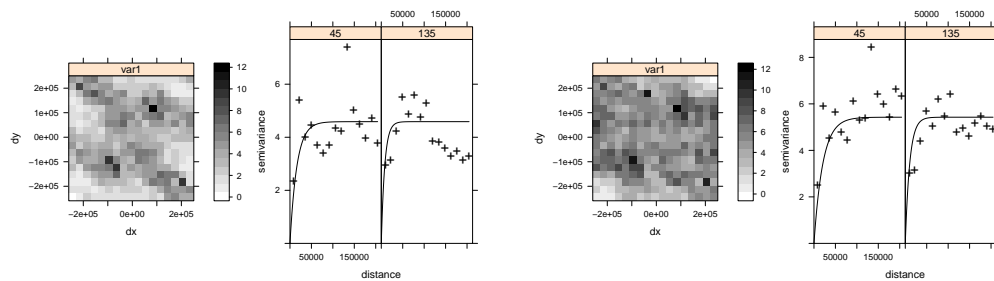


FIGURA 3.7: Mapas del variograma anisotrópico y modelos de variograma ajustados (azimut del semieje mayor es  $135^\circ$  y azimut del semieje menor es de  $45^\circ$ ) para los residuales de la temperatura media terrestre en los modelos clásico (dos paneles de izquierda) y DB (dos paneles de la derecha)

El variograma experimental asociado a los residuos que presentaron el mejor ajuste en los métodos clásicos y DB fue la media recortada, considerando un recorte del 10%. El modelo teórico ajustado fue en ambos casos el de Matérn con parámetros:  $\tau_1^2 = 1.197$ ,  $\sigma_1^2 = 5.637$ ,  $\phi_1 = 36775.12$  y  $k_1 = 0.5$  en el método clásico UK, y  $\tau_2^2 = 3.469$ ,  $\sigma_2^2 = 6.375$ ,  $\phi_2 = 176322.9$  y  $k_2=0.5$  en el método DBUK. Estos modelos mostraron el mas bajo cuadrado medio del error (Mean Square Error, MSE). Los parámetros fueron estimados utilizando los procedimientos de OLS, WLS y REML. En el caso de WLS, dos pesos se consideraron: la ponderación asignada por  $N_j/\gamma^2(h_j)$  y denotada por WLS, y los pesos dados por  $N_j/h_j^2$  y denotados por WLS1. WLS proporciona mejores

resultados que WLS1, con  $MSE = 1.035$  para los residuos en el modelo clásico, y  $MSE = 0.797$  para los residuos en DB (ver Figura 3.8).

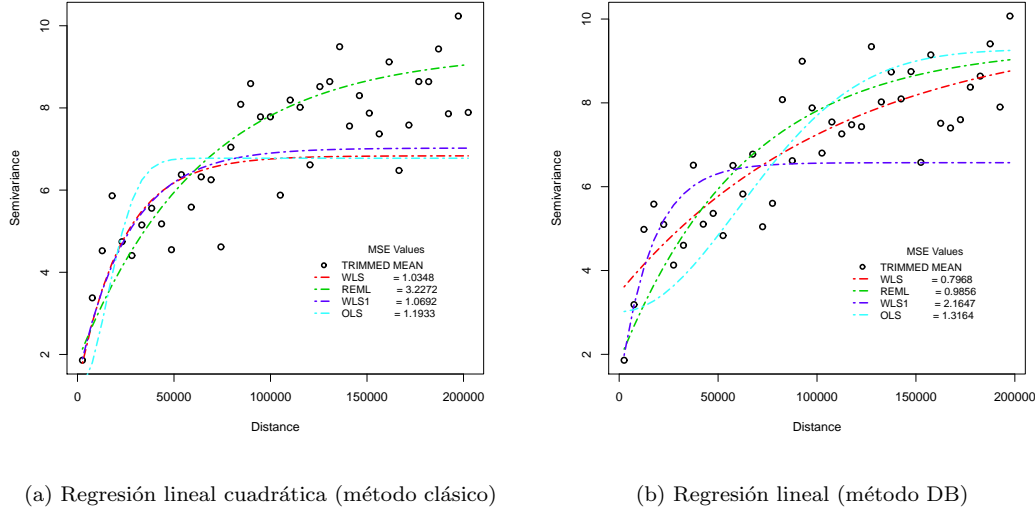


FIGURA 3.8: Variograma experimental de media recortada para los residuos, ajustando un modelo de Matérn por WLS, OLS y REML

Una vez que los variogramas fueron definidos, estos fueron utilizados en el krigado para la generación de los mapas de predicciones de la temperatura media terrestre y de las varianzas de los errores predichos. En el caso DBUK, se consideraron las coordenadas principales de la muestra y los puntos adicionales generados por el sistema de coordenadas original ( $w_x$  y  $w_y$ ). Posteriormente, se calcularon las predicciones teniendo en cuenta el sistema de coordenadas originales y los resultados obtenidos se presentan en la Figura 3.9, en cual se observa una alta coincidencia entre los métodos UK y DBUK.

La Figura 3.10 muestra los mapas de predicción de la varianza del error para los métodos UK y DBUK. Se observa que UK subestima la varianza en las zonas fronterizas. En general, la varianza del error de predicción para DBUK es más pequeño en toda la región de estudio.

Finalmente en esta aplicación, para evaluar las ventajas practicas de DBUK sobre UK con una tendencia modelada usando la regresión lineal, se realizaron 200 simulaciones sobre la muestra estudiada de las temperaturas media dia-

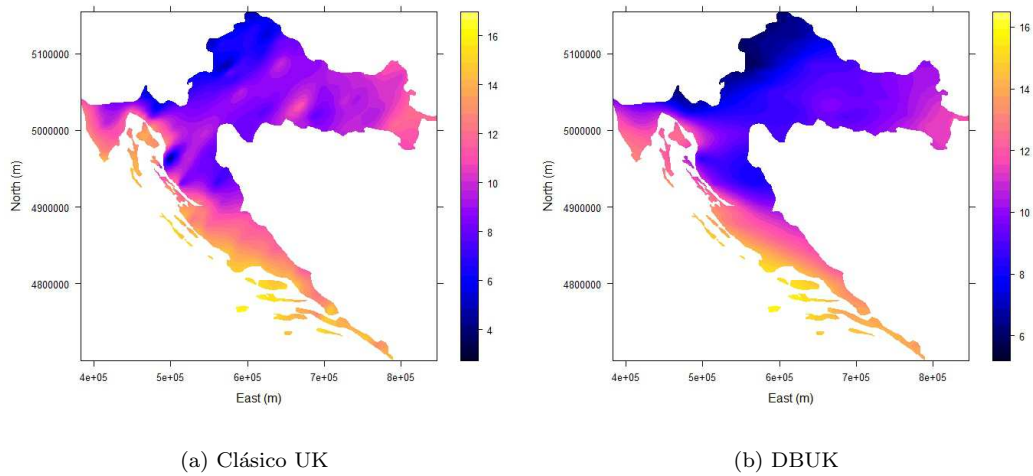


FIGURA 3.9: Mapas de predicción de la temperatura media diaria terrestre en Croacia

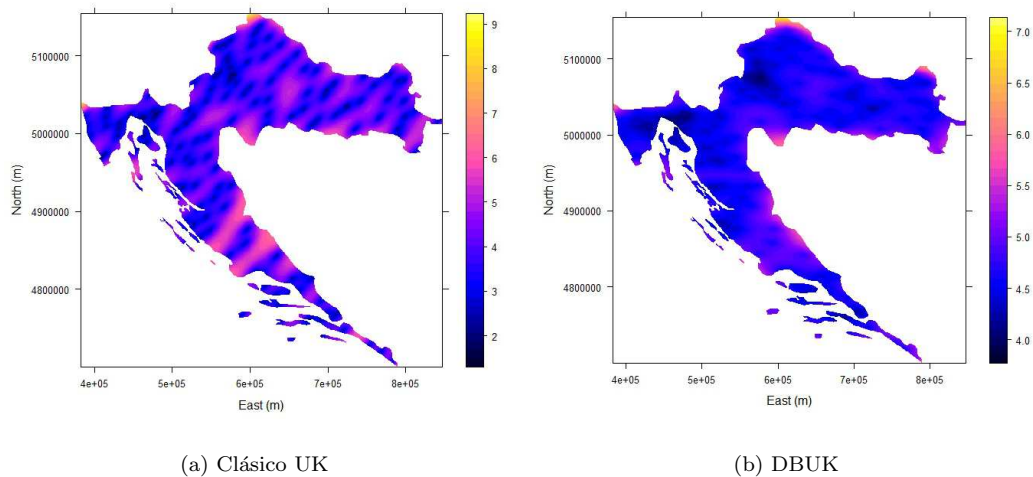


FIGURA 3.10: Mapas de predicción de las varianzas del error para la temperatura media diaria terrestre en Croacia

ria de las estaciones meteorológicas en Croacia. La muestra fue particionada en dos submuestras: el primer subconjunto de 110 datos fue utilizado para construir el modelo y el segundo subconjunto con 43 datos fue utilizado para evaluar el modelo, como se describe en (Bivand et al. 2008, Section 8.6). Las simulaciones consideraron 20 vecinos en la validación cruzada. Para los dos

métodos, se dejan fijos los valores empíricos de los parámetros asociados a los modelos de variograma.

En las 200 simulaciones, los promedios de RMSPEs, son 1.858 en UK y 1.820 en DBUK, y los promedios de  $R^2$  son 0.641 en UK y 0.657 en DBUK. Se nota una reducción del 3.8% para DBUK comparado con UK cuando se considera el RMSPE, y una ganancia en el  $R^2$  de 1.6%. Esta pequeña ganancia del DBUK con respecto al UK se obtiene debido a que sólo dos coordenadas principales fueron seleccionadas, pero si se hubiesen considerado más coordenadas principales, el método DBUK probablemente aumentaría su superioridad con respecto al método UK. Este hecho puede observarse en la siguiente aplicación del contenido de calcio.

### 3.4.2 Contenido de Calcio

En este conjunto de datos se consideran muestras de suelo recolectadas con una broca de tipo holandés en una malla regular incompleta a una distancia de aproximadamente 50 metros, con coordenadas geográficas: norte y este de 900 metros de distancia en ambas direcciones. Las muestras de suelo fueron tomadas de la capa de 0-20 cm de profundidad en cada una de las 178 localidades (ver Figura 3.11). El magnesio y el calcio se midieron en  $mmol_c/dm^3$ , pero en esta aplicación, se considera sólo el contenido de calcio. La región de estudio se dividió en tres sub-regiones ya que el muestreo regionalizado posibilita mapear la dirección de la variabilidad de las propiedades de textura y las propiedades químicas, lo que permite recortar el mapa de los rendimientos debido a las diferentes situaciones de la fertilidad del suelo y tipos de suelo. Por lo tanto, en estas sub-regiones se han experimentado los diferentes regímenes de manejo del suelo. Esta caracterización es ideal para aplicar el método propuesto en esta tesis.

En resumen, este conjunto de datos tiene información sobre el contenido de calcio, las coordenadas espaciales  $(w_x, w_y)$ , la altitud y la sub-región de cada muestra que se asocia con tres periodos de la fertilización en diferentes lugares o áreas.

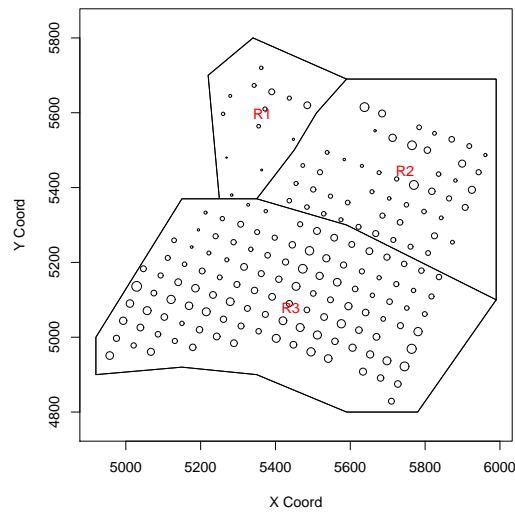


FIGURA 3.11: Gráfica de círculo de contenido de calcio con las líneas que delimitan las sub-regiones (lugares de muestreo)

Los datos son tomados de Capeche et al. (1997) y el principal objetivo del estudio fue la adecuada planificación del uso del suelo que permitiera una gestión racional y sostenible, evitando el proceso de erosión, con el fin de destinar subsidios en los campos experimentales para la realización de búsquedas que sean extrapoladas a suelos y zonas climáticas similares.

La Tabla 3.4 contiene los valores de los parámetros ajustados del variograma esférico y dos veces el log de la verosimilitud ( $2 \log L$ ), tanto para el método propuesto (DB) como para el método clásico. En el caso de DB, las coordenadas principales se construyeron utilizando las variables: coordenadas espaciales y la variable nominal que define la sub-región (la altitud no es considerada, al no existir información en los puntos no muestreados, esto con el fin de producir el mapa que se muestra más adelante). Los resultados presentados en la Tabla 3.4(a) muestran un aumento constante en  $2 \log L$  cuando se incrementa el número de coordenadas principales. Hay un aumento significativo en  $2 \log L$  ya que va desde -1272.03 hasta -1178.80 cuando el número de coordenadas principales va de 0 a 18. La Tabla 3.4(b) muestra los valores de  $2 \log L$  en el caso clásico; es claro que  $2 \log L$  aumenta cuando se considera la

variable de sub-región (tipo de suelo), pero no hay ganancia al adicionar las coordenadas espaciales o la altitud.

TABLA 3.4: Comparación entre los métodos DB y clásico con los valores de los parámetros ajustados del variograma esférico utilizando máxima verosimilitud

(a) Método DB				
Número de coordenadas principales con remoción tendencia DB	$\tau^2$	$\sigma^2$	$\phi$	$2\log L$
$\alpha_0$	23.23	111.69	244.90	-1272.03
2	0	87.08	107.65	-1260.00
4	0	80.56	102.50	-1253.61
8	0	67.48	89.39	-1236.01
17	0	51.29	83.11	-1193.20
18	0	46.98	81.52	-1178.80

(b) Método clásico				
Parámetros para remoción tendencia clásico	$\tau^2$	$\sigma^2$	$\phi$	$2\log L$
$\beta_0$	23.23	111.69	244.90	-1272.03
tipo de suelo	0	93.00	111.97	-1266.09
tipo de suelo, altitud	0	92.69	111.53	-1266.06
tipo de suelo, tendencia espacial lineal	0	87.53	107.45	-1261.18
tipo de suelo, altitud, tendencia espacial lineal	0	84.52	104.09	-1259.17

Para generar los mapas de contenido de calcio, se seleccionaron las variables: coordenadas espaciales (tendencia espacial lineal) y la sub-región (tipo de suelo). Además, con el fin de comparar los dos métodos (UK y DBUK), se consideran 17 coordenadas principales en el método DB (véase la Tabla 3.4(a)) para remover la tendencia porque las otras coordenadas principales obtenidas a partir del criterio (3.6) no fueron significativas a un nivel del 5%. Por otro lado, las variables explicativas (tipo de suelo y tendencia espacial lineal) se consideraron en el método clásico (véase la Tabla 3.4(b)). Una vez obtenidos los modelos de los variogramas, se llevan a cabo los métodos UK y DBUK, los cuales consideran la tendencia con las características mencionadas anteriormente para los dos métodos. Los mapas obtenidos se muestran en la Figura 3.12.

Las desviaciones estándar de los dos métodos analizados se muestran en la

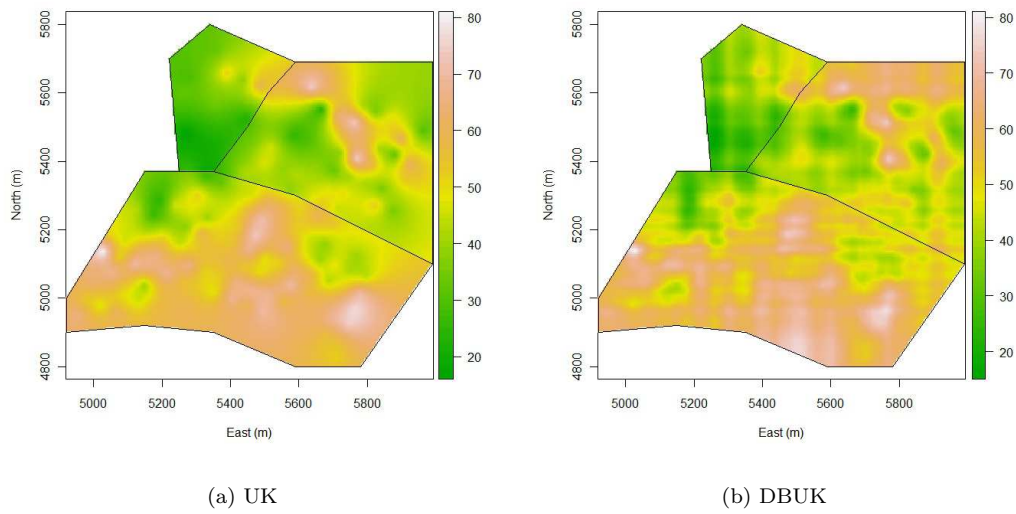


FIGURA 3.12: Mapas de predicción del contenido de calcio en el suelo incluyendo sub-región

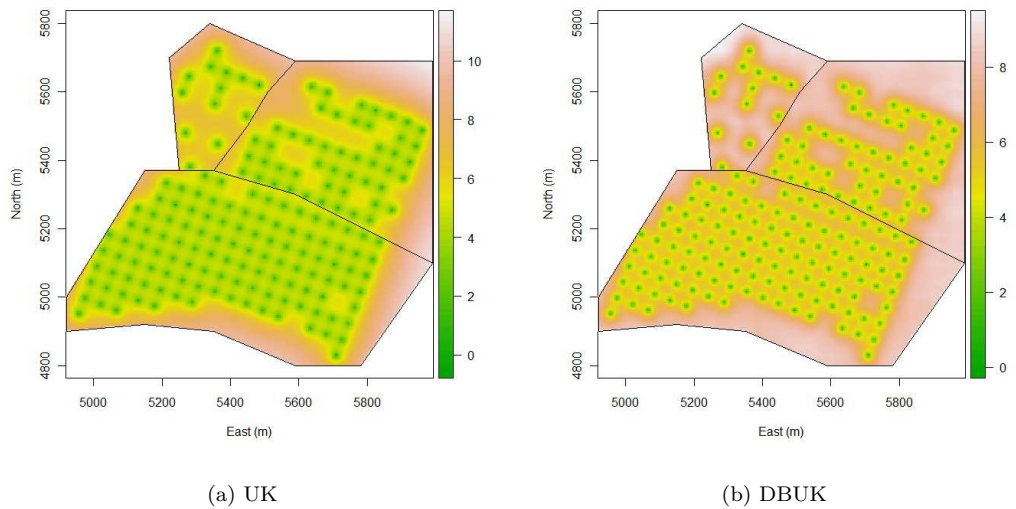


FIGURA 3.13: Mapas de predicción de los errores estándar para el contenido de calcio en el suelo, incluyendo sub-región

Figura 3.13. En esta, se observa que hay una reducción en las desviaciones estándar en el método DBUK con respecto al método UK. Los resultados de la validación cruzada se muestran en la Tabla 3.5, en donde se resalta un aumento



de alrededor del 10% del método DBUK propuesto sobre el método UK clásico.

TABLA 3.5: Comparación entre UK y DBUK para el contenido de calcio usando LOOCV

	UK	DBUK	DBUK-UK
RMSPE	7.734	7.011	-0.723
$R^2$	0.510	0.566	0.056

Una variedad de estudios para detectar la variabilidad entre regiones es prácticamente imposible, por lo cual se espera que el método propuesto sea útil en estos casos, ya que aprovecha al máximo la información existente. A pesar que la correlación sea baja con respecto a la variable a modelar, lo relevante en el método propuesto es la correlación entre las coordenadas principales (construidas con las variables existentes) y la variable respuesta espacial.

# Capítulo 4

## Modelo basado en distancias para la predicción espacial utilizando funciones de base radial

### 4.1 Introducción

Debido a que hoy en día existe un gran desarrollo de instrumentos de medición en tiempo real y de recursos de almacenamiento de datos, las funciones generadas a partir de experimentos aleatorios se pueden observar y procesar. De esta manera, los métodos globales (tales como el análisis de tendencias de superficie) utilizan todos los datos disponibles para la predicción, mientras que los métodos locales como las funciones de base radial (radial basis functions, RBF), los kriging y la distancia inversa ponderada, suelen utilizar sólo un subconjunto de los datos para hacer cada predicción. Una de las ventajas de los métodos locales es que el tiempo de cálculo se reduce en la predicción, al trabajar con los datos asociados a vecindarios. Algunos métodos hacen uso de todos los datos disponibles, pero únicamente tienen en cuenta las distancias a partir de la localización de la predicción. Estos métodos todavía se pueden considerar locales, es decir que muchas de las técnicas de interpolación utilizadas son

métodos locales.

Las funciones de base radial (RBF) tales como la multicuadrática (MQ) o completamente regularizada spline (CRS) son útiles en la construcción de modelos digitales de elevación (DEM), como se muestra en Mitášová & Hofierka (1993), en el que se incorpora el spline en un sistema de información geográfica para estudiar la erosión del suelo. Una variación de la función multicuadrática se llama la función inversa multicuadrática (IMQ), introducida por Hardy & Gopfert (1975). Luego, el spline capa delgada (TPS) fue introducido en el diseño geométrico por Duchon (1976). El nombre TPS se refiere a una analogía física que implica la flexión de una hoja delgada de metal. (Franke 1982) desarrolló un programa de ordenador para la solución del problema de interpolación de datos dispersos; el algoritmo se basa en una suma ponderada de TPS definidos localmente, obteniéndose una función de interpolación que es diferenciable. Más tarde, Thiébaux & Pedder (1987) describió la TPS como una superficie de dos dimensiones llamada spline cúbico. Otra variante popular de la TPS es la aproximación Gaussiana (GAU) utilizada por Schagen (1979). Otra función de base radial es la interpolación spline cúbica y exponencial (EXP) que permite evitar los puntos de inflexión y contiene splines cúbicos como un caso especial (Späh 1969). Por último, Mitáš & Mitášová (1988), Mitášová & Hofierka (1993) y Mitášová & Mitáš (1993) desarrollan la formulación de spline con tensión (ST) e implementan un algoritmo de segmentación con un tamaño flexible de la superposición del vecindario.

El enlace entre splines y kriging fue llamado equivalentemente “cercano” (Cressie 1989) porque el TPS corresponde a una covarianza generalizada específica, mientras que el estimador kriging y el interpolador RBF sólo requieren el uso de un kernel con propiedades adecuadas como la de definida positiva. En general, esto permite adaptar la función kernel a un conjunto de datos particular (Cressie 1989, Myers 1992). La mayor diferencia es que el usuario establece el parámetro de suavizamiento en los splines, mientras en el caso de kriging, el suavizamiento se determina de forma objetiva.

Investigaciones recientes utilizan RBF sobre dominios irregulares en dos dimensiones a través del proceso de conformación de trasplante (Heryudono &

Driscoll 2010). Zhang (2011) desarrolla un algoritmo rápido para el estimador de suavizado spline univariado en una regresión multivariante mediante el uso de funciones de base radial de soporte compacto. Yavuz & Erdoğan (2012) realiza un análisis de tendencias de las precipitaciones mensuales y anuales, utilizando los métodos de interpolación kriging ordinario, distancia inversa ponderada y spline completamente regularizado. Estos estudios demuestran la utilidad de trabajar con RBF.

Además, en los estudios anteriormente presentados con frecuencia se tiene que lidiar con variables explicativas de diferente naturaleza asociadas con una variable respuesta espacial. Dichas variables independientes pueden ser: categóricas, binarias y continuas; sin embargo, los métodos mencionados anteriormente no son totalmente apropiados cuando se modela una mezcla de variables explicativas.

Por lo tanto, el objetivo en este capítulo es presentar un enfoque unificado que utiliza RBF en donde las variables explicativas son de naturaleza mixta. En este sentido, se propone un nuevo método utilizando distancias entre los individuos, tales como la distancia de Gower (1968), aunque alguna otra distancia Euclidiana se puede utilizar. Por consiguiente, el método de interpolación espacial basada en distancias con funciones de base radial (distance-based spatial interpolation with radial basis functions, DBSIRBF) se aplica en el modelo geoestadístico para predecir la tendencia y estimar la estructura de covarianza cuando las variables explicativas son mixtas. La tendencia se incorpora en una RBF de acuerdo con un procedimiento de eliminación de la tendencia.

Este capítulo se desarrolla de la siguiente forma: en la Sección 4.2 se desarrolla la propuesta metodológica introduciendo la tendencia lineal local basada en distancias, se construyen las RBFs a partir de la tendencia basada en distancias, se describen algunas RBFs y se hace una aproximación a partir de la interpolación spline al método kriging para la predicción. Con el fin de evaluar la eficacia del método propuesto, en la Sección 4.3 se lleva a cabo un estudio de simulación para una variedad de escenarios prácticos que incluyen cinco funciones distintas de base radial e incorpora las coordenadas principales. Por último, en la Sección 4.4 se ilustra el método propuesto con una aplicación de

la predicción de concentración de calcio medido a una profundidad de 0-20 cm en Brasil, seleccionando el parámetro de suavizamiento mediante validación cruzada.

## 4.2 Modelo geoestadístico basado en distancias con funciones de base radial

Supóngase que se está interesado en relacionar una variable respuesta continua con variables georeferenciadas explicativas medidas en cada sitio de muestreo, estas variables pueden ser del tipo: latitud y longitud, binarias, categóricas y continuas. Al igual que en la Sección 3.2, sea  $\mathbf{s} \in \mathbb{R}^d$  una ubicación en un espacio Euclidiano  $d$ -dimensional y supóngase que  $Z(\mathbf{s})$  es un vector aleatorio en cada ubicación espacial  $\mathbf{s}$ . Realizando el mismo procedimiento que se presenta en la Sección 3.2, la idea es hacer una transformación de las variables explicativas utilizando el método basado en distancias. Para ello se definen las medidas de similitud o distancia Euclidiana presentadas en la Subsección 2.7.1, que dependen de las características de las variables explicativas. Una vez seleccionada alguna de las distancias presentadas allí, se realiza el proceso de descomposición espectral y se selecciona las coordenadas principales que más se relacionan con la variable respuesta, realizando cualquiera de los cuatro métodos presentados al final de la Sección 3.2. Por lo tanto, las  $X_{k+1}, \dots, X_{n-1}$  coordenadas principales deben ser removidas ya que son las menos relevantes.

Por otra parte, en la interpolación espacial, existen métodos que no requieren información de un modelo de dependencia espacial, tales como el variograma o covariograma, éstos se llaman deterministas y son los de interés en esta sección. El modelo (3.1) utilizando un formato basado en distancias se puede expresar en forma general por

$$Z(\mathbf{s}_i) = g(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \quad i = 1, \dots, n \quad (4.1)$$

donde  $g(\mathbf{s}_i)$  es una función de valor-real, dada por

$$g(\mathbf{s}_i) = \sum_{l=0}^k \nu_l f_l(\mathbf{s}_i) + \sum_{j=1}^n \omega_j \phi(\mathbf{s}_i - \mathbf{s}_j), \quad i = 1, \dots, n$$

o en forma matricial,

$$\mathbf{g}_s = F_s \boldsymbol{\nu}_s + \Phi_s \boldsymbol{\omega}_s \quad (4.2)$$

donde  $\mathbf{g}_s = (g(\mathbf{s}_1), \dots, g(\mathbf{s}_n))'$ ,  $F_s = (\mathbf{1}, F_1, \dots, F_k)$  es una matriz  $n \times (k + 1)$  con elementos  $\mathbf{1}$  y  $F_l = (f_l(\mathbf{s}_1), \dots, f_l(\mathbf{s}_n))'$ ,  $l = 1, \dots, k$ , y con cada  $f_l(\mathbf{s}_i)$  una función de valor real;  $\boldsymbol{\nu}_s = (\nu_0, \nu_1, \dots, \nu_k)'$  donde cada  $\nu_l$  corresponde al  $l$ -ésimo coeficiente del modelo de tendencia;  $\Phi_s$  es una matriz  $n \times n$  con elementos  $\phi(\mathbf{s}_i - \mathbf{s}_j)$ , el cual es una función de base radial, es decir una función escalar de la distancia Euclidiana entre  $\mathbf{s}_i$  y  $\mathbf{s}_j$ ; finalmente,  $\boldsymbol{\omega}_s = (\omega_1, \dots, \omega_n)'$ , con  $\omega_i$  un peso desconocido.

Los parámetros  $\boldsymbol{\nu}_s$  y  $\boldsymbol{\omega}_s$  pueden ser estimados por mínimos cuadrados penalizados, minimizando la siguiente expresión

$$\sum_{i=1}^n [Z(\mathbf{s}_i) - g(\mathbf{s}_i)]^2 + \rho \int_{\mathbb{R}^2} J_m(g(\mathbf{s})) d\mathbf{s} \quad (4.3)$$

donde  $J_m(g(\mathbf{s}))$  es una medida de la rugosidad de la función spline  $g$  (definida en términos de las  $m$ -ésimas derivadas de  $g$ ) y  $\rho > 0$  actúa como un parámetro de suavizamiento.

La expresión (4.3) se puede expresar al hacer los respectivos reemplazos como

$$\begin{aligned} L(\boldsymbol{\nu}_s, \boldsymbol{\omega}_s) &= (\mathbf{Z}_s - F_s \boldsymbol{\nu}_s - \Phi_s \boldsymbol{\omega}_s)' (\mathbf{Z}_s - F_s \boldsymbol{\nu}_s - \Phi_s \boldsymbol{\omega}_s) + \rho \int_{\mathbb{R}^2} [g''(\mathbf{s})]^2 d\mathbf{s} \\ &= (\mathbf{Z}_s - F_s \boldsymbol{\nu}_s - \Phi_s \boldsymbol{\omega}_s)' (\mathbf{Z}_s - F_s \boldsymbol{\nu}_s - \Phi_s \boldsymbol{\omega}_s) + \rho \|P_s \hat{g}_s\|^2 \end{aligned}$$

donde  $P_s$  es el espacio que genera  $\Phi_s$  y  $\|P_s \hat{g}_s\|^2 = \langle P_s \hat{g}_s, P_s \hat{g}_s \rangle = \boldsymbol{\omega}'_s \mathbf{q}_s \mathbf{q}'_s \boldsymbol{\omega}_s = \boldsymbol{\omega}'_s \Phi_s \boldsymbol{\omega}_s$  con  $\Phi_s = \mathbf{q}_s \mathbf{q}'_s$ . Por lo tanto,

$$\begin{aligned} L(\boldsymbol{\nu}_s, \boldsymbol{\omega}_s) &= (\mathbf{Z}_s - F_s \boldsymbol{\nu}_s - \Phi_s \boldsymbol{\omega}_s)' (\mathbf{Z}_s - F_s \boldsymbol{\nu}_s - \Phi_s \boldsymbol{\omega}_s) + \rho \boldsymbol{\omega}'_s \Phi_s \boldsymbol{\omega}_s \\ &= \mathbf{Z}'_s \mathbf{Z}_s - 2 \mathbf{Z}'_s F_s \boldsymbol{\nu}_s - 2 \mathbf{Z}'_s \Phi_s \boldsymbol{\omega}_s + \boldsymbol{\nu}'_s F'_s F_s \boldsymbol{\nu}_s + 2 \boldsymbol{\omega}'_s \Phi'_s F_s \boldsymbol{\nu}_s \\ &\quad + \boldsymbol{\omega}'_s \Phi'_s \Phi_s \boldsymbol{\omega}_s + \rho \boldsymbol{\omega}'_s \Phi_s \boldsymbol{\omega}_s \end{aligned}$$

Al derivar parcialmente con respecto a los vectores  $\boldsymbol{\nu}_s$  y  $\boldsymbol{\omega}_s$  e igualar a cero,

se encuentra que

$$\begin{aligned} \frac{\partial L(\boldsymbol{\nu}_s, \boldsymbol{\omega}_s)}{\partial \boldsymbol{\nu}_s} &= -2F'_s \mathbf{Z}_s + 2F'_s F_s \boldsymbol{\nu}_s + 2F'_s \Phi_s \boldsymbol{\omega}_s = 0 \\ F_s \boldsymbol{\nu}_s + \Phi_s \boldsymbol{\omega}_s &= \mathbf{Z}_s \end{aligned} \quad (4.4)$$

$$\begin{aligned} \frac{\partial L(\boldsymbol{\nu}_s, \boldsymbol{\omega}_s)}{\partial \boldsymbol{\omega}_s} &= -2\Phi'_s \mathbf{Z}_s + 2\Phi'_s F_s \boldsymbol{\nu}_s + 2\Phi'_s \Phi_s \boldsymbol{\omega}_s + 2\rho \Phi'_s \boldsymbol{\omega}_s = 0 \\ F_s \boldsymbol{\nu}_s + (\Phi_s + \rho I) \boldsymbol{\omega}_s &= \mathbf{Z}_s \end{aligned} \quad (4.5)$$

donde  $I$  es la matriz identidad de orden  $n \times n$  y  $\rho$  puede ser interpretado como ruido blanco adicionado a las varianzas en las localizaciones de los datos, pero no la varianza en la localización donde se predice (Wackernagel 2003).

Noté aquí que si  $\Phi_s$  es definida positiva entonces hay unicidad de los coeficientes en el interpolador  $\hat{g}(\mathbf{s}_i)$ . Con el fin de generalizar los interpoladores, es necesario considerar las formas más generales de las matrices definidas positivas.

**Definición 4.1.** Sean  $f_0, f_1, \dots, f_k$  funciones linealmente independientes de valor-real definidas sobre  $\mathbb{R}^d$  y  $\Phi_s$  una matriz simétrica real. Luego,  $\Phi_s$  es definida positiva con respecto a  $f_0, f_1, \dots, f_k$  si y sólo si para todos los conjuntos de puntos  $\mathbf{s}_1, \dots, \mathbf{s}_n$  en  $\mathbb{R}^d$  se tiene que  $\sum_{i=1}^n \sum_{j=1}^n q_i q_j \phi(\mathbf{s}_i - \mathbf{s}_j) \geq 0$  para todo  $q_i$  ( $i = 1, \dots, n$ ), donde  $q_i$  es un escalar (no todos cero), y tales que  $\sum_{j=1}^n f_l(\mathbf{s}_j) q_j = 0$  para  $l = 1, \dots, k$ .

Como por la definición 4.1,  $\Phi_s$  es definida positiva, entonces  $\Phi_s + \rho I$  es invertible y así (4.5) se puede escribir como:

$$\boldsymbol{\omega}_s = (\Phi_s + \rho I)^{-1} (\mathbf{Z}_s - F_s \boldsymbol{\nu}_s) \quad (4.6)$$

Reemplazando (4.6) en (4.4) se obtiene

$$\begin{aligned} F_s \boldsymbol{\nu}_s + \Phi_s (\Phi_s + \rho I)^{-1} (\mathbf{Z}_s - F_s \boldsymbol{\nu}_s) &= \mathbf{Z}_s \\ [I - \Phi_s (\Phi_s + \rho I)^{-1}] F_s \boldsymbol{\nu}_s &= [I - \Phi_s (\Phi_s + \rho I)^{-1}] \mathbf{Z}_s \end{aligned} \quad (4.7)$$

Observe que

$$[I - \Phi_s (\Phi_s + \rho I)^{-1}] = \left( I + \frac{1}{\rho} \Phi_s \right)^{-1} = \rho (\Phi_s + \rho I)^{-1} \quad (4.8)$$

Premultiplicando por  $F'_s$  la expresión (4.7) y reemplazando por (4.8), se encuentra que

$$\hat{\boldsymbol{\nu}}_s = [F'_s(\Phi_s + \rho I)^{-1}F_s]^{-1} F'_s(\Phi_s + \rho I)^{-1} \mathbf{Z}_s \quad (4.9)$$

y reemplazando (4.8) en (4.6), se encuentra finalmente que

$$\hat{\boldsymbol{\omega}}_s = (\Phi_s + \rho I)^{-1} \left\{ I - F_s [F'_s(\Phi_s + \rho I)^{-1}F_s]^{-1} F'_s(\Phi_s + \rho I)^{-1} \right\} \mathbf{Z}_s \quad (4.10)$$

Al premultiplicar (4.10) se obtiene

$$F'_s \boldsymbol{\nu}_s = \mathbf{0}$$

y al combinarlo con el sistema (4.5), se encuentra que  $(\boldsymbol{\omega}_s, \boldsymbol{\nu}_s)$  son la solución del siguiente sistema de ecuaciones lineales

$$\begin{pmatrix} \Phi_s + \rho I & F_s \\ F'_s & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}_s \\ \boldsymbol{\nu}_s \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_s \\ \mathbf{0} \end{pmatrix}$$

Si no hay tendencia,  $F_s$  se convierte en un vector de unos y  $\boldsymbol{\nu}_s$  en un parámetro de sesgo.

Finalmente, en esta subsección se presenta en la Tabla 4.1 algunas RBFs consideradas en esta investigación y que utilizan el enfoque basado en distancias. El parámetro de suavizamiento óptimo  $\eta$ , el cual es un parámetro de libre elección, se encuentra al minimizar la raíz del cuadrado medio del error de predicción (RMSPE) haciendo uso de la validación cruzada. Algunas descripciones adicionales de RBFs y sus relaciones con los splines y kriging se pueden encontrar en Bishop (1995, p. 164), Chilès & Delfiner (1999, pag. 272) y Cressie (1993, pag. 180).

### 4.2.1 Predicción espacial basada en distancias con funciones de base radial

Una vez se han estimado los parámetros  $\boldsymbol{\nu}_s$  y  $\boldsymbol{\omega}_s$ , se pueden discutir las técnicas espaciales para predecir el valor de un campo aleatorio en una nueva localización espacial,  $\mathbf{s}_0$ , a partir de las observaciones cercanas, y en donde se



TABLA 4.1: Formas funcionales de algunas RBFs

RBF	Forma funcional	RBF	Forma funcional
EXP	$\phi(\delta) = e^{-\eta\delta}, \quad \eta \neq 0$	GAU	$\phi(\delta) = e^{-\eta\delta^2}, \quad \eta \neq 0$
MQ	$\phi(\delta) = \sqrt{\eta^2 + \delta^2}, \quad \eta \neq 0$	IMQ	$\phi(\delta) = 1/\sqrt{\eta^2 + \delta^2}, \quad \eta \neq 0$
RBF	Forma funcional		
TPS	$\phi(\delta) = \begin{cases} (\eta \cdot \delta)^2 \log(\eta \cdot \delta) & \text{si } \delta \neq 0, \eta > 0 \\ 0 & \text{si } \delta = 0 \end{cases}$		
CRS	$\phi(\delta) = \begin{cases} \ln(\eta \cdot \delta/2)^2 + E_1(\eta \cdot \delta/2)^2 + C_E & \text{si } \delta \neq 0, \eta > 0 \\ 0 & \text{si } \delta = 0 \end{cases}$ <p>donde <math>\ln</math> es el logaritmo natural, <math>E_1(x)</math> es la función integral exponencial y <math>C_E</math> es la constante de Euler.</p>		
ST	$\phi(\delta) = \begin{cases} \ln(\eta \cdot \delta/2) + K_0(\eta \cdot \delta) + C_E & \text{si } \delta \neq 0 \\ 0 & \text{si } \delta = 0 \end{cases}$ <p>donde <math>K_0(x)</math> es la función modificada de Bessel y <math>C_E</math> es la constante de Euler.</p>		

observan un conjunto de variables explicativas mixtas. Para conseguirlo, se utiliza el método de kriging universal con la finalidad de construir a partir de la tendencia basada en distancias las predicciones espaciales.

Por lo tanto, al igual que en la Sección 3.2, supóngase que un nuevo individuo ( $n + 1$ ) es observado con sus respectivas variables explicativas mixtas, es decir  $v(\mathbf{s}_0) = (v_1(\mathbf{s}_0), \dots, v_p(\mathbf{s}_0))'$  es conocido. Entonces, las distancias entre el nuevo individuo y cada uno de los individuos involucrados en el modelo propuesto en (3.2) se pueden calcular como  $\delta_{0i} = \delta(v(\mathbf{s}_0), v(\mathbf{s}_i))$ ,  $i = 1, \dots, n$ . A partir de estas distancias, se puede hacer una predicción usando un resultado de Gower (1971) y Cuadras & Arenas (1990), que relaciona el vector  $\boldsymbol{\delta}_0 = (\delta_{01}^2, \dots, \delta_{0n}^2)'$  de cuadrados de las distancias con el vector  $x(\mathbf{s}_0) = (x_1(\mathbf{s}_0), \dots, x_k(\mathbf{s}_0))'$  de coordenadas principales asociado al nuevo in-

dividido mediante la expresión

$$\delta_{0i}^2 = [x(\mathbf{s}_0) - x(\mathbf{s}_i)]' [x(\mathbf{s}_0) - x(\mathbf{s}_i)]$$

con  $i = 1, \dots, n$ . Luego, se encuentra que

$$x(\mathbf{s}_0) = \frac{1}{2} \Lambda^{-1} X'(b - \delta_0)$$

donde  $b = (b_{11}, \dots, b_{nn})'$  y  $b_{ii} = x(\mathbf{s}_i)'x(\mathbf{s}_i)$ ,  $i = 1, \dots, n$ .

Ahora, el próximo objetivo es predecir el valor de  $Z(\mathbf{s}_0)$  basado en un conjunto de observaciones  $\mathbf{Z}_s$ . Para esto, el predictor de la función de base radial esta dado por

$$\hat{Z}(\mathbf{s}_0) = \hat{g}(\mathbf{s}_0) = \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i) = \boldsymbol{\varphi}'_s \mathbf{Z}_s \quad (4.11)$$

sujeto a la condición

$$\sum_{i=1}^n \varphi_i f_l(\mathbf{s}_i) = \boldsymbol{\varphi}'_s \mathbf{f}_s = f_l(\mathbf{s}_0), \quad l = 0, \dots, k$$

donde  $\boldsymbol{\varphi}_s = (\varphi_1, \dots, \varphi_n)'$ ,  $\mathbf{Z}_s = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$  y  $\mathbf{f}_s = (f_l(\mathbf{s}_1), \dots, f_l(\mathbf{s}_n))'$ .

El error esperado es igual a cero, es decir,

$$E\left(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)\right) = 0$$

y el error cuadrático medio de la predicción del krigado,  $\sigma_K^2$ , al utilizar la aproximación con funciones de base radial esta dado por

$$\begin{aligned} \sigma_K^2(\mathbf{s}_0) &= E\left\{\left[\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)\right]^2\right\} \\ &\cong -\sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \phi(\mathbf{s}_i - \mathbf{s}_j) + 2 \sum_{i=1}^n \varphi_i \phi(\mathbf{s}_i - \mathbf{s}_0) \\ &\cong -\boldsymbol{\varphi}'_s \Phi_s \boldsymbol{\varphi}_s + 2\boldsymbol{\varphi}'_s \boldsymbol{\phi}_0 \end{aligned} \quad (4.12)$$

donde  $\boldsymbol{\phi}_0 = (\phi(\mathbf{s}_1 - \mathbf{s}_0), \dots, \phi(\mathbf{s}_n - \mathbf{s}_0))'$  corresponde al vector de función de base radial evaluado entre los vecinos y el punto donde se quiere predecir, es decir  $\phi(\mathbf{s}_i - \mathbf{s}_0)$ .

Los pesos se determinan minimizando la siguiente expresión penalizada

$$l(\boldsymbol{\varphi}_s, \boldsymbol{\alpha}_s) = \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \phi(\mathbf{s}_i - \mathbf{s}_j) - 2 \sum_{i=1}^n \varphi_i \phi(\mathbf{s}_i - \mathbf{s}_0) + \rho \int_{\mathbb{R}^2} J_m(g(\mathbf{s})) d\mathbf{s} \\ + 2 \sum_{l=0}^k \alpha_l \left( \sum_{i=1}^n \varphi_i f_l(\mathbf{s}_i) - f_l(\mathbf{s}_0) \right)$$

o equivalentemente, la expresión anterior se convierte en

$$l(\boldsymbol{\varphi}_s, \boldsymbol{\alpha}_s) = \boldsymbol{\varphi}'_s (\Phi_s + \rho I) \boldsymbol{\varphi}_s - 2\boldsymbol{\varphi}'_s \boldsymbol{\phi}_0 + 2\boldsymbol{\alpha}'_s (F'_s \boldsymbol{\varphi}_s - \mathbf{f}_s(\mathbf{s}_0))$$

donde  $\boldsymbol{\alpha}_s = (\alpha_0, \dots, \alpha_k)'$  es el vector de  $k + 1$  multiplicadores de Lagrange asociado con la restricción insesgamiento,  $F_s$  fue definida en (4.2) y  $\mathbf{f}_s(\mathbf{s}_0) = (f_0(\mathbf{s}_0), \dots, f_k(\mathbf{s}_0))'$ .

Después de diferenciar con respecto a  $\boldsymbol{\varphi}_s$  y  $\boldsymbol{\alpha}_s$ , igualando el resultado a cero y realizando algunos procedimientos algebraicos, el siguiente sistema matricial se encuentra

$$\begin{pmatrix} \Phi_s + \rho I & F_s \\ F'_s & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\varphi}_s \\ \boldsymbol{\alpha}_s \end{pmatrix} = \begin{pmatrix} \boldsymbol{\phi}_0 \\ \mathbf{f}_s(\mathbf{s}_0) \end{pmatrix} \quad (4.13)$$

Resolviendo el sistema, los coeficientes para  $\boldsymbol{\varphi}_s$  y  $\boldsymbol{\alpha}_s$  están dados por

$$\hat{\boldsymbol{\varphi}}'_s = \left\{ \boldsymbol{\phi}_0 + F_s [F'_s(\Phi_s + \rho I)^{-1} F_s]^{-1} [\mathbf{f}_s(\mathbf{s}_0) - F'_s(\Phi_s + \rho I)^{-1} \boldsymbol{\phi}_0] \right\}' (\Phi_s + \rho I)^{-1} \\ \hat{\boldsymbol{\alpha}}_s = - [F'_s(\Phi_s + \rho I)^{-1} F_s]^{-1} [\mathbf{f}_s(\mathbf{s}_0) - F'_s(\Phi_s + \rho I)^{-1} \boldsymbol{\phi}_0] \quad (4.14)$$

Por otro lado, para obtener una expresión aproximada del error cuadrático de la predicción, se premultiplica la parte superior de (4.13) por  $\boldsymbol{\varphi}'_s$  y se encuentra que  $\boldsymbol{\varphi}'_s (\Phi_s + \rho I) \boldsymbol{\varphi}_s + \boldsymbol{\varphi}'_s F_s \boldsymbol{\alpha}_s = \boldsymbol{\varphi}'_s \boldsymbol{\phi}_0$ , éste término se reemplaza en la expresión (4.12) y se llega a

$$\begin{aligned} \sigma_K^2(\mathbf{s}_0) &\cong - \boldsymbol{\varphi}'_s \Phi_s \boldsymbol{\varphi}_s + 2\boldsymbol{\varphi}'_s \boldsymbol{\phi}_0 \\ &\cong - \boldsymbol{\varphi}'_s \boldsymbol{\phi}_0 + \rho \boldsymbol{\varphi}'_s \boldsymbol{\varphi}_s + \boldsymbol{\varphi}'_s F_s \boldsymbol{\alpha}_s + 2\boldsymbol{\varphi}'_s \boldsymbol{\phi}_0 \\ &\cong \boldsymbol{\varphi}'_s \boldsymbol{\phi}_0 + \rho \boldsymbol{\varphi}'_s \boldsymbol{\varphi}_s + \mathbf{f}'_s(\mathbf{s}_0) \boldsymbol{\alpha}_s \end{aligned}$$

donde  $F'_s \boldsymbol{\varphi}_s = \mathbf{f}'_s(\mathbf{s}_0)$ .

Una vez estimados  $\varphi_s$  y  $\alpha_s$  en (4.11), una expresión aproximada del error cuadrático de la predicción estimado se puede escribir como

$$\hat{\sigma}_K^2(\mathbf{s}_0) \cong \sum_{i=1}^n \hat{\varphi}_i \phi(\mathbf{s}_i - \mathbf{s}_0) + \rho \sum_{i=1}^n \hat{\varphi}_i^2 + \sum_{l=0}^k \hat{\alpha}_l f_l(\mathbf{s}_0) \quad (4.15)$$

El procedimiento presentado en esta sección se puede resumir en los siguientes pasos:

1. Obtener las coordenadas principales utilizando la descomposición espectral de la matriz de similitudes (o distancias) calculada a partir de las variables explicativas.
2. Seleccionar las coordenadas principales más correlacionadas o significativas con la variable regionalizada  $\mathbf{Z}_s$ . En este paso, se recomienda utilizar el criterio dado en (3.5) para hacer una primera selección con el fin de remover las coordenadas principales pobremente correlacionadas con las variable regionalizada, y luego, emplear los criterios (3.6) o (3.7) para seleccionar las coordenadas principales mas significativas utilizando la regresión DB.
3. Optimizar los parámetros  $\eta$  del interpolador espacial basado en distancias con funciones de base radial (DBSIRBF) y  $\rho$ , mediante el uso del RMSPE establecido en la expresión (3.16) por medio de la validación cruzada (leave-one-out), empleando las expresiones (4.9) y (4.10) en los diferentes vecindarios de un tamaño prefijado. El tamaño del vecindario,  $n_h$ , también se puede escoger dentro del mismo proceso de optimización.
4. Hacer las predicciones en los puntos muestreados y no muestreados para generar el mapa de predicción usando el método DBSIRBF, es decir, haciendo  $\hat{Z}(\mathbf{s}_0) = \hat{\varphi}'_s \mathbf{Z}_s$ .

En el caso en que se desee evaluar el ajuste de la DBSIRBF o comparar ajustes entre DBSIRBF se recomienda hacer uso de la validación cruzada (leave-one-out), empleando la expresión (3.16).

### 4.3 Estudio de simulación y discusión

Esta sección describe un estudio de simulación para evaluar la eficiencia del método propuesto bajo diferentes condiciones asociadas con los parámetros de suavizamiento y las funciones de base radial utilizando el método basado en distancias. Los escenarios están diseñados teniendo en cuenta la propuesta de Wand (2000), con algunas adaptaciones a un espacio bidimensional. En particular se estudia los efectos de: (i) el nivel de ruido, (ii) la densidad de diseño, (iii) el grado de variación espacial y (iv) la función de varianza. Estas configuraciones y escenarios se presentan en la Tabla 4.2.

TABLA 4.2: Escenarios considerados en los experimentos espaciales simulados

Factor	Forma genérica
Nivel de ruido	$z_j(\mathbf{s}_i) = \beta_0 + \beta_1 V_i + \beta_2 D_{i2} + \beta_3 D_{i3} + f(w_{x_i}) + f(w_{y_i}) + f(w_{x_i})f(w_{y_i}) + \sigma_j \varepsilon(\mathbf{s}_i)$ $\sigma_j = 0.02 + 0.04(j - 1)^2$
Densidad de diseño	$z_j(\mathbf{s}_i) = \beta_0 + \beta_1 V_i + \beta_2 D_{i2} + \beta_3 D_{i3} + f(X_{ji}) + f(Y_{ji}) + f(X_{ji})f(Y_{ji}) + \sigma \varepsilon(\mathbf{s}_i)$ $\sigma = 0.1, X_{ji} = F_j^{-1}(X_i), Y_{ji} = F_j^{-1}(Y_i)$
Variación espacial	$z_j(\mathbf{s}_i) = \beta_0 + \beta_1 V_i + \beta_2 D_{i2} + \beta_3 D_{i3} + f_j(w_{x_i}) + f_j(w_{y_i}) + f_j(w_{x_i})f_j(w_{y_i}) + \sigma \varepsilon(\mathbf{s}_i)$ $\sigma = 0.2, f_j(l_i) = \sqrt{l_i(1-l_i)} \sin \left[ \frac{2\pi \{1 + 2^{(9-4j)/5}\}}{l_i + 2^{(9-4j)/5}} \right]$
Función de varianza	$z_j(\mathbf{s}_i) = \beta_0 + \beta_1 V_i + \beta_2 D_{i2} + \beta_3 D_{i3} + f(w_{x_i}) + f(w_{y_i}) + f(w_{x_i})f(w_{y_i})$ $+ \sqrt{v_j(w_{x_i}) + v_j(w_{y_i}) + v_j(w_{x_i})v_j(w_{y_i})} \varepsilon(\mathbf{s}_i)$ con $v_j(l_i) = \{0.15 [1 + 0.4(2j - 7)(l_i - 0.5)]\}^2$
Los supuestos y otras elecciones	
$V_i \sim Bi(n, 0.4); i = 1, \dots, 100; n = 50, 100, 150; \varepsilon(\mathbf{s}_i) \stackrel{iid}{\sim} N(0, 0.1); X_i, Y_i \stackrel{iid}{\sim} Uniform(0, 1)$ $F_j$ es la Beta $\left(\frac{j+4}{5}, \frac{11-j}{5}\right); j = 1, 3; f(l_i) = 1.5f_1\left(\frac{l_i-0.5}{0.15}\right) - f_1\left(\frac{l_i-0.8}{0.04}\right); f_1(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right);$ $l_i = w_{x_i}, w_{y_i}$	

Este estudio considera las simulaciones en dos dimensiones  $(w_x, w_y)$ , además una variable aleatoria binomial  $V_i \sim Bi(n, 0.4)$ , con tamaños de muestra  $n = 50, 100, 150$ , tamaños de vecindario  $n_h = 8, 32$ , parámetros de suavizamiento  $\eta = 0.01, 0.1$  y  $j = 1, 3$  en el factor de varianza. Adicionalmente, supóngase que se tiene una variable nominal asociada a tres regiones específicas en el cuadrado de longitud uno, como se muestra en la Figura 3.1. Puesto que hay tres regiones, sólo dos variables dummy ( $D_2$  y  $D_3$ ) se consideran para evitar problemas de singularidad. Además,  $\varepsilon(\mathbf{s}_i)$  se construye asumiendo un campo aleatorio Gaussiano asociado con la pepita de  $\tau^2 = 0.1$ . Para los parámetros de

tendencia, se asume los siguientes valores  $\beta_0 = 10$ ,  $\beta_1 = -4$ ,  $\beta_2 = 2$  y  $\beta_3 = -4$  asociados a las coordenadas espaciales  $w_{x_i}$  y  $w_{y_i}$ , donde  $i$  es la  $i$ -ésima observación simulada. En la Tabla 4.3, se presentan los escenarios simulados. El método propuesto se prueba con cinco RBFs: MQ, TPS, CRS, ST y EXP. Un total de 120 escenarios fueron simulados, y para cada uno de ellos, el proceso se repitió 100 veces.

TABLA 4.3: Escenarios espaciales simulados (los números naturales en las últimas cinco columnas (de 1 a 120) representan el número del escenario)

Parámetros del modelo			$n$	Función de base radial				
$\eta$	$j$	$n_h$		MQ	TPS	CRS	ST	EXP
0.01	1	8	50	1	25	49	73	97
			100	2	26	50	74	98
			150	3	27	51	75	99
		32	50	4	28	52	76	100
			100	5	29	53	77	101
			150	6	30	54	78	102
	3	8	50	7	31	55	79	103
			100	8	32	56	80	104
			150	9	33	57	81	105
		32	50	10	34	58	82	106
			100	11	35	59	83	107
			150	12	36	60	84	108
0.1	1	8	50	13	37	61	85	109
			100	14	38	62	86	110
			150	15	39	63	87	111
		32	50	16	40	64	88	112
			100	17	41	65	89	113
			150	18	42	66	90	114
	3	8	50	19	43	67	91	115
			100	20	44	68	92	116
			150	21	45	69	93	117
		32	50	22	46	70	94	118
			100	23	47	71	95	119
			150	24	48	72	96	120

Para cada conjunto de datos simulados, se evaluó la calidad del ajuste mediante el RMSPE obtenido por el método LOOCV. Los resultados se presentan en las Tablas 4.4 y 4.5. Inicialmente se consideró un parámetro positivo para  $\rho$ , pero los valores de RMSPE no mostraron diferencias significativas con los obtenidos cuando  $\rho = 0$ ; en particular, cuando fueron utilizadas las funciones

multicuadrática (MQ) y exponencial (EXP).

TABLA 4.4: Promedios de RMSPEs bajo el método DBSIRBF en los escenarios espaciales presentados en la Tabla 4.3 (casos nivel de ruido y densidad de diseño)

Parámetro			$n$	Nivel de ruido					Densidad de diseño				
$\eta$	$j$	$n_h$		MQ	TPS	CRS	ST	EXP	MQ	TPS	CRS	ST	EXP
0.01	1	8	50	3.50	3.58	4.10	3.88	3.61	3.40	4.07	3.69	3.90	3.43
			100	6.67	6.61	6.26	6.24	6.20	68.01	34.21	68.01	68.01	68.01
			150	10.35	10.98	8.74	8.65	9.66	9.16	9.47	8.98	9.06	9.04
		32	50	2.10	2.62	1.96	2.14	2.05	1.79	2.29	1.51	1.81	1.74
			100	2.27	2.27	1.58	1.73	1.89	2.15	2.13	1.41	1.54	1.69
			150	2.21	2.40	1.72	1.81	2.04	1.97	2.15	1.53	1.64	1.81
	3	8	50	4.08	4.30	4.59	4.51	4.17	6.57	7.03	6.83	6.82	6.54
			100	6.74	6.67	6.31	6.29	6.26	39.48	21.86	39.48	39.48	39.48
			150	10.10	10.72	8.52	8.43	9.43	7.34	7.64	6.35	6.27	6.91
		32	50	2.11	2.63	1.97	2.15	2.06	1.91	2.42	1.70	1.93	1.86
			100	2.28	2.28	1.59	1.74	1.90	2.24	2.29	1.42	1.63	1.82
			150	2.22	2.41	1.72	1.82	2.04	2.02	2.19	1.55	1.64	1.85
0.1	1	8	50	3.67	3.59	3.92	3.59	3.61	4.03	3.98	4.79	4.02	3.43
			100	10.22	7.11	6.46	6.63	6.19	68.01	34.02	68.01	68.01	68.01
			150	14.82	11.02	7.68	10.35	9.66	10.49	9.43	9.43	9.34	9.04
		32	50	2.84	2.61	2.05	2.58	2.05	2.42	2.27	1.60	2.25	1.74
			100	11.08	3.45	1.61	2.59	1.89	9.76	3.40	1.46	2.51	1.69
			150	5.17	2.59	1.78	2.46	2.04	4.61	2.33	1.57	2.20	1.81
	3	8	50	4.35	4.29	4.64	4.30	4.17	7.10	6.95	7.07	6.94	6.54
			100	10.23	7.17	6.50	6.69	6.26	39.48	21.86	39.48	39.48	39.48
			150	14.42	10.75	7.53	10.11	9.43	10.16	7.69	5.59	7.25	6.91
		32	50	2.85	2.61	2.07	2.59	2.06	2.60	2.41	1.79	2.38	1.86
			100	11.10	3.47	1.62	2.60	1.90	10.26	3.50	1.50	2.63	1.82
			150	5.18	2.6	1.78	2.47	2.04	4.94	2.39	1.61	2.25	1.85

Las Tablas 4.4 y 4.5 muestran los valores promedios de RMSPE para los 120 casos descritos en la Tabla 4.3 en 100 simulaciones por caso. El método DBSIRBF funciona bien para vecindarios grandes, lo que indica una ganancia (vista en un decrecimiento) de 70% de los valores promedios de RMSPE cuando  $n_h = 32$  con respecto a  $n_h = 8$ . Sin embargo, cuando  $j = 3$  los valores promedios de RMSPE fueron 1.92 veces mayores que los obtenidos cuando  $j = 1$ . Al tener en cuenta el parámetro  $\eta$ , en general se presenta una ligera reducción de 4.8% en los valores promedios de RMSPE cuando  $\eta = 0.01$ , comparado con  $\eta = 0.1$ .

Por otro lado, analizando las formas genéricas, los valores más bajos de

TABLA 4.5: Promedios de RMSPEs bajo el método DBSIRBF en los escenarios espaciales presentados en la Tabla 4.3 (casos varianza espacial y función de varianza)

Parámetro			$n$	Variación espacial					Función de varianza				
$\eta$	$j$	$n_h$		MQ	TPS	CRS	ST	EXP	MQ	TPS	CRS	ST	EXP
0.01	1	8	50	5.85	5.72	6.90	6.55	6.05	5.09	5.22	5.80	5.58	5.19
			100	4.83	5.77	3.73	3.69	4.10	6.70	6.66	6.23	6.21	6.20
			150	4.73	5.15	3.70	3.82	4.38	9.56	10.17	8.12	7.95	8.91
		32	50	2.02	2.55	1.83	2.05	1.96	2.11	2.63	1.97	2.15	2.06
			100	2.13	2.15	1.47	1.63	1.78	2.28	2.28	1.58	1.74	1.90
			150	2.16	2.33	1.63	1.74	1.98	2.22	2.41	1.72	1.82	2.05
	3	8	50	18.87	18.76	18.22	18.32	18.63	4.84	4.98	5.51	5.32	4.94
			100	91.92	91.94	91.84	91.84	91.86	6.76	6.69	6.33	6.31	6.27
			150	83.52	77.94	85.59	85.30	83.82	9.96	10.57	8.39	8.31	9.29
		32	50	1.98	2.58	1.65	2.00	1.92	2.11	2.63	1.97	2.15	2.06
			100	2.15	2.17	1.49	1.63	1.76	2.28	2.28	1.59	1.74	1.90
			150	2.23	2.42	1.68	1.82	2.03	2.22	2.41	1.72	1.82	2.05
0.1	1	8	50	5.87	5.78	5.87	5.78	6.05	5.31	5.24	5.41	5.24	5.19
			100	9.31	6.02	3.98	5.39	4.10	10.19	7.17	6.42	6.68	6.20
			150	6.71	5.21	4.08	4.99	4.38	13.82	10.20	7.18	9.60	8.91
		32	50	2.71	2.54	1.89	2.52	1.96	2.84	2.61	2.07	2.59	2.06
			100	10.03	3.18	1.52	2.41	1.78	11.13	3.46	1.62	2.60	1.90
			150	5.23	2.53	1.70	2.39	1.98	5.19	2.60	1.79	2.47	2.05
	3	8	50	19.42	19.06	18.54	19.09	18.63	5.06	4.99	5.21	5.00	4.94
			100	92.35	90.02	91.90	91.95	91.86	10.26	7.19	6.52	6.71	6.27
			150	84.19	21.39	85.26	83.69	83.82	14.22	10.61	7.44	9.98	9.29
		32	50	2.81	2.56	1.77	2.53	1.92	2.85	2.62	2.07	2.59	2.06
			100	9.79	3.30	1.55	2.49	1.76	11.12	3.47	1.63	2.60	1.90
			150	5.98	2.67	1.78	2.49	2.03	5.19	2.60	1.79	2.47	2.05

RMSPE correspondieron a los casos de nivel de ruido y variación espacial, con valores promedios de RMSPE de 4.79 y 4.92, respectivamente. Para los casos, densidad del diseño y función de varianza, los valores promedios de RMSPE fueron de 11.55 y 18.22, respectivamente. En cuanto al método DBSIRBF se encuentra que: i) para el caso *nivel de ruido*, el método DBSIRBF que produce el valor promedio de RMSPE más bajo fue el CRS, mientras que la MQ muestra el valor más alto, ii) para el caso *densidad de diseño*, el método DBSIRBF con promedio de RMSPE más pequeño fue el TPS, mientras que MQ volvió a mostrar el más alto, iii) en términos de *variación espacial*, el valor promedio de RMSPE más bajo se observó con la CRS, mientras que la MQ demostró una vez más los valores promedio de RMSPE mas altos, y iv)



en el caso *función de varianza* el método DBSIRBF con mejores resultados en términos de promedios de RMSPEs fue el TPS, y de nuevo, la función MQ utilizada en el método DBSIRBF muestra los peores resultados.

Debido a que los valores promedios de RMSPEs fueron mayores en tamaños de vecindarios  $n_h = 8$  con respecto a los casos con  $n_h = 32$ , se muestran los bloxplot por separado (véanse las Figuras 4.1 y 4.2). Analizando la Figura 4.1 se resalta lo siguiente: i) en términos de *nivel de ruido*, se encontró una menor variabilidad cuando  $j = 1$ , aunque los valores promedios de RMSPEs fueron muy similares para ambos casos  $j = 1$  y  $j = 3$ ; ii) en el caso de la *función de diseño*, el método DBSIRBF que mostró la mayor variabilidad fue el TPS, mientras que los otros presentaron comportamientos similares; iii) bajo los escenarios de *variación espacial*, cuando  $j = 3$  y  $n = 150$  se generó en general una gran variabilidad, excepto para el TPS cuando  $\eta = 0.1$ ; y iv) para el caso *función de varianza*, la variabilidad fue mayor cuando  $n = 50$  y disminuyó cuando  $n = 100$  y  $n = 150$ , es de resaltar aquí que los únicos casos en los cuales la MQ no disminuyó fue cuando  $\eta = 0.1$ , lo cual se esperaba ya que la función de base radial MQ en general trabaja mejor en valores pequeños de  $\eta$ .

De acuerdo con la Figura 4.2, se observa que: i) para el *nivel de ruido*, se encontró una mayor variabilidad cuando  $j = 3$ , incrementándose el valor promedio de RMSPE en  $\eta = 0.1$  y  $n = 100$ , especialmente cuando se utilizan las funciones de base radial MQ, TPS y ST; ii) en términos de la *función de diseño*, las funciones de base radial MQ y ST mostraron mayor variabilidad cuando  $j = 3$  y  $n = 100$ , mientras para las funciones de base radial EXP y TPS, la mayor variabilidad se mostró en  $j = 1$  y  $n = 100$ , y los valores promedios de RMSPEs mayores se presentan en las funciones de base radial MQ, TPS y ST; iii) para la *variación espacial* y la *función de varianza*, la mayor variabilidad se presentó cuando  $j = 3$ , un poco más grande en las funciones de base radial CRS bajo el caso de *variación espacial* que la observada en el caso de la *función de diseño*. En general, el valor promedio de RMSPE fue menor cuando  $n = 100$ , excepto en los casos de funciones de base radial MQ.

En términos generales, el método DBSIRBF fue lo suficientemente robusto ante diferentes tamaños de muestra debido a que la variabilidad fue similar y

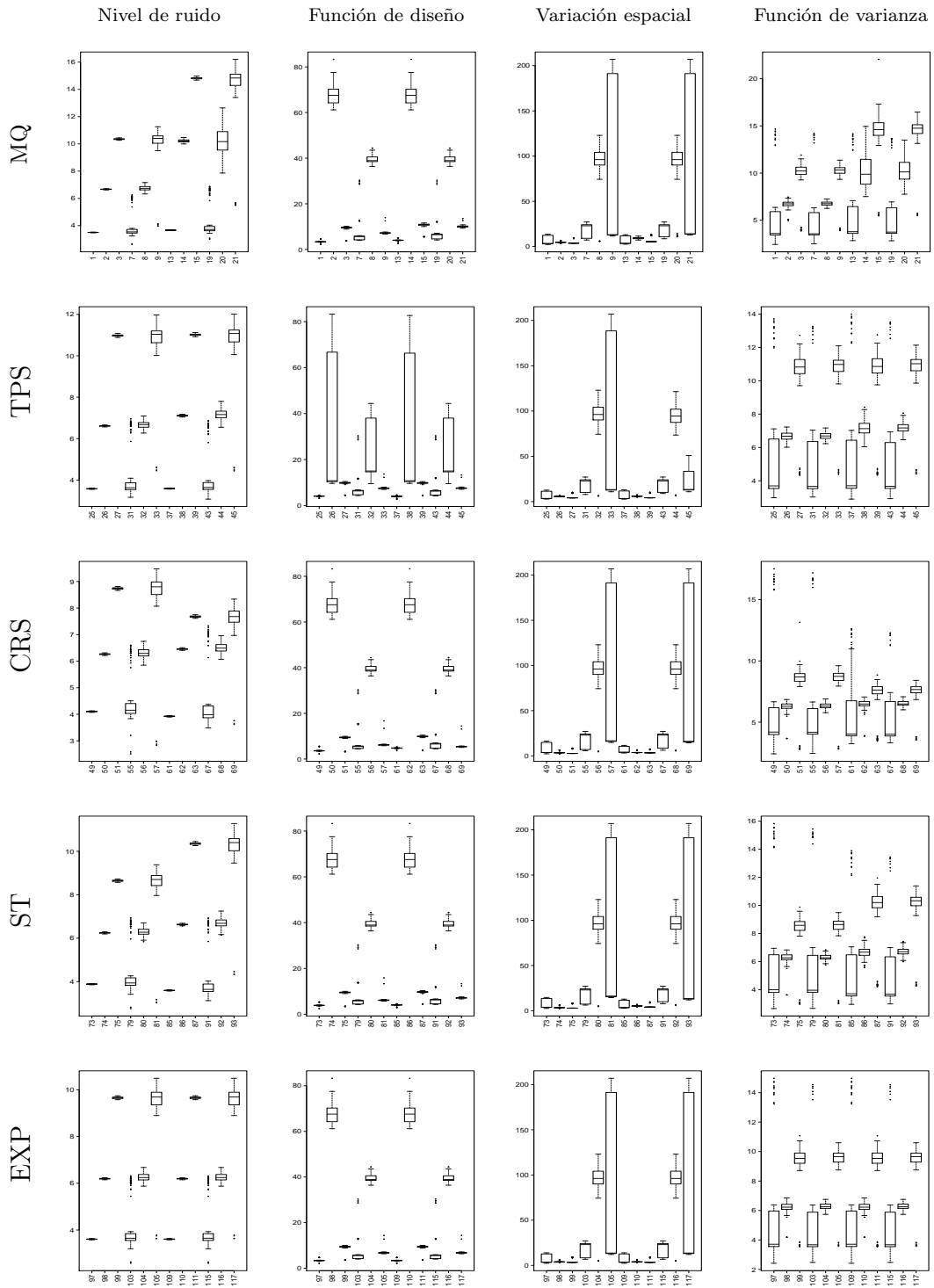


FIGURA 4.1: RMSPE para los escenarios espaciales simulados cuando  $n_h = 8$

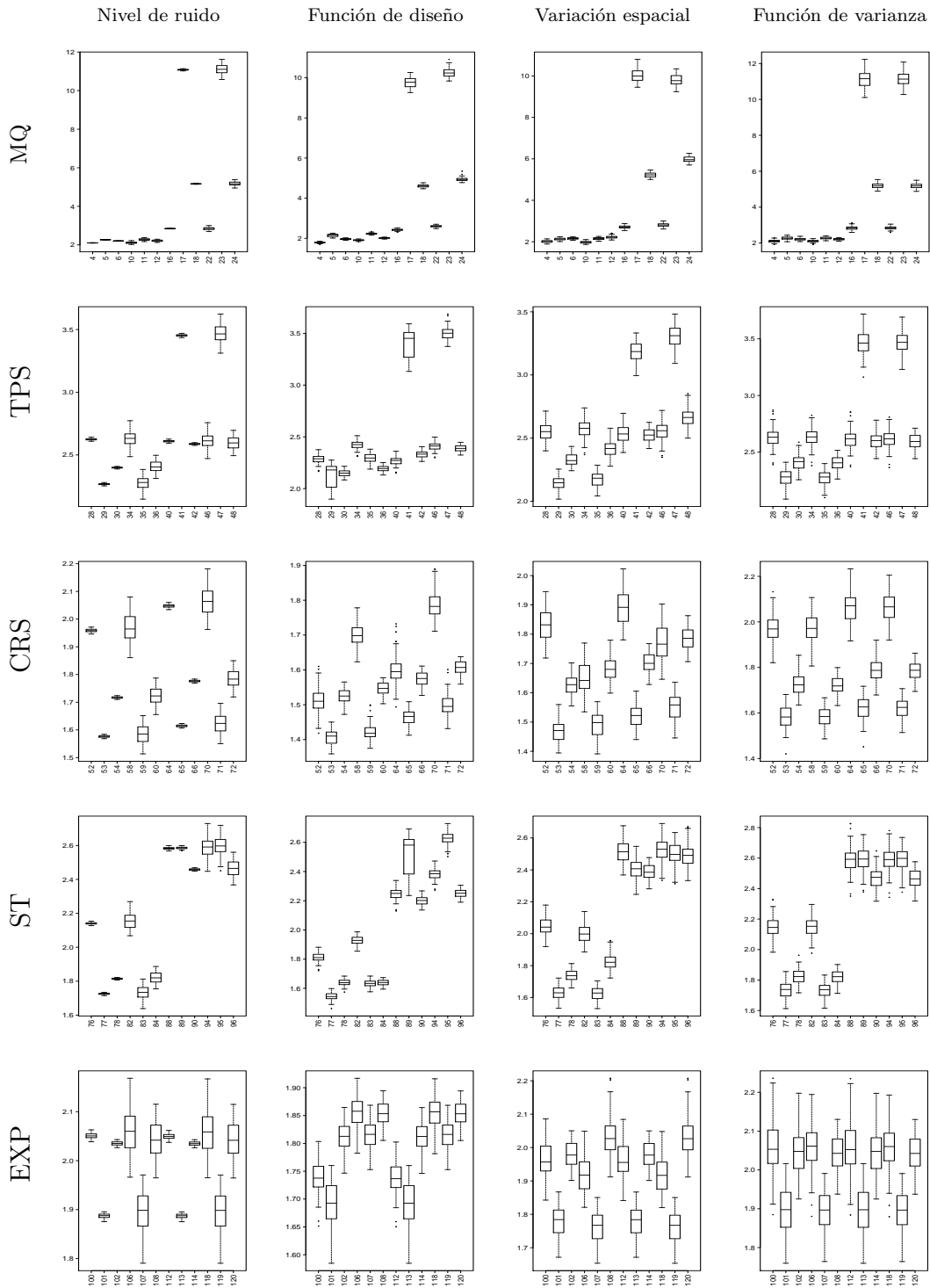


FIGURA 4.2: RMSPE para los escenarios espaciales simulados cuando  $n_h = 32$

homogénea en todos los escenarios analizados.

## 4.4 Aplicación

El conjunto de datos empleado en esta aplicación corresponde al utilizado en la Subsección 3.4.2 del Capítulo 3 y que fue estudiado por Capeche et al. (1997), la distribución de los puntos observados se presenta en las Figuras 3.11 y 4.3(a). Al igual que en la Subsección 3.4.2, las coordenadas principales se construyeron a partir de las variables: coordenadas espaciales y la variable nominal que define la sub-región. Para la selección de las coordenadas principales, se utiliza el criterio dado en (3.5) para hacer una primera selección con el fin de remover las coordenadas principales pobremente correlacionadas con las variable regionalizada, y luego, se emplea el criterio (3.6) para seleccionar las coordenadas principales mas significativas utilizando la regresión DB.

Para obtener los mapas de contenido de calcio, se utilizan las coordenadas espaciales (tendencia espacial lineal) y la variable de sub-región (tipo de suelo) categorizada. De acuerdo a los resultados encontrados en la Subsección 3.4.2, diecisiete coordenadas principales fueron consideradas en la eliminación de la tendencia y la validación cruzada (leave-one-out) se llevó a cabo para seleccionar el método DBSIRBF que mostrará el RMSPE más bajo.

TABLA 4.6: Comparación de algunos métodos DBSIRBFs para el contenido de calcio utilizando LOOCV

RBF	Optimización					
	optim		bobyqa			
	$\eta$	RMSPE	$\eta$	$\rho$	iter	RMSPE
EXP	0.03	7.32	0.03	0.00	234	7.32
MQ	0.00	7.38	0.00	0.00	19	7.38
IM	27.36	7.30	27.36	0.00	45	7.30
TPS	0.81	7.62	0.20	1.24	23	7.61
CRS	0.10	7.33	0.30	2.38	31	7.33
ST	0.08	7.34	0.58	2.04	34	7.33
GAU	0.20	8.01	0.20	1.00	28	8.01

Los resultados de la validación cruzada se muestran en la Tabla 4.6, donde

se observa que las funciones de base radial que mejor trabajan en el método propuesto son la EXP y la CRS mostrando los valores de RMSPE más bajos. Como detalle práctico, las funciones de optimización usadas para  $\rho$  y  $\eta$  fueron “bobyqa” del paquete “minqa” y “optim” del paquete “stats” del programa R Development Core Team (2012), dependiendo del método que proporcionara resultados más estables. Esta información también se muestra en la Tabla 4.6. Los mapas de predicción correspondientes se muestran en la Figura 4.3(b)-(h).

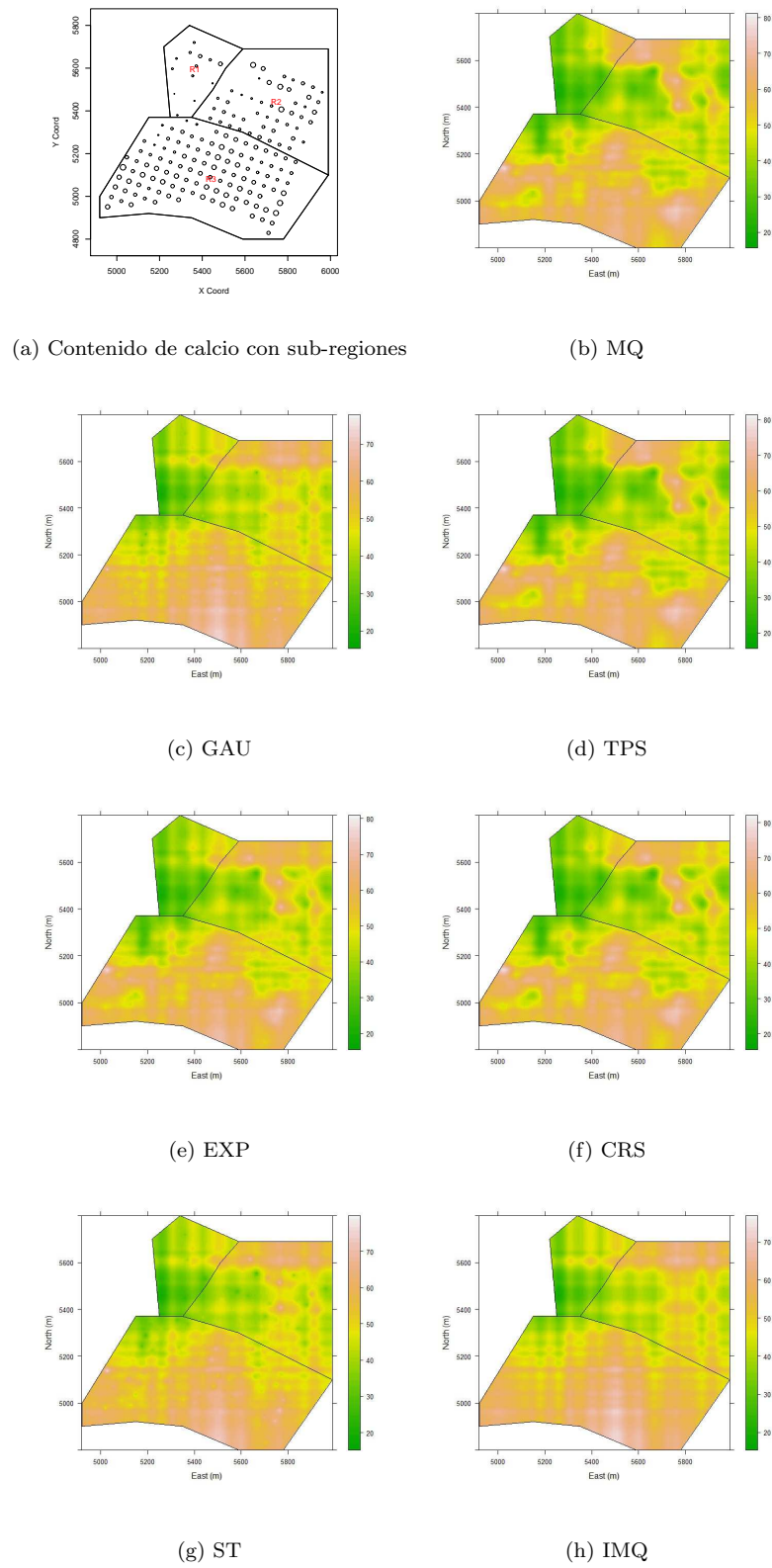


FIGURA 4.3: Localizaciones de muestreo y mapas de predicción bajo el método DBSIRBF para el contenido de calcio en el suelo, incluyendo sub-región (tipo de suelo)



# Capítulo 5

## Modelo basado en distancias para la predicción espacio-temporal usando funciones de base radial

### 5.1 Introducción

En los últimos años ha habido un enorme crecimiento de los modelos y técnicas para la realización del análisis de datos espacio-temporal. En Cressie & Huang (1999), Christakos (2000), Myers et al. (2002), Mateu et al. (2003), Kolovos et al. (2004), Banerjee et al. (2004), Sahu & Mardia (2005), Chen et al. (2006) y Gneiting et al. (2007), entre otros, se puede encontrar un resumen de las principales técnicas para modelos espacio-temporales, junto con numerosas aplicaciones prácticas para una variedad de fenómenos naturales. Por ejemplo, se encuentran estudios de: contaminación del aire (De Cesare et al. 1997, De Cesare et al. 2001*b*, Şen et al. 2006), precipitación (Yavuz & Erdoğan 2012), temperatura terrestre (Perčec Tadić 2010, Hengl et al. 2012), hidrología (Rouhani & Hall 1989, Rouhani & Myers 1990), ecología (Bellier et al. 2007, Planque et al. 2007), monitoreo y seguimiento de población de la fauna silvestre (Kondoh et al. 2011), medicina en el análisis de imágenes del cerebro (Ye



2008, Ye et al. 2011) y análisis económico de los precios inmobiliarios (Chica et al. 2007).

Las mediciones de contaminación del aire, precipitación y temperatura terrestre a menudo se observan a diario en más de un centenar de lugares en el mundo y los datos de los últimos años suelen estar disponibles, sumado a esto, esta disponibilidad de información de imágenes de satélite y el software diseñado para el análisis (estadístico y geográfico), motivan el estudio del problema en el presente capítulo. Los métodos geoestadísticos espacio-temporales, sin embargo, requieren de conjuntos de datos grandes, que complican su manipulación por el tiempo de procesamiento y por la difícil tarea de adaptarse a modelos realistas y complejos.

Por otro lado, las funciones de base radial tales como la multicuadrática o spline completamente regularizado son útiles en la construcción de modelos digitales de elevación (DEM), como se muestra en (Mitášová & Hofierka 1993). Una variación de la función multicuadrática se llama la multicuadrática inversa, introducida por (Hardy & Gopfert 1975). En Späh (1969) se describe un método que permite evitar puntos de inflexión y contiene splines cúbicos como un caso especial, utilizando interpolación spline cúbica y exponencial. Más tarde, la spline capa delgada se introdujo en el diseño geométrico por (Duchon 1976) y la aproximación Gaussiana utilizada por (Schagen 1979) se presenta como una variante popular de la TPS. Por último, (Mitáš & Mitášová 1988, Mitášová & Hofierka 1993, Mitášová & Mitáš 1993) desarrollan la formulación del spline con tensión e implementan un algoritmo de segmentación con un tamaño flexible en la superposición del vecindario.

En este capítulo, basados en la combinación de funciones de base radial y las coordenadas principales obtenidas a través del método basado en distancias, se propone un nuevo método llamado interpolación espacio-temporal basado en distancias usando funciones de base radial (distance-based spatio-temporal interpolation using radial based functions, DBSITRBFs). El método propuesto considera principalmente la interpolación espacio-temporal de las funciones de base radial en un modelo métrico espacio-temporal, con tendencia obtenida a partir de las coordenadas principales, las cuales se obtienen a partir de las

variables explicativas mixtas mediante el método de descomposición espectral que se realiza a las distancias entre individuos.

Especialmente, en el análisis de datos espacio-temporal, a menudo se tiene que lidiar con variables de diversa naturaleza que están asociadas con la variable respuesta: las variables categóricas y binarias tales como el tipo de suelo o estación del año, y las variables continuas (por ejemplo, las coordenadas espaciales o la precipitación). El objetivo aquí es presentar un enfoque unificado que utiliza las RBFs en tales contextos espacio-temporales donde las variables explicativas son de naturaleza mixta. Por lo tanto, este trabajo propone un nuevo método utilizando las distancias entre los individuos, tales como la distancia de Gower (1968); aunque alguna otra distancia Euclidiana también se puede llegar a utilizar.

La propuesta espacio-temporal basada en distancias esta soportada en los métodos desarrollados por Cuadras & Arenas (1990) y Cuadras et al. (1996), quienes como se ha dicho en los capítulos anteriores presentaron algunos resultados de un modelo DB para la predicción con variables mixtas. Esta estrategia es una excelente alternativa, ya que aprovecha al máximo la información obtenida debido a la relación entre las observaciones, la cual puede ser establecida a través del uso de la descomposición espectral, utilizando cualquier distancia Euclidiana. En consecuencia, este enfoque permite incluir en el modelo mas coordenadas principales que variables explicativas en los puntos de muestreo para mejorar las predicciones generales.

Las coordenadas principales obtenidas mediante el método de distancias se encuentran a partir de las covariables asociadas con la variable respuesta y las coordenadas espacio-temporales. La selección de las coordenadas principales se lleva a cabo utilizando los valores de la prueba- $t$  significativos estadísticamente y considerando una caída significativa en la falta de predictibilidad, es decir, las coordenadas principales que están más asociadas con el variable respuesta. De esta manera, para evaluar la exactitud del interpolador del método propuesto, se realizaron simulaciones incondicionales para cinco funciones de base radial en diferentes escenarios prácticos. Los resultados muestran que las RBFs utilizando el método DB tienen ventajas como la de trabajar con variables mixtas

en el tendencia y el no requerir de la estimación de un variograma espacio-temporal, que normalmente requiere mucho tiempo computacional. Además, ofrece flexibilidad en el ajuste de los parámetros de suavizamiento, ya que al ser un método local trabaja sólo con vecindarios que el investigador puede modificar de acuerdo a su conocimiento.

Este capítulo está dividido en 3 secciones principales. En la Sección 5.2 se desarrolla la propuesta metodológica introduciendo la tendencia lineal local basada en distancias; se construyen las RBFs a partir de la tendencia basada en distancias, se describen algunas RBFs espacio-temporales y se presenta una aproximación de la interpolación spline para el modelo propuesto utilizando el método de interpolación kriging. En la Sección 5.3 se presenta un estudio de simulación basado en algunos modelos spline espacio-temporales. En la Sección 5.4 se desarrolla una aplicación de la temperatura media mensual en Croacia, donde se incorpora la tendencia a partir de las coordenadas principales obtenidas en función de las coordenadas espacio-temporales, distancia al mar, elevación y estación climática del año.

## 5.2 Modelo espacio-temporal basado en distancias con tendencia lineal local

Sea  $\{Z(\mathbf{s}, t), \mathbf{s} \in D, t \in T\}$  un proceso espacio-temporal aleatorio, donde  $\mathbf{s}$  varía sobre un conjunto dado  $D \subseteq \mathbb{R}^d$  y  $T \subseteq \mathbb{Z}$  ó  $\mathbb{R}$ , de manera que los modelos desarrollados son adecuados tanto para tiempo-discreto como para tiempo-continuo. Sin pérdida de generalidad, se toma  $T \subseteq \mathbb{R}$ . Supóngase que este proceso se observó en un conjunto de localizaciones espacio-temporales  $\{(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)\} \in D \times T$  obteniendo un conjunto de valores  $\{Z(\mathbf{s}_1, t_1), \dots, Z(\mathbf{s}_n, t_n)\}$ .

Supóngase que el proceso estocástico espacio-temporal,  $Z(\mathbf{s}_i, t_i)$ , sigue un modelo de función aleatoria, que puede descomponerse como

$$Z(\mathbf{s}_i, t_i) = \mu(\mathbf{s}_i, t_i) + \varepsilon(\mathbf{s}_i, t_i) \quad (\mathbf{s}_i, t_i) \in \mathbb{R}^d \times \mathbb{R} \quad (5.1)$$

con  $i = 1, \dots, n$  y donde  $\mu(\mathbf{s}_i, t_i) = E[Z(\mathbf{s}_i, t_i)]$  es una función determinística

asociada con la tendencia y  $\varepsilon(\mathbf{s}_i, t_i)$  es un proceso estocástico de media cero y variograma  $2\gamma(\cdot, \cdot)$ . Este proceso caracteriza la dependencia espacio-temporal y modeliza las fluctuaciones espacio-temporales de  $Z(\mathbf{s}_i, t_i)$  alrededor de su media  $\mu(\mathbf{s}_i, t_i)$ . La tendencia espacial está formada por las variables categóricas, continuas y binarias, y se modela como

$$\mu(\mathbf{s}_i, t_i) = \theta_0 + v'(\mathbf{s}_i, t_i)\boldsymbol{\theta} \quad (5.2)$$

donde  $v(\mathbf{s}_i, t_i) = (v_1(\mathbf{s}_i, t_i), \dots, v_p(\mathbf{s}_i, t_i))'$  es un vector que contiene variables explicativas asociadas a la localización espacio-temporal  $(\mathbf{s}_i, t_i)$ ,  $\theta_0$  es el parámetro desconocido asociado al intercepto y  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)'$  es un vector de parámetros desconocidos.

En forma matricial el modelo (5.1), se puede expresar como:

$$\mathbf{Z}_{st} = \mathbf{1}\theta_0 + V\boldsymbol{\theta} + \boldsymbol{\varepsilon}_{st} \quad (5.3)$$

donde  $\mathbf{Z}_{st} = (Z(\mathbf{s}_1, t_1), \dots, Z(\mathbf{s}_n, t_n))'$ ,  $\mathbf{1}$  es un vector de dimensión  $n \times 1$  asociado al intercepto,  $V = (V_1, \dots, V_n)$  es la matriz de diseño de dimensión  $n \times p$  con  $p$  variables explicativas  $V_j = (v_j(\mathbf{s}_1, t_1), \dots, v_j(\mathbf{s}_n, t_n))'$  de dimensión  $n \times 1$ ,  $j = 1, \dots, p$ ; además,  $\boldsymbol{\varepsilon}_{st} = (\varepsilon(\mathbf{s}_1, t_1), \dots, \varepsilon(\mathbf{s}_n, t_n))'$ .

Ahora, la idea es hacer una transformación de las variables explicativas utilizando el método basado en distancias. Para ello se definen las medidas de similitud o distancia Euclidiana presentadas en la Subsección 2.7.1, que dependen de las características de las variables explicativas.

Si el vector  $v(\mathbf{s}_i, t_i)$  dado en (5.2) está formado por variables binarias, categóricas y continuas, entonces la similitud de acuerdo a Gower (1971) se puede definir para variables mixtas como la expresión presentada en (2.17). En el caso que las variables explicativas sean binarias o categóricas, como se mencionó en la Subsección 2.7.1, la similitud se puede expresar mediante las expresiones presentadas en (2.16). Por medio de la transformación

$$\delta_{ij} = \sqrt{1 - m_{ij}}$$

es posible obtener las distancias Euclidianas. Si todas las variables explicativas en (5.2) son continuas, la distancia al cuadrado se define como

$$\delta_{ij}^2 = (v(\mathbf{s}_i, t_i) - v(\mathbf{s}_j, t_j))'(v(\mathbf{s}_i, t_i) - v(\mathbf{s}_j, t_j))$$

o alternativamente por la distancia absoluta  $\delta_{ij}^2 = \sum_{l=1}^p |v_l(\mathbf{s}_i, t_i) - v_l(\mathbf{s}_j, t_j)|$ . Entonces, en el caso de sólo disponer información de las coordenadas espacio-temporales,  $(w_x, w_y, t)$ , las distancias espacio-temporales estarán dadas por

$$\delta_{ij} = \sqrt{(w_{x_i} - w_{x_j})^2 + (w_{y_i} - w_{y_j})^2 + (t_i - t_j)^2}$$

Expresiones para la similaridad de Gower como la dada en la ecuación (2.17) serán útiles en la medida de disponer de información asociada con las variables mixtas, no sólo para los puntos muestreados sino también para los no muestreados, lo cual restringe su uso en las zonas no muestreadas.

Realizando el mismo procedimiento que se presentó en la Sección 3.2, la idea es hacer una transformación de las variables explicativas utilizando el método basado en distancias. Para ello una vez seleccionada alguna distancia Euclidiana, se realiza el proceso de descomposición espectral y se selecciona las coordenadas principales que más se relacionan con la variable respuesta, realizando cualquiera de los cuatro métodos presentados al final de la Sección 3.2. Por lo tanto, las  $X_{k+1}, \dots, X_{n-1}$  coordenadas principales deben ser removidas ya que son las menos relevantes.

### 5.2.1 Tendencia basada en distancias con funciones de base radial

En la interpolación espacio-temporal, existen métodos que no requieren un modelo de dependencia espacio-temporal, como el variograma o covariograma, éstos se llaman deterministas y son los que interesan en esta subsección. El modelo (5.1) utilizando un formato basado en distancias se puede expresar en forma general por

$$Z(\mathbf{s}_i, t_i) = g(\mathbf{s}_i, t_i) + \varepsilon(\mathbf{s}_i, t_i), \quad i = 1, \dots, n \quad (5.4)$$

donde  $g(\mathbf{s}_i, t_i)$  es una función de valor real, dada por

$$g(\mathbf{s}_i, t_i) = \sum_{l=0}^k \nu_l X_l(\mathbf{s}_i, t_i) + \sum_{j=1}^n \omega_j \phi(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j), \quad i = 1, \dots, n$$

o en forma matricial,

$$\mathbf{g}_{st} = \mathbf{X}_{st}\boldsymbol{\nu}_{st} + \Phi_{st}\boldsymbol{\omega}_{st} \quad (5.5)$$

donde  $\mathbf{g}_{st} = (g(\mathbf{s}_1, t_1), \dots, g(\mathbf{s}_n, t_n))'$ ,  $\mathbf{X}_{st} = (\mathbf{1}, X) = (\mathbf{1}, X_1, \dots, X_k)$  es una matriz  $n \times (k + 1)$  con elementos  $\mathbf{1}$  y  $X_l = (x_l(\mathbf{s}_1, t_1), \dots, x_l(\mathbf{s}_n, t_n))'$ ,  $l = 1, \dots, k$ ;  $\boldsymbol{\nu}_{st} = (\nu_0, \dots, \nu_k)'$  donde cada  $\nu_l$  corresponde al  $l$ -ésimo coeficiente del modelo de tendencia;  $\Phi$  es una matriz  $n \times n$  con elementos  $\phi(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j)$ , los cuales son funciones de base radial, es decir funciones escalares de la distancia Euclidiana entre las coordenadas espacio-temporales  $(\mathbf{s}_i, t_i)$  y  $(\mathbf{s}_j, t_j)$ ; finalmente,  $\boldsymbol{\omega}_{st} = (\omega_1, \dots, \omega_n)'$ , con  $\omega_i$  un peso desconocido.

Los parámetros  $\boldsymbol{\nu}_{st}$  y  $\boldsymbol{\omega}_{st}$  pueden ser estimados por mínimos cuadrados penalizados, minimizando la siguiente expresión

$$\sum_{i=1}^n [Z(\mathbf{s}_i, t_i) - g(\mathbf{s}_i, t_i)]^2 + \rho \int_{\mathbb{R}^2 \times \mathbb{R}} J_m(g(\mathbf{s}, t)) d(\mathbf{s}, t) \quad (5.6)$$

donde  $J_m(g(\mathbf{s}, t))$  es una medida de la rugosidad de la función spline  $g$  (definida en términos de las  $m$ -ésimas derivadas de  $g$ ) y  $\rho > 0$  actúa como un parámetro de suavizamiento.

La expresión (5.6) se puede expresar al hacer los respectivos reemplazos como

$$\begin{aligned} L(\boldsymbol{\nu}_{st}, \boldsymbol{\omega}_{st}) &= (\mathbf{Z}_{st} - \mathbf{X}_{st}\boldsymbol{\nu}_{st} - \Phi_{st}\boldsymbol{\omega}_{st})' (\mathbf{Z}_{st} - \mathbf{X}_{st}\boldsymbol{\nu}_{st} - \Phi_{st}\boldsymbol{\omega}_{st}) \\ &\quad + \rho \int_{\mathbb{R}^2 \times \mathbb{R}} [g''(\mathbf{s}, t)]^2 d(\mathbf{s}, t) \\ &= (\mathbf{Z}_{st} - \mathbf{X}_{st}\boldsymbol{\nu}_{st} - \Phi_{st}\boldsymbol{\omega}_{st})' (\mathbf{Z}_{st} - \mathbf{X}_{st}\boldsymbol{\nu}_{st} - \Phi_{st}\boldsymbol{\omega}_{st}) + \rho \|P_{st}\hat{g}_{st}\|^2 \end{aligned}$$

donde  $P_{st}$  es el espacio que genera  $\Phi_{st}$  y  $\|P_{st}\hat{g}_{st}\|^2 = \langle P_{st}\hat{g}_{st}, P_{st}\hat{g}_{st} \rangle = \boldsymbol{\omega}'_{st} \mathbf{q}_{st} \mathbf{q}'_{st} \boldsymbol{\omega}_{st} = \boldsymbol{\omega}'_{st} \Phi_{st} \boldsymbol{\omega}_{st}$  con  $\Phi_{st} = \mathbf{q}_{st} \mathbf{q}'_{st}$ . Por lo tanto,

$$\begin{aligned} L(\boldsymbol{\nu}_{st}, \boldsymbol{\omega}_{st}) &= (\mathbf{Z}_{st} - \mathbf{X}_{st}\boldsymbol{\nu}_{st} - \Phi_{st}\boldsymbol{\omega}_{st})' (\mathbf{Z}_{st} - \mathbf{X}_{st}\boldsymbol{\nu}_{st} - \Phi_{st}\boldsymbol{\omega}_{st}) + \rho \boldsymbol{\omega}'_{st} \Phi_{st} \boldsymbol{\omega}_{st} \\ &= \mathbf{Z}'_{st} \mathbf{Z}_{st} - 2\mathbf{Z}'_{st} \mathbf{X}_{st} \boldsymbol{\nu}_{st} - 2\mathbf{Z}'_{st} \Phi_{st} \boldsymbol{\omega}_{st} + \boldsymbol{\nu}'_{st} \mathbf{X}'_{st} \mathbf{X}_{st} \boldsymbol{\nu}_{st} \\ &\quad + 2\boldsymbol{\omega}'_{st} \Phi'_{st} \mathbf{X}_{st} \boldsymbol{\nu}_{st} + \boldsymbol{\omega}'_{st} \Phi'_{st} \Phi_{st} \boldsymbol{\omega}_{st} + \rho \boldsymbol{\omega}'_{st} \Phi_{st} \boldsymbol{\omega}_{st} \end{aligned}$$

Al derivar parcialmente con respecto a los vectores  $\boldsymbol{\nu}_{st}$  y  $\boldsymbol{\omega}_{st}$  e igualar a

cero, se encuentra que

$$\frac{\partial L(\boldsymbol{\nu}_{st}, \boldsymbol{\omega}_{st})}{\partial \boldsymbol{\nu}_{st}} = -2\mathbf{X}'_{st}\mathbf{Z}_{st} + 2\mathbf{X}'_{st}\mathbf{X}_{st}\boldsymbol{\nu}_{st} + 2\mathbf{X}'_{st}\Phi_{st}\boldsymbol{\omega}_{st} = 0$$

$$\mathbf{X}_{st}\boldsymbol{\nu}_{st} + \Phi_{st}\boldsymbol{\omega}_{st} = \mathbf{Z}_{st} \quad (5.7)$$

$$\frac{\partial L(\boldsymbol{\nu}_{st}, \boldsymbol{\omega}_{st})}{\partial \boldsymbol{\omega}_{st}} = -2\Phi'_{st}\mathbf{Z}_{st} + 2\Phi'_{st}\mathbf{X}_{st}\boldsymbol{\nu}_{st} + 2\Phi'_{st}\Phi_{st}\boldsymbol{\omega}_{st} + 2\rho\Phi'_{st}\boldsymbol{\omega}_{st} = 0$$

$$\mathbf{X}_{st}\boldsymbol{\nu}_{st} + (\Phi_{st} + \rho I)\boldsymbol{\omega}_{st} = \mathbf{Z}_{st} \quad (5.8)$$

donde  $I$  es la matriz identidad de orden  $n \times n$  y  $\rho$  puede ser interpretado como ruido blanco adicionado a las varianzas en las localizaciones de los datos, pero no la varianza en la localización donde se predice (Wackernagel 2003).

Noté aquí que si  $\Phi_{st}$  es definida positiva entonces hay unicidad de los coeficientes en el interpolador  $\hat{g}(\mathbf{s}_i, t_i)$ . El problema es cómo construir una forma funcional  $\phi(\mathbf{s}_i, t_i)$  con la condición apropiada de definida positiva. Con el fin de generalizar los interpoladores, es necesario considerar las formas más generales de las matrices definidas positivas.

**Definición 5.1.** Sean  $\mathbf{1}, X_1, \dots, X_k$  funciones linealmente independientes de valor-real definidas sobre  $\mathbb{R}^d \times \mathbb{R}$  y sea  $\Phi_{st}$  una matriz simétrica real. Luego,  $\Phi_{st}$  es definida positiva con respecto a  $\mathbf{1}, X_1, \dots, X_k$  si y sólo si para todos los conjuntos de puntos  $(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)$  en  $\mathbb{R}^d \times \mathbb{R}$  se tiene que  $\sum_{i=1}^n \sum_{j=1}^n q_i q_j \phi(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) \geq 0$  para todo  $q_i$  ( $i = 1, \dots, n$ ), donde  $q_i$  es un escalar (no todos cero), y tales que  $\sum_{j=1}^n X_l(\mathbf{s}_j) q_j = 0$  para  $l = 1, \dots, k$ .

Como por la definición 5.1,  $\Phi_{st}$  es definida positiva, entonces  $\Phi_{st} + \rho I$  es invertible y así (5.8) se puede escribir como:

$$\boldsymbol{\omega}_{st} = (\Phi_{st} + \rho I)^{-1}(\mathbf{Z}_{st} - \mathbf{X}_{st}\boldsymbol{\nu}_{st}) \quad (5.9)$$

Reemplazando (5.9) en (5.7) se obtiene

$$\mathbf{X}_{st}\boldsymbol{\nu}_{st} + \Phi_{st}(\Phi_{st} + \rho I)^{-1}(\mathbf{Z}_{st} - \mathbf{X}_{st}\boldsymbol{\nu}_{st}) = \mathbf{Z}_{st}$$

$$[I - \Phi_{st}(\Phi_{st} + \rho I)^{-1}] \mathbf{X}_{st}\boldsymbol{\nu}_{st} = [I - \Phi_{st}(\Phi_{st} + \rho I)^{-1}] \mathbf{Z}_{st} \quad (5.10)$$

Observe que

$$[I - \Phi_{st}(\Phi_{st} + \rho I)^{-1}] = \left( I + \frac{1}{\rho} \Phi_{st} \right)^{-1} = \rho(\Phi_{st} + \rho I)^{-1} \quad (5.11)$$

Premultiplicando por  $\mathbf{X}'_{st}$  la expresión (5.10) y al reemplazar por (5.11), se encuentra que

$$\widehat{\boldsymbol{\nu}}_{st} = [\mathbf{X}'_{st}(\Phi_{st} + \rho I)^{-1}\mathbf{X}_{st}]^{-1} \mathbf{X}'_{st}(\Phi_{st} + \rho I)^{-1}\mathbf{Z}_{st} \quad (5.12)$$

y reemplazando (5.11) en (5.9), se encuentra finalmente que

$$\widehat{\boldsymbol{\omega}}_{st} = (\Phi_{st} + \rho I)^{-1} \{I - \mathbf{X}_{st} [\mathbf{X}'_{st}(\Phi_{st} + \rho I)^{-1}\mathbf{X}_{st}]^{-1} \} \mathbf{Z}_{st} \quad (5.13)$$

$$\mathbf{X}'_{st}(\Phi_{st} + \rho I)^{-1} \mathbf{Z}_{st} \quad (5.14)$$

Alternativamente, al premultiplicar (5.13) se obtiene

$$\mathbf{X}'_{st}\boldsymbol{\nu}_{st} = \mathbf{0}$$

y al combinarlo con el sistema (5.8), se encuentra que  $(\boldsymbol{\omega}_{st}, \boldsymbol{\nu}_{st})$  son la solución del siguiente sistema de ecuaciones lineales

$$\begin{pmatrix} \Phi_{st} + \rho I & \mathbf{X}_{st} \\ \mathbf{X}'_{st} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}_{st} \\ \boldsymbol{\nu}_{st} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{st} \\ \mathbf{0} \end{pmatrix}$$

Si no hay tendencia,  $\mathbf{X}_{st}$  se convierte en un vector de unos y  $\boldsymbol{\nu}_{st}$  en un parámetro de sesgo. En el caso de trabajar con el anterior sistema, es necesario considerar la proposición 5.1.

**Proposición 5.1.** *Sea  $\Phi_{st} + \rho I$  definida positiva y todos los conjuntos de puntos  $(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)$  en  $\mathbb{R}^d \times \mathbb{R}$ , luego la matriz*

$$\begin{pmatrix} \Phi_{st} + \rho I & \mathbf{X}_{st} \\ \mathbf{X}'_{st} & \mathbf{0} \end{pmatrix}$$

*es no singular.*

*Proof.* Supóngase por contradicción que la matriz es singular. Entonces, existe un vector  $(U_1' U_2')'$  no idénticamente cero tal que

$$\begin{pmatrix} \Phi_{st} + \rho I & \mathbf{X}_{st} \\ \mathbf{X}'_{st} & \mathbf{0} \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$$

Por lo tanto,  $(\Phi_{st} + \rho I)U_1 + \mathbf{X}_{st}U_2 = \mathbf{0}$  y  $\mathbf{X}'_{st}U_1 = \mathbf{0}$  con  $U_1 \neq \mathbf{0}$  y  $U_2 \neq \mathbf{0}$ . Claramente,  $\mathbf{X}'_{st}U_1 = \mathbf{0}$  implica  $U_1'\mathbf{X}_{st} = \mathbf{0}$ , y por lo tanto,  $U_1'\mathbf{X}_{st}U_2 = \mathbf{0}$ ,



lo cual implica que  $U_1'(\Phi_{st} + \rho I)U_1 + U_1'\mathbf{X}_{st}U_2 = U_1'(\Phi_{st} + \rho I)U_1 = \mathbf{0}$  lo cual contradice el carácter de definido positivo de  $(\Phi_{st} + \rho I)$  a menos que  $U_1$  sea un vector cero. Si  $U_1$  es el vector cero, entonces  $(\Phi_{st} + \rho I)U_1 = \mathbf{0}$  y de aquí,  $(\Phi_{st} + \rho I)U_1 + \mathbf{X}'_{st}U_2 = \mathbf{X}'_{st}U_2 = \mathbf{0}$ , pero las funciones matriciales escalares ( $\mathbf{X}_{st}$  esta asociada a la tendencia) son linealmente independientes, y así,  $\mathbf{X}'_{st}U_2 = \mathbf{0}$  implica que  $U_2 = \mathbf{0}$ . Dado que  $U_1$  y  $U_2$  son vectores cero, la matriz original debe ser no singular.  $\square$

Ahora, el variograma además de ser un medio para caracterizar la estructura espacio-temporal, se utiliza en kriging para asignar ponderaciones a las observaciones y predecir el valor de alguna variable en las localizaciones no muestreadas  $(\mathbf{s}_i, t_i)$ , o donde hay un interés en predecir en un soporte diferente o cuadrícula (Lloyd 2010). Para utilizar el variograma en kriging, se le debe ajustar a éste un modelo matemático, de tal manera que los coeficientes puedan luego utilizarse en el sistema de ecuaciones kriging. En este capítulo, se trabaja con el modelo métrico espacio-temporal, este modelo espacio-temporal tiene covarianza estacionaria y anisotropía geométrica en  $\mathbb{R}^d \times \mathbb{R}$ . Por lo tanto, una métrica en el espacio-tiempo que utiliza directamente modelos isotrópicos se define como

$$C_{st}(\delta_s, \delta_t) = C(\delta_{st}^2) = C(q_1^2\delta_s^2 + q_2^2\delta_t^2) \quad (5.15)$$

donde  $\delta_{st}^2 = q_1^2\delta_s^2 + q_2^2\delta_t^2$ ,  $q_1, q_2 \in \mathbb{R}$  son las constantes que definen la métrica espacio-temporal,  $\delta_s$  y  $\delta_t$  son las usuales distancias Euclidianas en espacio y tiempo, respectivamente. Este modelo supone la misma estructura de dependencia en el espacio y el tiempo, y sólo permitir cambios en el rango de las dos funciones de covarianza. Algunas aplicaciones de este modelo se pueden encontrar en Armstrong & Hubert (1993) y Snepvangers & Huisman (2003).

Finalmente, en esta subsección se presenta en la Tabla 5.1 algunas RBFs espacio-temporales consideradas en esta investigación y que utilizan el enfoque basado en distancias. El parámetro de suavizamiento óptimo  $\eta$ , el cual es un parámetro de libre elección, se encuentra al minimizar la raíz del cuadrado medio del error de predicción (RMSPE) haciendo uso de la validación cruzada.

TABLA 5.1: Formas funcionales de algunas RBFs espacio-temporales

RBF	Forma funcional	RBF	Forma funcional
EXP	$\phi(\delta_{st}) = e^{-\eta\delta_{st}}, \quad \eta \neq 0$	GAU	$\phi(\delta_{st}) = e^{-\eta\delta_{st}^2}, \quad \eta \neq 0$
MQ	$\phi(\delta_{st}) = \sqrt{\eta^2 + \delta_{st}^2}, \quad \eta \neq 0$	IMQ	$\phi(\delta_{st}) = 1/\sqrt{\eta^2 + \delta_{st}^2}, \quad \eta \neq 0$
RBF	Forma funcional		
TPS	$\phi(\delta_{st}) = \begin{cases} (\eta \cdot \delta_{st})^2 \log(\eta \cdot \delta_{st}) & \text{si } \delta_{st} \neq 0, \eta > 0 \\ 0 & \text{si } \delta_{st} = 0 \end{cases}$		
CRS	$\phi(\delta_{st}) = \begin{cases} \ln(\eta \cdot \delta_{st}/2)^2 + E_1(\eta \cdot \delta_{st}/2)^2 + C_E & \text{si } \delta_{st} \neq 0, \eta > 0 \\ 0 & \text{si } \delta_{st} = 0 \end{cases}$ <p>donde <math>\ln</math> es el logaritmo natural, <math>E_1(x)</math> es la función integral exponencial y <math>C_E</math> es la constante de Euler.</p>		
ST	$\phi(\delta_{st}) = \begin{cases} \ln(\eta \cdot \delta_{st}/2) + K_0(\eta \cdot \delta_{st}) + C_E & \text{si } \delta_{st} \neq 0 \\ 0 & \text{si } \delta_{st} = 0 \end{cases}$ <p>donde <math>K_0(x)</math> es la función modificada de Bessel y <math>C_E</math> es la constante de Euler.</p>		

## 5.2.2 Predicción espacio-temporal usando funciones de base radial basada en distancias

Una vez se han estimado los parámetros  $\boldsymbol{\nu}_{st}$  y  $\boldsymbol{\omega}_{st}$ , se pueden discutir las técnicas espacio-temporales para predecir el valor en una determinada localización,  $(\mathbf{s}_0, t_0)$ , a partir de las observaciones más cercanas y donde se han observado un conjunto de variables explicativas mixtas. Para conseguirlo, se utiliza el método kriging universal con la finalidad de construir a partir de la tendencia basada en distancias las predicciones espacio-temporales.

Ahora, las coordenadas  $x(\mathbf{s}_0, t_0) = (x_1(\mathbf{s}_0, t_0), \dots, x_k(\mathbf{s}_0, t_0))'$  se obtienen suponiendo que las observaciones de las variables explicativas mix-

tas están disponibles para un nuevo individuo, es decir,  $v(\mathbf{s}_0, t_0) = (v_1(\mathbf{s}_0, t_0), \dots, v_p(\mathbf{s}_0, t_0))$  es conocido. Luego, se pueden calcular las distancias entre el nuevo individuo y cada uno de los individuos involucrados en el modelo (5.1), es decir,  $\delta_{0i} = \delta(v(\mathbf{s}_0, t_0), v(\mathbf{s}_i, t_i))$ ,  $i = 1, \dots, n$ . A partir de estas distancias, una predicción puede hacerse usando un resultado propuesto por Gower (1971) y Cuadras & Arenas (1990), que relaciona el vector  $\boldsymbol{\delta}_0 = (\delta_{01}^2, \dots, \delta_{0n}^2)'$  de los cuadrados de las distancias y el vector  $x(\mathbf{s}_0, t_0)$  de coordenadas principales asociado al nuevo individuo mediante la expresión

$$\delta_{0i}^2 = [x(\mathbf{s}_0, t_0) - x(\mathbf{s}_i, t_i)]' [x(\mathbf{s}_0, t_0) - x(\mathbf{s}_i, t_i)]$$

donde  $x(\mathbf{s}_i, t_i) = (x_1(\mathbf{s}_i, t_i), \dots, x_k(\mathbf{s}_i, t_i))'$ ,  $i = 1, \dots, n$ . A continuación, se encuentra que

$$x(\mathbf{s}_0, t_0) = \frac{1}{2} \Lambda^{-1} X'(\mathbf{b} - \boldsymbol{\delta}_0)$$

donde  $\mathbf{b} = (b_{11}, \dots, b_{nn})$  y  $b_{ii} = x'(\mathbf{s}_i, t_i) x(\mathbf{s}_i, t_i)$ ,  $i = 1, \dots, n$ .

Ahora, el siguiente objetivo es predecir el valor de  $Z(\mathbf{s}_0, t_0)$  basado en un conjunto de observaciones  $\mathbf{Z}_{st}$ . Para ello, el predictor de la RBF está dado por

$$\hat{Z}(\mathbf{s}_0, t_0) = \hat{g}(\mathbf{s}_0, t_0) = \sum_{i=1}^n \varphi_i Z(\mathbf{s}_i, t_i) = \boldsymbol{\varphi}' \mathbf{Z}_{st} \quad (5.16)$$

sujeto a

$$\sum_{i=1}^n \varphi_i x_l(\mathbf{s}_i, t_i) = \boldsymbol{\varphi}'_{st} X_l = x_l(\mathbf{s}_0, t_0), \quad l = 0, \dots, k$$

donde  $\boldsymbol{\varphi}_{st} = (\varphi_1, \dots, \varphi_n)'$ ,  $\mathbf{Z}_{st} = (Z(\mathbf{s}_1, t_1), \dots, Z(\mathbf{s}_n, t_n))'$  y  $X_l = (x_l(\mathbf{s}_1, t_1), \dots, x_l(\mathbf{s}_n, t_n))$ .

El error esperado es igual a cero

$$E\left(\hat{Z}(\mathbf{s}_0, t_0) - Z(\mathbf{s}_0, t_0)\right) = 0$$

y el error cuadrático medio de la predicción del krigeado,  $\sigma_K^2$ , al utilizar la aproximación con funciones de base radial esta dado por

$$\begin{aligned} \sigma_K^2(\mathbf{s}, t) &= E\left\{\left[\hat{Z}(\mathbf{s}_0, t_0) - Z(\mathbf{s}_0, t_0)\right]^2\right\} \\ &\cong \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \phi(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) - 2 \sum_{i=1}^n \varphi_i \phi(\mathbf{s}_i - \mathbf{s}_0, t_i - t_0) \\ &\cong \boldsymbol{\varphi}'_{st} \Phi_{st} \boldsymbol{\varphi}_{st} - 2 \boldsymbol{\varphi}'_{st} \boldsymbol{\phi}_0 \end{aligned} \quad (5.17)$$

donde  $\phi_0 = (\phi(\mathbf{s}_1 - \mathbf{s}_0, t_1 - t_0), \dots, \phi(\mathbf{s}_n - \mathbf{s}_0, t_n - t_0))'$  y  $\Phi_{st}$  es una matriz  $n \times n$  con elementos  $\phi(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j)$ ,  $i, j = 1, \dots, n$ . Además,  $\phi_0$  corresponde al vector de función de base radial evaluado entre los vecinos y el punto donde se quiere predecir, es decir  $\phi(\mathbf{s}_i - \mathbf{s}_0, t_i - t_0)$ .

Los pesos se determinan minimizando la siguiente expresión penalizada

$$l(\varphi_{st}, \alpha_{st}) = \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j \phi(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j) - 2 \sum_{i=1}^n \varphi_i \phi(\mathbf{s}_i - \mathbf{s}_0, t_i - t_0) \\ + \rho \int_{\mathbb{R}^d \times \mathbb{R}} J_m(g(\mathbf{s}, t)) d(\mathbf{s}, t) + 2 \sum_{l=0}^k \alpha_l \left( \sum_{i=1}^n \varphi_i x_l(\mathbf{s}_i, t_i) - x_l(\mathbf{s}_0, t_0) \right)$$

donde  $\alpha_{st} = (\gamma_0, \dots, \gamma_k)'$  es el vector de  $(k+1)$  multiplicadores de Lagrange asociados con la restricción de insesgamiento.

En la forma matricial, la expresión anterior se convierte en

$$l(\varphi_{st}, \alpha_{st}) = \varphi'_{st} (\Phi_{st} + \rho I) \varphi_{st} - 2\varphi'_{st} \phi_0 + 2\alpha'_{st} (\mathbf{X}'_{st} \varphi_{st} - \mathbf{x}(\mathbf{s}_0, t_0))$$

donde  $\mathbf{X}_{st}$  fue definida en (5.5), y  $\mathbf{x}(\mathbf{s}_0, t_0) = (1, x'(\mathbf{s}_0, t_0))' = (1, x_1(\mathbf{s}_0, t_0), \dots, x_k(\mathbf{s}_0, t_0))'$ .

Después de diferenciar con respecto a  $\varphi_{st}$  y  $\gamma_{st}$ , igualando el resultado a cero y realizando algunos procedimientos algebraicos, el siguiente sistema matricial se encuentra

$$\begin{pmatrix} \Phi_{st} + \rho I & \mathbf{X}_{st} \\ \mathbf{X}'_{st} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \varphi_{st} \\ \alpha_{st} \end{pmatrix} = \begin{pmatrix} \phi_0 \\ \mathbf{x}(\mathbf{s}_0, t_0) \end{pmatrix} \quad (5.18)$$

Resolviendo el sistema, los coeficientes para  $\varphi_{st}$  y  $\alpha_{st}$  están dadas por

$$\widehat{\varphi}'_{st} = \{ \phi_0 + \mathbf{X}_{st} [\mathbf{X}'_{st} (\Phi_{st} + \rho I)^{-1} \mathbf{X}_{st}]^{-1} [\mathbf{x}(\mathbf{s}_0, t_0) \\ - \mathbf{X}'_{st} (\Phi_{st} + \rho I)^{-1} \phi_0] \}' (\Phi_{st} + \rho I)^{-1} \quad (5.19) \\ \widehat{\alpha}_{st} = - [\mathbf{X}'_{st} (\Phi_{st} + \rho I)^{-1} \mathbf{X}_{st}]^{-1} [\mathbf{x}(\mathbf{s}_0, t_0) - \mathbf{X}'_{st} (\Phi_{st} + \rho I)^{-1} \phi_0]$$

Por otro lado, para obtener una expresión aproximada del error cuadrático de la predicción, se premultiplica la parte superior de (5.18) por  $\varphi'_{st}$  y se encuentra que  $\varphi'_{st} (\Phi_{st} + \rho I) \varphi_{st} + \varphi'_{st} \mathbf{X}_{st} \alpha_{st} = \varphi'_{st} \phi_0$ , éste término se reemplaza

en la expresión (5.17) y se llega a

$$\begin{aligned}\sigma_K^2(\mathbf{s}_0, t_0) &\cong -\boldsymbol{\varphi}'_{st}\Phi_{st}\boldsymbol{\varphi}_{st} + 2\boldsymbol{\varphi}'_{st}\boldsymbol{\phi}_0 \\ &\cong -\boldsymbol{\varphi}'_{st}\boldsymbol{\phi}_0 + \rho\boldsymbol{\varphi}'_{st}\boldsymbol{\varphi}_{st} + \boldsymbol{\varphi}'_{st}\mathbf{X}_{st}\boldsymbol{\alpha}_{st} + 2\boldsymbol{\varphi}'_{st}\boldsymbol{\phi}_0 \\ &\cong \boldsymbol{\varphi}'_{st}\boldsymbol{\phi}_0 + \rho\boldsymbol{\varphi}'_{st}\boldsymbol{\varphi}_{st} + \mathbf{x}'_{st}(\mathbf{s}_0, t_0)\boldsymbol{\alpha}_{st}\end{aligned}$$

donde  $\mathbf{X}'_{st}\boldsymbol{\varphi}_{st} = \mathbf{x}(\mathbf{s}_0, t_0)$ .

Una vez estimados  $\boldsymbol{\varphi}_{st}$  y  $\boldsymbol{\alpha}_{st}$  en (5.19), una expresión aproximada del error cuadrático de la predicción estimado se puede escribir como

$$\widehat{\sigma}_K^2(\mathbf{s}_0, t_0) \cong \sum_{i=1}^n \widehat{\varphi}_i \phi(\mathbf{s}_i - \mathbf{s}_0, t_i - t_0) + \rho \sum_{i=1}^n \widehat{\varphi}_i^2 + \sum_{l=0}^k \widehat{\alpha}_l x_l(\mathbf{s}_0, t_0) \quad (5.20)$$

El procedimiento presentado en esta sección se puede resumir en los siguientes pasos:

1. Obtener las coordenadas principales utilizando la descomposición espectral de la matriz de similitudes (o distancias) calculada a partir de las variables explicativas.
2. Seleccionar las coordenadas principales más correlacionadas o significativas con la variable regionalizada  $\mathbf{Z}_{st}$ . En este paso, se recomienda utilizar el criterio dado en (3.5) para hacer una primera selección con el fin de remover las coordenadas principales pobremente correlacionadas con las variable regionalizada, y luego, emplear los criterios (3.6) o (3.7) para seleccionar las coordenadas principales mas significativas utilizando la regresión DB.
3. Optimizar los parámetros  $\eta$  del interpolador espacio-temporal basado en distancias usando funciones de base radial (DBSTIRBF) y  $\rho$ , por medio de validación cruzada (leave-one-out) mediante el uso de la expresión

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^n \left( \widehat{Z}_{[i]}(\mathbf{s}_i, t_i) - Z(\mathbf{s}_i, t_i) \right)^2}{n}} \quad (5.21)$$

y empleando las expresiones (5.12) y (5.13) en los diferentes vecindarios de un tamaño prefijado. En la expresión (5.21),  $\widehat{Z}_{[i]}(\mathbf{s}_i, t_i)$  es el

valor predicho obtenido de la validación cruzada y  $Z(\mathbf{s}_i, t_i)$  es el valor muestreado en la localización  $(\mathbf{s}_i, t_i)$ . El tamaño del vecindario,  $n_h$ , también se puede escoger dentro del mismo proceso de optimización.

4. Hacer las predicciones en los puntos muestreados y no muestreados para generar el mapa de predicción usando el método DBSTIRBF, es decir, haciendo  $\hat{Z}(\mathbf{s}_0, t_0) = \hat{\varphi}'_{st} \mathbf{Z}_{st}$ .

En el caso en que se desee evaluar el ajuste de la DBSTIRBF o comparar ajustes entre DBSTIRBF se recomienda hacer uso de LOOCV, empleando también la expresión (5.21).

### 5.3 Estudio de simulación y discusión

Al igual que en la Sección 4.3, en esta sección se describe un estudio de simulación para evaluar la eficacia del método propuesto, DBSTIRBF, bajo diferentes condiciones asociadas a los parámetros de suavizado y las funciones de base radial. En particular, se estudian los efectos de: (i) el nivel de ruido, (ii) la densidad de diseño, (iii) el grado de variación espacio-temporal y (iv) la función de varianza. Estas configuraciones y escenarios se presentan en la Tabla 5.2.

Este estudio considera las simulaciones en tres dimensiones  $(w_x, w_y, t)$ , además una variable aleatoria binomial  $V_1 \sim Bi(n, 0.4)$ , tamaños de muestra  $n = 150$  y  $n = 250$  asociados a 25 puntos en el espacio en cada uno de los casos y 6 y 10 puntos en el tiempo, respectivamente, tamaños de vecindario  $n_h = 8, 32$ , parámetros de suavizamiento  $\eta = 0.01, 0.1$  y  $j = 1, 3$  en el factor de varianza. Adicionalmente, se asume una variable nominal asociada a tres regiones específicas en el cuadrado de longitud uno, como se muestra en la Figura 5.1. Dado que hay tres regiones, se consideran sólo dos variables dummy ( $D_2$  y  $D_3$ ) para evitar problemas de singularidad. Además,  $\varepsilon(\mathbf{s}_i, t_i)$  se construye asumiendo un campo aleatorio Gaussiano, para un modelo espacio-temporal no separable asociado con una pepita  $\tau^2 = 1$  y media 0. Para los parámetros de tendencia, se asume los siguientes valores  $\beta_0 = 10$ ,  $\beta_1 = -4$ ,

TABLA 5.2: Escenarios considerados en los experimentos simulados espacio-temporales

Factor	Forma genérica
Nivel de ruido	$z_j(\mathbf{s}_i, t_i) = \beta_0 + \beta_1 V_i + \beta_2 D_{i2} + \beta_3 D_{i3} + \beta_4 t_i + f(w_{x_i}) + f(w_{y_i}) + f(t_i) + f(w_{x_i})f(w_{y_i}) + f(w_{x_i})f(t_i) + f(w_{y_i})f(t_i) + \sigma_j \varepsilon(\mathbf{s}_i, t_i)$ donde $\sigma_j = 0.02 + 0.04(j - 1)^2$
Densidad de diseño	$z_j(\mathbf{s}_i, t_i) = \beta_0 + \beta_1 V_i + \beta_2 D_{i2} + \beta_3 D_{i3} + \beta_4 t_i + f(X_{ji}) + f(Y_{ji}) + f(T_{ji}) + f(X_{ji})f(Y_{ji}) + f(X_{ji})f(T_{ji}) + f(Y_{ji})f(T_{ji}) + \sigma \varepsilon(\mathbf{s}_i, t_i)$ donde $\sigma = 0.1$ , $X_{ji} = F_j^{-1}(X_i)$ , $Y_{ji} = F_j^{-1}(Y_i)$ , $T_{ji} = F_j^{-1}(T_i)$ con $T_i = t_i/t_{max}$ , $t_{max} = 6, 10$
Variación espacio-temporal	$z_j(\mathbf{s}_i, t_i) = \beta_0 + \beta_1 V_i + \beta_2 D_{i2} + \beta_3 D_{i3} + \beta_4 t_i + f(w_{x_i}) + f(w_{y_i}) + f(t_i) + f(w_{x_i})f(w_{y_i}) + f(w_{x_i})f(t_i) + f(w_{y_i})f(t_i) + \sigma \varepsilon(\mathbf{s}_i, t_i)$ donde $\sigma = 0.2$ , $f_j(l_i) = \sqrt{l_i(1-l_i)} \sin \left[ \frac{2\pi \left\{ 1 + 2^{(9-4j)/5} \right\}}{l_i + 2^{(9-4j)/5}} \right]$
Función de variación	$z_j(\mathbf{s}_i, t_i) = \beta_0 + \beta_1 V_i + \beta_2 D_{i2} + \beta_3 D_{i3} + \beta_4 t_i + f(w_{x_i}) + f(w_{y_i}) + f(w_{x_i})f(w_{y_i}) + \sqrt{\varsigma_1 + \varsigma_2 + \varsigma_3 + \varsigma_1 \varsigma_2 + \varsigma_1 \varsigma_3 + \varsigma_2 \varsigma_3} \varepsilon(\mathbf{s}_i, t_i)$ donde $\varsigma_1 = v_j(w_{x_i})$ , $\varsigma_2 = v_j(w_{y_i})$ , $\varsigma_3 = v_j(t_i)$ , $v_j(l_i) = \{0.15 [1 + 0.4(2j - 7)(l_i - 0.5)]\}^2$
Los supuestos y otras elecciones	
$V_i \sim Binomial(n, 0.4); \varepsilon(\mathbf{s}_i, t_i) \stackrel{iid}{\sim} N(0, 0.1); n = 150 \text{ (25 puntos en el espacio y 6 puntos en el tiempo)}$ y $n = 250$ (25 puntos en el espacio y 10 puntos en el tiempo); $F_j$ es la $Beta \left( \frac{j+4}{5}, \frac{11-j}{5} \right)$ ; $j = 1, 3$ ; $f(l_i) = 1.5 f_1 \left( \frac{l_i - 0.5}{0.15} \right) - f_1 \left( \frac{l_i - 0.8}{0.04} \right)$ ; $f_1(u) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{u^2}{2} \right)$ ; $X_i, Y_i, T_i \stackrel{iid}{\sim} Uniform(0, 1)$ ; $l_i = w_{x_i}, w_{y_i}, t_i; i = 1, \dots, n$	

$\beta_2 = 2$  y  $\beta_3 = -4$ , con  $w_{x_i}$  y  $w_{y_i}$  asociados a las coordenadas espaciales, y  $t_i$  asociado al tiempo, donde  $i$  es la  $i$ -ésima observación simulada. En la Tabla 5.3, se presentan los escenarios simulados. El método propuesto se prueba con cinco RBFs (MQ, TPS, CRS, ST y EXP), considerando el modelo métrico dado en (5.15) con  $q_1 = q_2 = 1$ . Un total de 80 escenarios fueron simulados, y para cada uno de ellos, el proceso se repitió 100 veces.

Para cada conjunto de datos simulados, se evaluó la calidad del ajuste con el RMSPE obtenido mediante el método de validación cruzada (leave-one-out). Los resultados se muestran en las Tablas 5.4 y 5.5. Inicialmente se consideró usar un parámetro positivo para  $\rho$ , pero los valores de RMSPE no mostraron diferencias significativas con los obtenidos cuando  $\rho = 0$ ; en particular, cuando las funciones de base radial MQ, EXP, CRS y GAU fueron utilizadas.

Las Tablas 5.4 y 5.5 muestran los valores medios de RMSPEs obtenidos de 100 simulaciones por caso y para los 80 casos descritos en la Tabla 5.3. El método DBSTIRBF funciona bien para vecindarios grandes, lo que indica

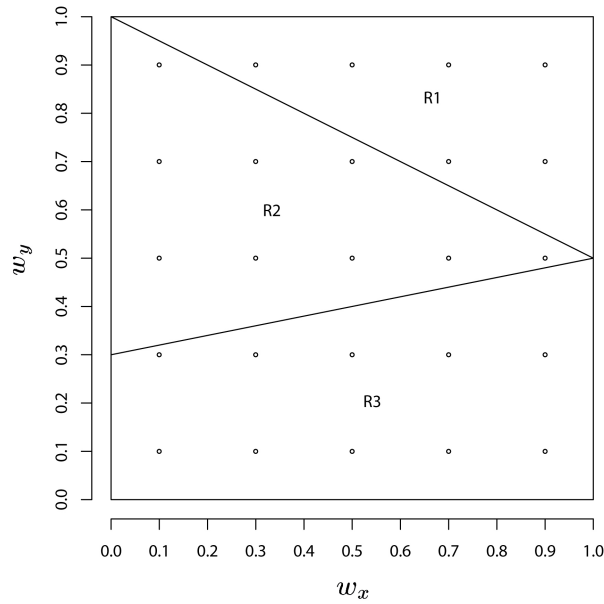


FIGURA 5.1: Localización de los puntos de muestreo y regiones asociadas a la definición de la variable nominal

TABLA 5.3: Escenarios espacio-temporales simulados (los números naturales en las últimas cinco columnas (de 1 a 80) representan el número del escenario)

Parámetros del modelo			$n$	Función de base radial				
$\eta$	$j$	$n_h$		MQ	TPS	CRS	ST	EXP
0.01	1	8	150	1	17	33	49	65
			250	2	18	34	50	66
		32	150	3	19	35	51	67
	250		4	20	36	52	68	
	3	8	150	5	21	37	53	69
			250	6	22	38	54	70
32		150	7	23	39	55	71	
	250	8	24	40	56	72		
0.1	1	8	150	9	25	41	57	73
			250	10	26	42	58	74
		32	150	11	27	43	59	75
			250	12	28	44	60	76
	3	8	150	13	29	45	61	77
			250	14	30	46	62	78
		32	150	15	31	47	63	79
			250	16	32	48	64	80

una ganancia (vista en un decrecimiento) de 70.6% de los valores medios de



TABLA 5.4: Promedios de RMSPEs bajo el método DBSTIRBF en los escenarios espacio-temporales presentados en la Tabla 5.3 (casos nivel de ruido y densidad de diseño)

Parámetro			n	Nivel de ruido					Densidad de diseño					
$\eta$	j	$n_h$		MQ	TPS	CRS	ST	EXP	MQ	TPS	CRS	ST	EXP	
0.01	1	8	150	2.49	2.48	2.49	2.48	2.48	2.61	8.06	8.05	8.05	8.05	
			250	1.71	1.71	1.70	1.70	1.70	1.63	1.67	1.66	1.66	1.65	
		32	150	1.58	1.78	1.41	1.60	1.56	1.50	1.75	1.33	1.56	1.48	
			250	1.54	1.58	1.56	1.55	1.54	1.42	1.53	1.29	1.46	1.41	
		3	8	150	2.55	2.54	2.55	2.55	2.55	3.43	3.56	3.56	3.56	3.55
				250	1.75	1.74	1.74	1.74	1.74	1.67	11.77	11.85	11.77	5.03
	32		150	1.59	1.78	1.44	1.61	1.57	1.48	1.72	1.35	1.54	1.46	
			250	1.54	1.58	1.55	1.55	1.54	1.44	1.53	1.38	1.48	1.44	
	0.1	1	8	150	2.49	2.48	2.52	2.48	2.49	2.60	8.05	8.13	8.06	2.60
				250	1.71	1.71	1.70	1.71	1.70	1.64	1.66	2.10	1.66	1.63
			32	150	1.82	1.79	1.54	1.78	1.56	1.74	1.74	1.39	1.74	1.48
				250	1.59	1.58	1.56	1.58	1.54	1.52	1.52	25.56	1.52	1.41
3			8	150	2.55	2.54	2.58	2.54	2.55	3.45	3.56	3.64	3.56	3.45
				250	1.75	1.74	1.74	1.74	1.74	1.92	5.69	11.84	5.69	1.92
		32	150	1.82	1.79	1.52	1.78	1.57	1.72	1.72	1.44	1.72	1.46	
			250	1.59	1.58	1.56	1.58	1.53	1.53	1.53	1.49	1.53	1.44	

RMSPE cuando  $n_h = 32$  con respecto a  $n_h = 8$ . Cuando  $\eta = 0.1$ , hay en general una ligera reducción de 2.5% comparado con  $\eta = 0.01$  en los valores medios de RMSPE. Teniendo en cuenta el parámetro  $j$ , hubo en general una ligera reducción (pérdida) de 1.43% en los valores medios de RMSPE cuando  $j = 3$  comparado con  $j = 1$ . Mientras que para cuando  $n = 150$ , los valores medios de RMSPE fueron 15% más grandes que los obtenidos cuando  $n = 250$ .

En cuanto a los escenarios asociados a las formas genéricas, los valores más bajos de RMSPE correspondieron a los casos de nivel de ruido y función de varianza, con valores medios de RMSPE de 1.86 y 1.90, respectivamente. Para los casos, densidad del diseño y función de varianza espacio-temporal, los valores medios de RMSPE fueron de 3.38 y 2.16, respectivamente. En cuanto al método DBSTIRBF, se encuentra que: i) para el caso *el nivel de ruido*, el método DBSTIRBF que produjo el valor promedio de RMSPE más bajo fue la CRS, mientras que la TPS presentó el más alto, ii) para el caso *densidad del diseño*, el método DBSTIRBF con promedio de RMSPE más

TABLA 5.5: Promedios de RMSPEs bajo el método de DBSTIRBF de los escenarios espacio-temporales presentados en la Tabla 5.3 (casos variación espacio-temporal y función de varianza)

Parámetro			$n$	Variación espacio-temporal					Función de varianza					
$\eta$	$j$	$n_h$		MQ	TPS	CRS	ST	EXP	MQ	TPS	CRS	ST	EXP	
0.01	1	8	150	3.00	3.77	3.78	3.77	3.78	2.61	2.60	2.61	2.60	2.61	
			250	1.74	1.79	1.78	1.79	1.76	1.81	1.84	1.83	1.83	1.83	
		32	150	1.49	1.74	1.36	1.55	1.48	1.56	1.76	1.45	1.60	1.55	
			250	1.45	1.54	1.39	1.48	1.44	1.50	1.57	1.51	1.53	1.50	
		3	8	150	2.55	3.35	3.34	3.34	2.94	2.60	2.59	2.60	2.59	2.60
				250	1.88	3.05	3.04	3.04	2.02	1.80	1.81	1.81	1.81	1.81
	32		150	1.55	1.85	1.37	1.64	1.53	1.56	1.76	1.44	1.60	1.55	
			250	1.52	1.64	1.41	1.57	1.51	1.50	1.57	1.51	1.53	1.50	
	0.1	1	8	150	3.00	3.77	3.81	3.77	3.01	2.60	2.60	2.64	2.60	2.61
				250	1.74	1.78	1.78	1.79	1.75	1.83	1.84	1.83	1.84	1.83
			32	150	1.73	1.73	1.48	1.73	1.48	1.78	1.77	1.53	1.76	1.55
				250	1.53	1.53	1.52	1.53	1.44	1.57	1.57	1.55	1.57	1.50
3			8	150	2.57	3.36	3.42	3.35	2.56	2.59	2.59	2.62	2.59	2.59
				250	1.94	3.04	3.17	3.05	1.95	1.81	1.81	1.80	1.81	1.80
		32	150	1.83	1.84	1.52	1.84	1.52	1.78	1.77	1.53	1.76	1.55	
			250	1.64	1.64	1.58	1.64	1.51	1.57	1.57	1.55	1.57	1.49	

pequeño fue el construido con la MQ, mientras que con CRS mostró una vez más el RMSPE más alto, iii) en términos de *variación espacio-temporal*, el valor promedio de RMSPE más bajo se observó con la MQ, mientras que la TPS ha mostrado una vez los valores medios de RMSPE más altos, y iv) en el caso *función de varianza* el método DBSTIRBF con mejores resultados en términos de promedios de RMSPE fue el construido con la CRS, y otra vez, el método DBSTIRBF construido con la función TPS muestra los peores resultados.

Dado que los valores promedios de RMSPEs fueron mayores en aquellos casos con un tamaño de muestra  $n = 150$  con respecto a los casos con el tamaño de la muestra  $n = 250$ , estos casos se muestran en diagramas de caja por separado, véanse las Figuras 5.2 y 5.3. En estos gráficos, se encuentra que los valores medios de RMSPE son menores para  $n_h = 32$  con respecto a  $n_h = 8$ . Además, de acuerdo a la Figura 5.2, se tiene lo siguiente: i) en términos de *nivel de ruido*, se encontró una menor variabilidad cuando  $j = 1$  y  $n_h = 8$ , mientras que cuando  $n_h = 32$  se encuentran menores valores de RMSPE, ii)

en el caso de *densidad del diseño*, el método DBSTIRBF que mostró la mayor variabilidad fue el TPS, mientras que las funciones de base radial MQ y EXP muestran una menor variabilidad, especialmente para  $j = 1$  y  $n_h = 32$ , iii) bajo el escenario de *variación espacio-temporal*, cuando  $j = 1$  y  $n_h = 8$  se observa en general una gran variabilidad, a excepción de la función base radial EXP cuando  $\eta = 0.1$ , y iv) para el caso *función de varianza*, la variabilidad fue similar en todas las funciones de prueba. En general, los valores promedio de RMSPE más grandes se encontraron para el tamaño de vecindario  $n_h = 8$ .

De acuerdo a la Figura 5.3, se nota que: i) cuando se considera el *nivel de ruido*, hay una mayor variabilidad con  $j = 3$ , incrementándose el valor promedio de RMSPE con  $n_h = 8$ ; ii) en el caso *densidad del diseño*, las funciones de base radial TPS, CRS y ST muestran mayor variabilidad con  $j = 3$  y  $n_h = 8$ , independientemente de  $\eta$ , mientras que para las funciones de base radial MQ y EXP, las mayores variabilidades se muestran para  $j = 3$  y  $n_h = 8$  pero con  $\eta = 0.1$  y  $\eta = 0.01$ , respectivamente, iii) en el caso de *variación espacio-temporal*, la mayor variabilidad fue para  $j = 3$  y  $n_h = 8$ , sobre todo para las funciones de base radial ST y TPS, y la menor variabilidad fue para la función de base radial CRS, y iv) para el caso *función de varianza*, la variabilidad fue menor para las funciones de base radial ST y TPS, en especial cuando  $n_h = 32$ . En general, el valor medio de RMSPE fue menor cuando  $n_h = 32$ .

## 5.4 Aplicación

El conjunto de datos empleado en esta aplicación corresponde al utilizado en la Subsección 3.4.1 del Capítulo 3 y que fue estudiado por Hengl (2009). En esta aplicación se analiza la temperatura media mensual terrestre en Croacia a partir de 155 estaciones meteorológicas. La temperatura media fue medida desde enero hasta diciembre de 2008. Tal como se explicó en la Sección 3.4.1 del Capítulo 3 en la mayoría de las estaciones meteorológicas, la temperatura se mide tres veces al día, a las 7 am, 1 pm y 9 pm, y la media de la temperatura diaria ( $\Delta T$  en un día) se calcula como un promedio ponderado, de acuerdo a

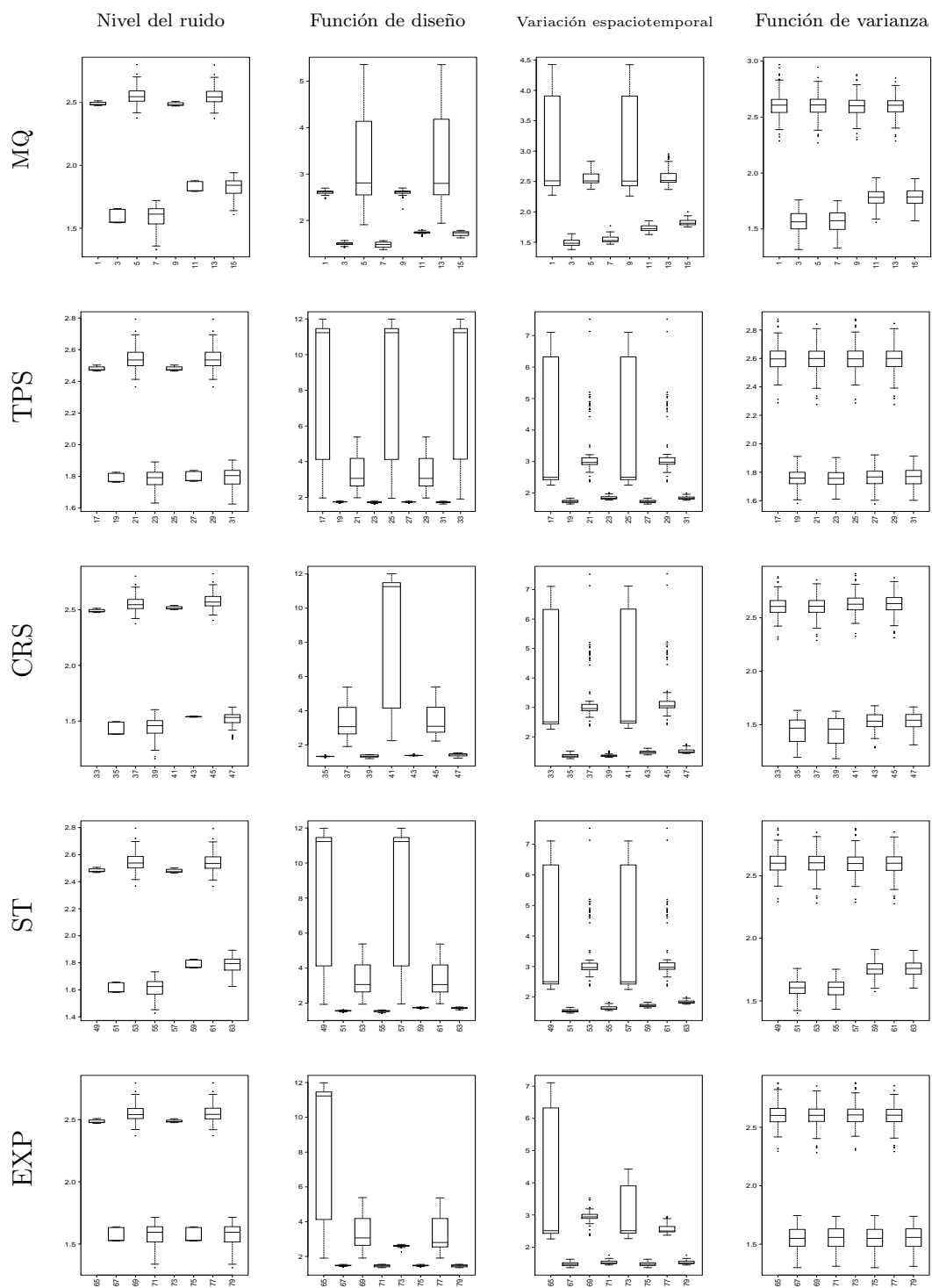


FIGURA 5.2: RMSPE para los escenarios espacio-temporales simulados con 6 tiempos

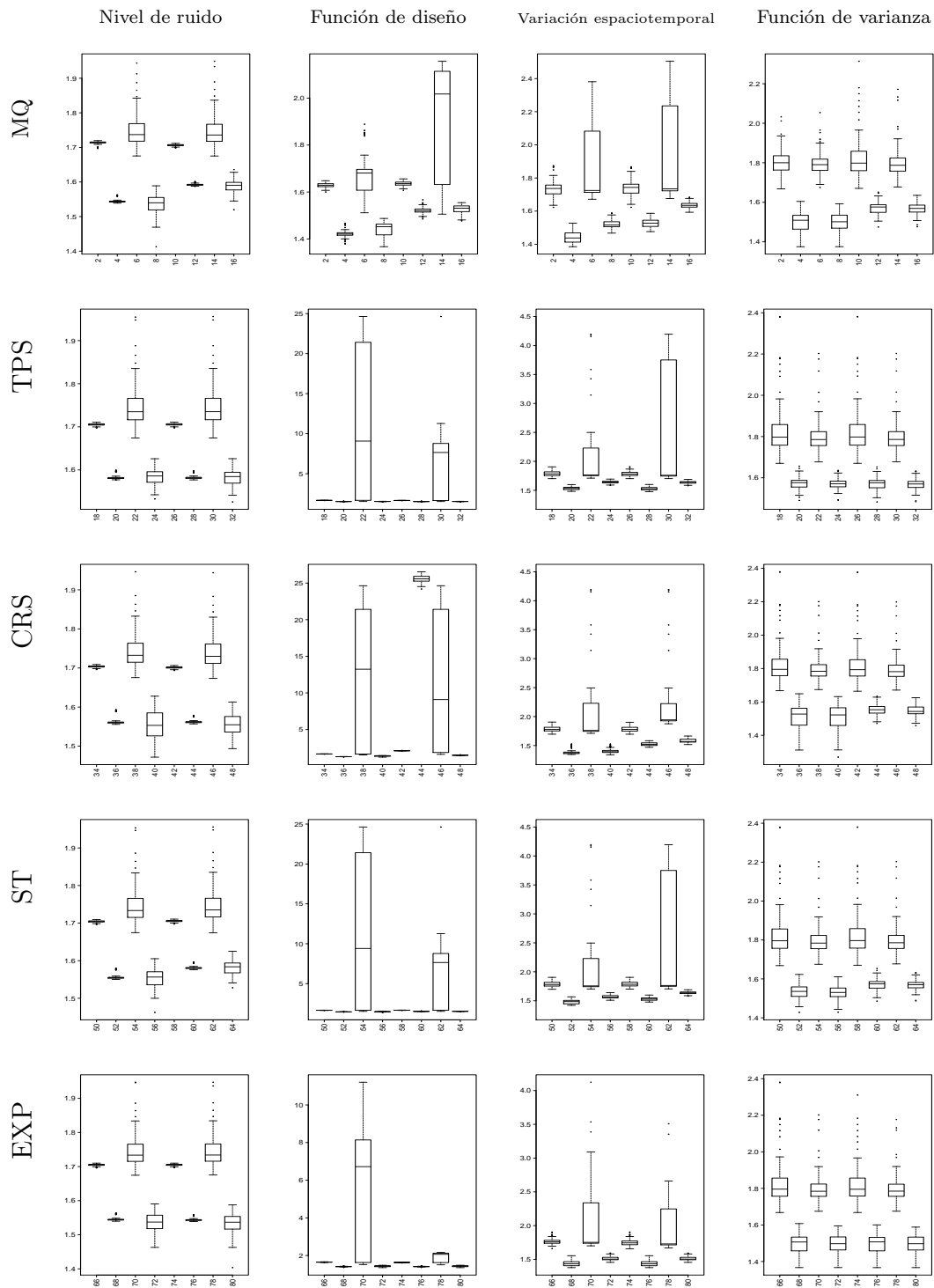


FIGURA 5.3: RMSPE para los escenarios espacio-temporales simulados con 10 tiempos

la siguiente expresión

$$\Delta T = \frac{T_{(7am)} + T_{(1pm)} + 2 \cdot T_{(9pm)}}{4}$$

Luego la temperatura media mensual se obtiene de la media diaria mencionada anteriormente, teniendo en cuenta que se dispone de una composición de imágenes (imágenes MODIS de 1 km de resolución, de 8 días, dispuestas al público) de la temperatura media diaria, es decir, de 3 a 4 registros mensuales. Las mediciones de temperatura se recogen automáticamente en 159 estaciones meteorológicas. Como cuatro estaciones meteorológicas no tenían registros disponibles para unos meses, entonces dichas estaciones se retiraron del análisis. Por lo tanto, se consideran sólo 155 estaciones y se calcula con los valores observados removiendo los datos faltantes (o perdidos) la temperatura media mensual. La distribución espacial de las estaciones no es óptima (Zaninovic et al. 2008, Perčec Tadić 2010), hay un cierto submuestreo a mayor altitud y en áreas con menor densidad de población; por razones prácticas, las zonas de mayor densidad de población se les dio prioridad. Por lo tanto, se podría esperar que la precisión de la cartografía sea menor a mayor altitud y en las tierras altas (Hengl 2009).

Las coordenadas  $w_x$  y  $w_y$  se obtuvieron a partir de una transformación de coordenadas geográficas (latitud y longitud) a un sistema de coordenadas cartesianas. La localización de las 155 estaciones meteorológicas se muestra en la Figura 5.4(a) (localizaciones espaciales). Las coordenadas principales se calculan a partir de la descomposición espectral generada por: las coordenadas espaciales ( $w_x, w_y$ ), el mes, el modelo digital de elevación (DEM, en metros), la distancia topográfica ponderada desde la línea a la costa (DSEA, en km), el índice de humedad topográfica (TWI) y la estación climatológica del año. Tanto las coordenadas espaciales como el tiempo se estandarizaron para dar igual peso a todas las dimensiones (espacio-tiempo). Además, una regresión basada en distancias espacio-temporal se realizó con las coordenadas principales y la temperatura media mensual terrestre. En este proceso, se encontró que cada una de las 10 primeras coordenadas principales tenían una alta significancia estadística, a un nivel del 5%, con la temperatura media de terrestre. La regresión basada en distancias espacio-temporal explicó el 96.1% de la va-

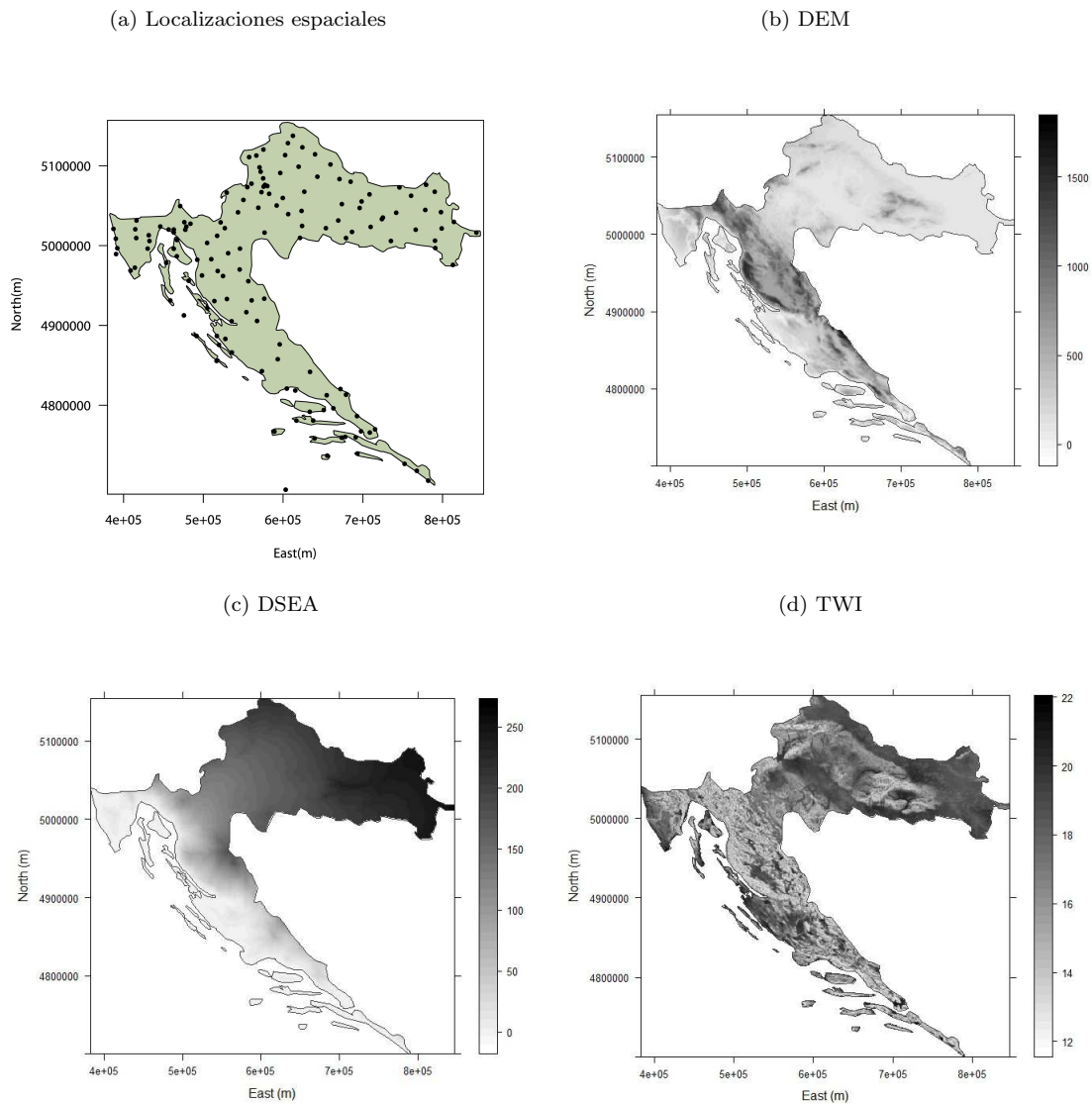


FIGURA 5.4: Localizaciones espaciales de las estaciones meteorológicas en Croacia y predictores estáticos topográficos: Modelo Digital de Elevación (DEM, en metros), la distancia topográfica ponderada desde la línea de costa (DSEA, en km) y el índice de humedad topográfica (TWI)

riabilidad de las temperaturas mensuales. Por lo tanto, estas 10 coordenadas principales se tuvieron en cuenta en la tendencia, y por último, LOOCV considerando los 30 vecinos mas cercanos se llevó a cabo para evaluar la calidad

en el ajuste del método DBSTIRBF, esto asociado al menor RMSPE.

Además, el método LOOCV utilizando el modelo métrico dado en (5.15) con los parámetros  $q_1 = q_2 = 1$  fue utilizado para comprobar el rendimiento de los seis modelos de esta aplicación. La Tabla 5.6 muestra los valores de RMSPE para los resultados de interpolación en las seis funciones de base radial espacio-temporales y las funciones de base radial que muestran los valores más pequeños de RMSPE son la CRS y la ST. Los mapas de predicción correspondientes se muestran en la Figura 5.5.

TABLA 5.6: Comparación de algunos métodos DBSTIRBF para las temperaturas promedios mensuales de 2008 en Croacia con LOOCV

	MQ	TPS	CRS	ST	EXP	GAU
$\eta$	0.001	0.001	0.200	0.001	0.001	0.010
$\rho$	0.000	0.001	0.000	0.100	0.000	0.000
RMSPE	2.333	2.284	2.242	2.251	2.311	2.274



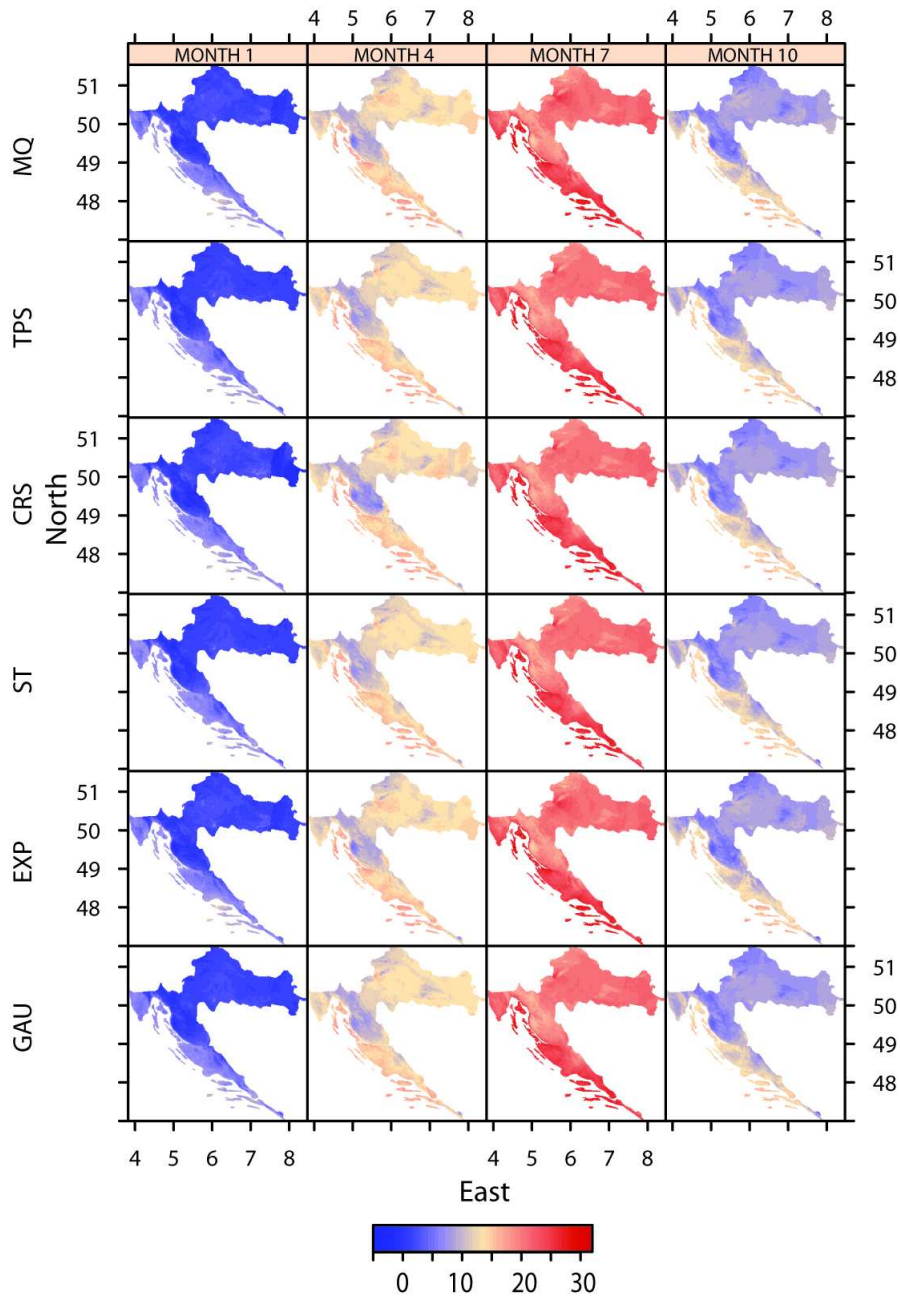


FIGURA 5.5: Mapas de predicción de la temperatura promedio mensual de la tierra en Croacia bajo el método DBSTIRBF en enero, abril, julio y octubre (unidades de las coordenadas este y norte en 100.000 metros)

# Capítulo 6

## Funciones geoestadísticas y funciones de base radial en el programa R: Paquete geospt

### 6.1 Introducción

El presente capítulo presenta una serie de funciones desarrolladas e implementadas en el programa estadístico R. En cuanto a las funciones de base radial implementadas y puestas en la librería **geospt**, estas no consideran tendencia, la cual es considerada en las funciones presentadas en los Capítulos 4 y 5 de la presente tesis, las funciones asociadas con la tendencia se presentan en el Apéndice A, las cuales esta fundamentalmente asociadas con las coordenadas principales. Se espera próximamente cargarlas con un instructivo en esta librería junto con las demás funciones implementadas en esta investigación.

Para realizar un análisis geoestadístico es necesario considerar una serie de pasos: un primer paso consiste en analizar la calidad y la cantidad de datos requeridos, es decir el muestreo espacial. El segundo paso es el análisis exploratorio, el cual se basa en el uso de técnicas estadísticas convencionales y en el análisis estructural de los datos, con el objetivo de identificar la presencia de anisotropía o isotropía y la tendencia. En el modelado del variograma, se

evalúa la relación espacial entre los valores de la variable regionalizada, y se ajusta un modelo de variograma al variograma experimental. Una vez el modelo de variograma se encuentra, los valores de predicción se pueden generar usando la interpolación Kriging para la construcción del mapa de predicción de la variable explicada. Sin embargo, hay métodos deterministas, donde el modelo de interpolación no requiere de un modelo de variograma, como es el caso de las RBFs explicadas en los capítulos previos. Después de esto, para elegir el mejor método de interpolación se utiliza la validación cruzada. El último paso consiste en la generación de los mapas de predicciones de la variable regionalizada y de las desviaciones estándar, junto con su interpretación y análisis.

Actualmente, la realización de estos procedimientos es viable gracias a los modernos programas informáticos existentes. Sin embargo, no se puede decir que exista un sólo programa informático que tenga implementadas todas las herramientas geoestadísticas, esto junto con la carencia de funciones en el programa R en cuanto a las funciones de base radial espaciales y espacio-temporales, y con respecto a el *pocket plot*, entre otras, motiva la realización de la librería expuesta aquí, la cual fue útil en el desarrollo de esta investigación.

Proponemos una serie de funciones que están diseñadas en el programa R. Estas permiten un análisis geoestadístico más completo junto con la ayuda de paquetes previamente diseñados en R, tales como: **geoR**, **gstat** y **sgeostat**, entre otros. De esta manera, estas contribuciones son: una función para la construcción del variograma experimental de la media recortada, una función para la construcción del *pocketplot* para datos grillados (útil para el análisis de estacionariedad local), y funciones de base radial (multicuadrática, multicuadrática inversa, spline con tensión, completamente regularizada spline y spline capa delgada) para optimizar, predecir y realizar validación cruzada en el espacio, una función para producir un gráfico que muestra el comportamiento del parámetro *ETA*, asociado con la función de base radial, y una función que genera una tabla con el resumen de las estadísticas de la validación cruzada para evaluar la exactitud de los métodos de interpolación (geoestadísticos y determinísticos) con base en los errores de predicción. Se describen breve-

mente algunas de las funciones, y luego se ilustra su funcionamiento con varios ejercicios. El paquete está implementado en el programa (R Development Core Team (2012)) y se encuentra disponible en el Comprehensive R Archive Network (CRAN) en <http://cran.r-project.org/web/packages/geospt>

## 6.2 Implementación de funciones geoestadísticas en R

En esta sección enfatizamos el uso de herramientas informáticas en el programa R, asociado con los conceptos teóricos definidos en algunas secciones anteriores.

### 6.2.1 Pocket plot

El Pocket Plot (llamado así debido a su uso en la detección de bolsillos de no estacionariedad) es una técnica necesaria para identificar un área localizada atípica con respecto al modelo de estacionariedad, es construida para aprovechar la naturaleza espacial de los datos a través de las coordenadas de filas y columnas (estas "x" y "y" respectivamente). Para la ilustración de este ejemplo, ver la siguiente Figura 6.1

En geoestadística se pretende estimar las relaciones espaciales entre los datos de los puntos (modelamiento del variograma). Luego este estimado es usado para el desarrollo del método kriging y para estimar la variabilidad del predictor. Aunque el estimador de Cressie & Hawkins (1980), ofrece una estimación robusta para el variograma, hay aun una fracción de las diferencias ( $Z_i - Z_j$ ), que resulta ser inapropiada en la estimación del variograma de Cressie. Las ubicaciones sobre la grilla que exhiben diferentes medidas del resto se deben identificar. Estos bolsillos de no estacionariedad, una vez descubiertos, pueden ser removidos de la estimación del variograma, pero naturalmente eventualmente deben ser modelados e incorporados en las apreciaciones finales del recurso analizado. El Pocket Plot "Gráfico de Bolsillo", es una simple idea

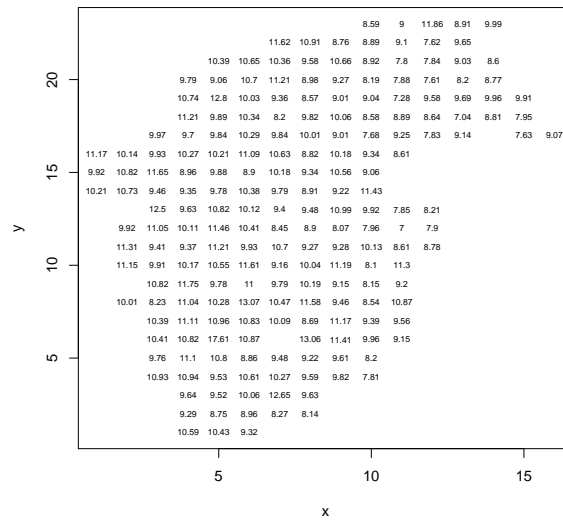


FIGURA 6.1: Ubicación espacial de una muestra de cenizas de carbón (coal-ash), las unidades están en % en ubicaciones reorientadas (Cressie, 1993)

que se ilustrará sobre las diferencias norte-sur de los datos de coal-ash <sup>1</sup> Concentrados sobre la fila  $j$  de la grilla, para alguna otra fila,  $k$  por ejemplo, hay un cierto número  $N_{jk}$ , de diferencias de datos definidas, cuyas localizaciones están a una distancia  $h = |j - k|$  en la dirección norte-sur. Sea  $\bar{Y}_{jk}$  la media de estas *diferencias*  $|^{1/2}$ , promediadas sobre los  $N_{jk}$  términos, y se define:

$$\bar{Y}_h = \frac{1}{|N(h)|} \sum_{N(h)} |Z_i - Z_j|^{1/2} \quad (6.1)$$

$\bar{Y}_h$  es una media ponderada de los  $\bar{Y}_{jk}$ s tales que  $|j - k| = h$ . Luego se define

$$P_{jk} = \bar{Y}_{jk} - \bar{Y}_h \quad (6.2)$$

( $P_{jk} : k = 1, 2, \dots$ ), es la contribución del residual de la fila  $j$ , al estimador del variograma en la diferencia de rezagos. Idealmente, estos puntos serán repartidos a ambos lados del cero, pero si hay algo inusual en la fila  $j$ , entonces

<sup>1</sup>“Este registro de datos del porcentaje de coal ash encontrado en muestras mineras originalmente reportadas por Gomez & Hazen (1970) y posteriormente utilizado Cressie (1993). Los datos se pueden descargar de la librería **gstat** o **sp** del programa R”



Para este caso, consideraremos la base de datos coalash mencionada anteriormente. La función requiere el nombre del data.frame, el tipo de gráfico asociado con la probabilidad o la varianza estandarizada del pocket plot en las direcciones sur-norte o este-oeste; pocketplot de probabilidades por fila, es decir, horizontal “sur-norte” “PPR”, pocketplot de probabilidades por columnas, es decir, vertical “este-oeste” “PPC”, pocketplot de varianza por filas, es decir horizontal “sur-norte” “PVR” y pocketplot de varianzas por columnas, es decir vertical “este-oeste” “PVC”, las coordenadas “X” y “Y”, el nombre de la variable a analizar “Z”, y la identificación de los atípicos (automática “F” o personal “T”). El siguiente código en R, describe la situación de un análisis de estacionariedad local en probabilidades del % ceniza de carbón en dirección sur-norte:

```
library(gstat)
library(geospt)
data(coalash)
pocket.plot(coalash, "PPR", coalash$x, coalash$y, coalash$coalash, F)
```

El gráfico obtenido se muestra en la Figura 6.2a y el gráfico asociado con varianzas estándar se muestra en la Figura 6.2b.

## 6.2.2 Variograma media recortada

Para este variograma, se programo modificando la suma en la formula de Cressie-Hawkins establecida en la expresión (1.6) por la media recortada, así

$$\hat{\gamma}(h) = \frac{\left[ \text{trim.m} \left( \left| \hat{Z}(s_i) - \hat{Z}(s_j) \right|^{\frac{1}{2}} \right) \right]^4}{0.457 + 0.494/N(h)} \quad (6.4)$$

En esta modificación del variograma el usuario puede escoger el porcentaje del recorte. Así, en el caso de un recorte del 50%, el variograma estimado coincidirá con el variograma de la mediana, el cual es mas robusto ante la presencia de atípicos, mientras que si el porcentaje de recorte es del 0%, el estimador para el variograma de la media recortada coincidirá con el estimador robusto de

Cressie-Hawkins. En Bárdossy (2001) se compara el estimador clásico, robusto y media recortada (con un recorte del 10%) y se considera un atípico para evaluar el funcionamiento de los 3 variogramas. Encontramos que el estimador para el variograma de la media recortada produce mejores resultados ante la presencia de atípicos y por lo tanto es más robusto, resultados similares se muestran por medio de simulaciones en Roustant et al. (2007).

La función propuesta *est.variograms()* esta estructurada a partir de la función *est.variogram()* del paquete **sgeostat** en <http://cran.rproject.org/web/packages/sgeostat>. Implementamos el variograma de la media recortada, adicionando en su funcionamiento la instrucción *trim*, correspondiente al porcentaje de recorte del variograma experimental en caja bandeja (bin). En este ejemplo consideramos la base de datos *maas* del paquete **sgeostat**, especificando un porcentaje de recorte del 10% como se explica a continuación en el programa R:

```
library(sgeostat)
data(maas)
maas.point <- point(maas)
maas.pair <- pair(maas.point, num.lags=24, maxdist=2000)
maas.v <- est.variograms(maas.point,maas.pair,'zinc',trim=0.1)
maas.v
```

La salida obtenida incluyendo a los variogramas experimentales ya mencionados el de la mediana es

	lags	bins	classic	robust	med	trimmed.mean	n
1	1	41.66667	101947.2	65465.76	36286.13	57015.22	31
2	2	125.00000	113158.9	61238.92	33444.66	51991.43	184
3	3	208.33333	143501.3	79790.82	53728.38	67770.61	279
4	4	291.66667	177257.6	101478.44	63406.79	86754.46	336
5	5	375.00000	239373.8	144476.65	103685.85	125286.53	367
6	6	458.33333	233764.5	145387.50	115946.06	125355.24	404
7	7	541.66667	273382.4	194285.17	186095.48	177289.00	421
8	8	625.00000	280300.4	197139.93	215218.63	180371.19	441



9	9	708.33333	308830.8	227925.27	273564.52	207709.69	455
10	10	791.66667	297263.4	225228.13	240608.52	210802.15	447
11	11	875.00000	337402.5	250439.56	276672.91	230168.09	461
12	12	958.33333	321287.9	226290.79	246422.02	199083.61	433
13	13	1041.66667	342465.0	252177.03	262795.80	229030.66	417
14	14	1125.00000	371965.3	289594.79	303591.84	271317.58	387
15	15	1208.33333	309236.5	232539.63	234756.15	212280.02	386
16	16	1291.66667	315844.0	239704.08	238300.05	217875.01	360
17	17	1375.00000	347594.5	239448.38	246261.11	210848.17	343
18	18	1458.33333	300932.6	226781.23	226889.51	203460.52	354
19	19	1541.66667	290834.7	210952.98	183415.61	190246.32	330
20	20	1625.00000	260444.7	197217.81	163738.82	174456.98	327
21	21	1708.33333	315371.1	228165.97	206878.84	205701.77	319
22	22	1791.66667	270525.7	198176.63	163732.14	181498.03	323
23	23	1875.00000	255374.6	174233.92	147363.74	155691.27	288
24	24	1958.33333	275440.4	193038.79	168454.22	171184.29	277

### 6.2.3 Resumen estadísticas de la validación cruzada

Para generar el resumen de estadísticas de la validación cruzada, proponemos la función `criterio.cv()`, Esta genera una matriz con los estadísticos presentados en la Sección 1.7, obtenidos mediante LOOCV. Para que esta trabaje correctamente, entramos un `data.frame` con las coordenadas de los datos, las predicciones de la variable analizada, la predicción de la varianza del error estimado del krigado, los valores observados de la variable analizada, los zscore (se obtiene de los residuales divididos por los errores estándar del kriging), y el fold. En el caso de usar la función `rbf.tcv`, la varianza de predicción y el zscore estarán compuestos por NA's. A continuación se muestra un ejemplo para un kriging ordinario y una función de base radial TPS:

```
data(meuse)
coordinates(meuse) <- ~x+y
m <- vgm(.59, "Sph", 874, .04)
# leave-one-out cross validation:
```

```
out <- krige.cv(log(zinc)~1, meuse, m, nmax = 40)
criterio.cv(out)
```

	MPE	ASEPE	RMSPE	MSPE	RMSSPE	R2
1	0.006674145	0.4188814	0.3873933	0.01150903	0.924489	0.7101429

```
data(preci)
attach(preci)
# optimal eta
tab <- rbf.tcv(eta=0.1461, z=prec, coordinates=preci[,2:3],
n.neigh=9, func="TPS")
criterio.cv(tab)
```

	MPE	ASEPE	RMSPE	MSPE	RMSSPE	R2
1	0.8148734	NA	4.02535	NA	NA	0.8071019

#### 6.2.4 Funciones rbf

La función  $rbf()$  es construida a partir de la expresión (4.13) de la Subsección 4.2.1 sin considerar tendencia, es decir,  $F_s = \mathbf{1}$  siendo  $\mathbf{1}$  un vector de tamaño  $n \times 1$ , y requiere para su funcionamiento; el parámetro de suavizamiento  $eta$ , la variable regionalizada  $z$ , las coordenadas de los puntos que fueron usados para generar las predicciones (es decir la muestra)  $coordinates$ , las coordenadas de los puntos a predecir o vector de puntos a predecir  $newdata$ , el número de vecinos más cercanos  $n.neigh$ , este último si se desea un cierto tamaño de vecindario, también es necesario especificar el tipo de RBF dado por  $func$ .

La función  $rbf.cv()$  construye un valor de RMSPE con la expresión dada en el ítem ii. de la Sección 1.7, y requiere;  $eta$ ,  $z$ ,  $coordinates$ ,  $n.neigh$ , y  $func$ . La función  $RBf.phi()$  requiere; una distancia entre el par de puntos en las localizaciones  $S_i$  y  $S_j$ ,  $eta$  y  $func$ . La función  $rbf.tcv()$  requiere  $eta$ ,  $z$ ,  $n.neigh$ , y  $func$ , esta última es necesaria para el mapeo de la predicción de la variable analizada, y para la construcción de las estadísticas de resumen de validación cruzada.

Para la función `graph.rbf()`, presentamos un ejemplo con la base de datos `preci` incluida en el paquete `geospt`. Los datos corresponden a una muestra empírica de precipitación. En la función se debe especificar;  $z$ , `coordinates`, `newdata`, `n.neigh`, `func`, `np` número de puntos donde se calcula la función de base radial, `n.eta` es factor de longitud del eje X, número de veces del parámetro de suavizado `eta`, y `P.T` es un operador lógico (T=True o F=False) para imprimir la tabla con los valores que generan el gráfico. A continuación se muestra la gráfica obtenida con cuatro funciones de base radial implementadas.

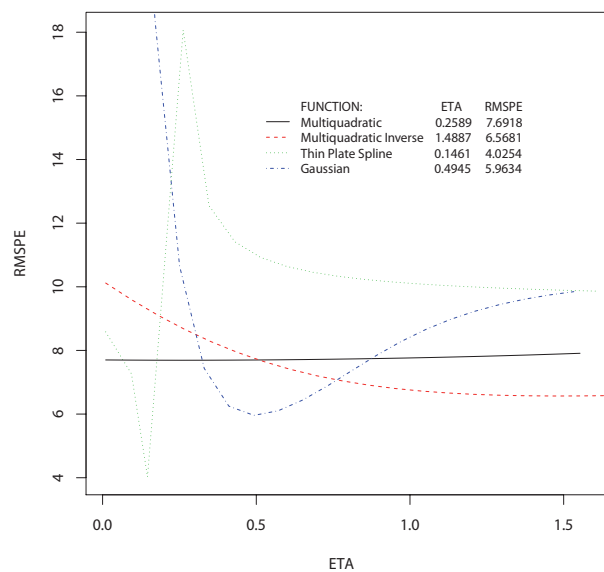


FIGURA 6.3: Optimización de `eta`, en funciones de base radial

Como se muestra en la Figura 6.3, el menor RMSPE se genera con la spline capa delgada.

La Tabla 6.1 provee una breve descripción de las funciones implementadas en el paquete `geospt`.

## 6.2.5 Mapa de predicciones

Con la muestra utilizada en el capítulo 3, realizamos un mapa de predicción de la temperatura media diaria terrestre para el 1 de diciembre de 2008, la dis-

Función	Descripción
<i>criterio.cv()</i>	Extrae un resumen de valores estadísticos obtenidos de LOOCV: <i>MPE</i> , <i>ASEPE</i> , <i>RMSPE</i> , <i>MSPE</i> , <i>RMSSPE</i> , $R^2$ .
<i>est.variograms()</i>	Calcula el variograma experimental: clásico, robusto, mediana, y media recortada (Cressie 1993, Bárdossy 2001)
<i>pocketplot()</i>	Gráfica el pocketplot de probabilidad o varianza estandarizada en las direcciones sur-norte y este-oeste, (ver (Cressie 1993))
<i>rbf()</i>	Extrae las predicciones a partir de las funciones: gaussiana ( <i>GAU</i> ), exponencial ( <i>EXPON</i> ), trigonométrica ( <i>TRI</i> ) spline capa delgada ( <i>TPS</i> ), spline completamente regularizada ( <i>CRS</i> ), spline con tensión ( <i>ST</i> ), inversa multicuadrática ( <i>IM</i> ), y multicuadrática ( <i>MQ</i> )
<i>rbf.cv()</i>	Genera un valor de <i>RMSPE</i> , resultado de LOOCV
<i>RBF.phi()</i>	Genera un valor numérico, obtenido de la función de base radial generado a partir de; una distancia, un parámetro eta ( $\eta$ ) de suavizamiento, y una función ‘‘GAU’’, ‘‘EXPON’’, ‘‘TRI’’, ‘‘TPS’’, ‘‘CRS’’, ‘‘ST’’, ‘‘IM’’ or ‘‘M’’
<i>rbf.tcv()</i>	Genera una tabla con las coordenadas de los datos, la predicción, los valores observados, los residuos, la varianza de predicción, y el zscore (residual dividido por el error estándar) de la variable analizada, los cuales están asociados a la LOOCV
<i>graph.rbf()</i>	Genera un gráfico que describe el comportamiento del parámetro eta optimizado, asociado con la función de base radial

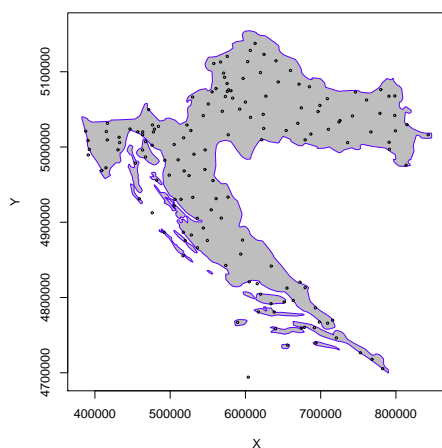
TABLA 6.1: Algunas funciones del paquete **geosp**

tribución de los datos se muestra en la Figura 6.4a. Inicialmente, realizamos la optimización del parámetro  $\eta$  a partir de los datos previamente mencionados. Para esto, trabajaremos con la función *graph.rbf()*. Esta función se utiliza para las funciones de base radial mencionadas previamente en la Tabla 6.1, encontrando que la función multicuadrática es la que mejor ajusta estos datos, así con el parámetro de suavizamiento  $\eta = 0.0001$ , tenemos un  $RMSPE = 1.873$

a partir de la expresión 3.16, esto se hace con la función *rbf.cv()*. La función *graph.rbf()* trabaja con las funciones *rbf.cv()* (función asociada al *RMSPE* de la función base radial) y *optimize()* (función optimizadora), (no se utilizó la función *optim()*, descrita en Mittelhammer et al. (2000) ya que no demandaba más tiempo que *optimize()*). Por último la función *optimize()* del programa R descrita en Brent (1973), busca el valor óptimo del parámetro  $\eta$  asociado con el *RMSPE* generado por la función *rbf.cv()*, en un intervalo para dicho parámetro establecido por el usuario, por ejemplo para multicuadráticas los óptimos para  $\eta$  suelen encontrarse cerca a 0, por lo cual un buen intervalo sería entre 0 y 1. Los datos y el shapefile son precargados. Seguidamente se genera una grilla de 70000 puntos dentro de la región analizada, con el fin de generar predicciones de la temperatura media terrestre. Esta grilla es obtenida usando la función *spsample()* del paquete **sp**. Las predicciones son generadas con la función *rbf()*, esta función requiere el valor del parámetro "eta", la variable a analizar "z", las coordenadas de los puntos muestreados *coordinates*, las coordenadas de los nuevos puntos *newdata*, el número de vecinos *n.neigh*, y el tipo de función de base radial *func*. Estas predicciones son luego convertidas a un objeto de clases *SpatialPixelsDataFrame* y *sp*, con la instrucción *coordinates()*, del paquete **sp**, y finalmente con la función *spplot()* se obtiene el mapa de las predicciones de la variable analizada, el cual se muestra en la Figura 6.4b.

```
library(gstat)
library(geoR)
# Consideramos los datos "dif.IDSTA3" de la aplicación del Capítulo 3
Datos <- read.table("../dif.IDSTA3.txt", header=T, sep=" ", dec = ",")
readShapePoly(file("../croatia.shp"))
rbf.cv(eta=0.0001, z=Datos$TEMP, coordinates=Datos[,3:4], n.neigh=15,
      func="M")
      1.873115
# prediction case a grid of points
pts <- spsample(croatia.shp, n=70000, type="regular")
pred.rbf <- rbf(eta=0.0001, z=Datos$TEMP, coordinates=Datos[,3:4],
              newdata=pts@coords, n.neigh=15, func="M")
coordinates(pred.rbf) = c("x", "y")
```

```
gridded(pred.rbf) <- TRUE
# muestra el mapa de predicción con la rbf multicuadrática
spplot(pred.rbf["var1.pred"], cuts=40, scales = list(draw =T),
        col.regions=bpy.colors(100), key.space=list(space="right", cex=0.8))
```



(a) Ubicación de estaciones meteorológicas

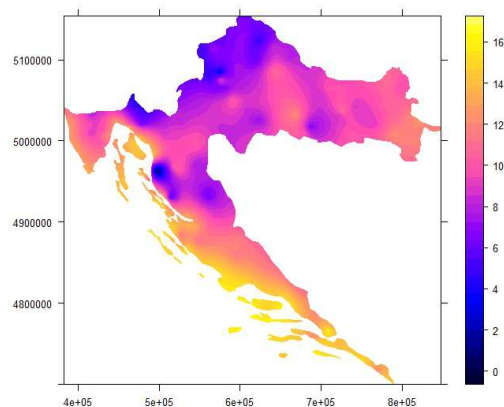
(b) Predicción de la temperatura media  
terrestre, el 1 de diciembre de 2008

FIGURA 6.4: Mapa de Croacia

Algunas funciones que se presentan en el Apéndice A, aun no han sido incorporadas en la librería y próximamente esperamos dejarlas disponibles para los usuarios del programa R.



# Capítulo 7

## Conclusiones y futuras líneas de investigación

### 7.1 Conclusiones

Se propusieron innovaciones en la predicción espacio y espacio-temporal, a partir de métodos geoestadísticos kriging y de funciones de base radial, considerando métodos basados en distancias. En este sentido, por medio de las distancias entre las variables explicativas, incorporadas específicamente en la regresión basada en distancias, se propusieron modificaciones en: el método kriging universal y en la interpolación con splines espacial y espacio-temporal usando las funciones de base radial.

Como consecuencia de lo anterior, las principales aportaciones del trabajo son: tres métodos propuestos para interpolación espacial y espacio-temporal explicados en los Capítulos 3, 4 y 5, una librería llamada **geospt** en el programa R para: interpolación espacial con funciones de base radial sin tendencia y análisis geoestadístico, descrita con detalle en el Capítulo 6. Además, una serie de funciones diseñadas para realizar los métodos propuestos, que permiten llevar a cabo interpolación espacial y espacio-temporal con tendencia basada en distancias, optimización y validación cruzada “leave-one-out”, las cuales se encuentran en el Apéndice A.



Los estudios de simulación permitieron validar el funcionamiento de los métodos propuestos mostrando ventajas y desventajas bajo los escenarios presentados en los Capítulos 3, 4 y 5. Es así como en el Capítulo 3, el estudio de simulación permitió comparar la capacidad predictiva del método tradicional kriging universal con respecto a kriging universal basado en distancias; mientras que en los Capítulos 4 y 5, los estudios de simulación permitieron comparar el funcionamiento de las funciones de base radial espaciales y espacio-temporales, considerando en la tendencia las coordenadas principales generadas a partir de las variables explicativas mixtas mediante el uso del método basado en distancias.

El método propuesto DBUK muestra, tanto en las simulaciones como en las aplicaciones, ventajas en la reducción del error con respecto al método clásico de krigeado universal. Esta reducción de los errores se asocia a una mejor modelización de la tendencia y a un menor error en el ajuste y modelado del variograma, al considerar las coordenadas principales obtenidas a partir de las variables explicativas mixtas. Entre muchas otras posibles causas, el error es generado por omisión de variables y por considerar formas funcionales incorrectas.

Los resultados de esta investigación muestran que el nuevo método DBUK es robusto a este tipo de situaciones en comparación con el método clásico de krigeado universal. El estudio de simulación muestra que el método propuesto DBUK es mejor que el método de krigeado universal tradicional ya que se encontró una notoria reducción del error, asociada a un RMSPE mas pequeño, esta reducción en general fue superior al 10%.

El método DBUK podría producir una mejor estimación de la variable regionalizada si el número de coordenadas principales se incrementa. Esto es posible, incluyendo las coordenadas principales más significativas tanto en modelo de tendencia como en el variograma; la segunda aplicación presentada en el Capítulo 3 ilustra este hecho. Sin embargo, un estudio más detallado de simulación debería realizarse en esta dirección.

Por otro lado, los métodos propuestos en esta investigación con funciones de base radial consideraron estacionariedad e isotropía. En los resultados pre-

sentados en los Capítulos 4 y 5, es decir, los métodos propuestos DBSIRBF en espacio y DBSTIRBF en espacio-tiempo analizados mediante una estructura de krigeado considerando en la tendencia las coordenadas principales, presentan un buen funcionamiento al trabajar con vecindarios grandes, indicando en general que se tendrá un menor error asociado a un RMSPE más pequeño.

En diversos estudios, la detección de variabilidad entre zonas es una tarea muy difícil, y por lo cual los métodos propuestos DBUK, DBSIRBF y DBSTIRBF son útiles de acuerdo a los resultados obtenidos en la presente investigación, ya que aprovechan al máximo la información existente asociada a las variables explicativas. Aunque la correlación de las variables explicativas puede ser baja con respecto a la variable respuesta, el punto clave en los métodos propuestos es la correlación entre las coordenadas principales (construida con las variables explicativas) y la variable respuesta.

Los métodos aquí desarrollados se aplicaron a datos agronómicos y climatológicos. Los resultados de validación cruzada “leave-one-out” mostraron un buen rendimiento de los predictores propuestos, lo cual indica que se pueden utilizar como métodos alternos y validos a los tradicionales para el modelado de variables correlacionadas espacialmente y espacio-temporalmente, considerando siempre covariables en la remoción de la tendencia.

## 7.2 Futuras líneas de investigación

Aunque los interpoladores locales como los propuestos en esta investigación, trabajan relativamente rápido al considerar vecindarios pequeños, la complejidad computacional sigue siendo un problema. Los métodos propuestos en esta tesis demandan mucho tiempo y pueden gastar varias horas para generar predicciones sobretodo en casos espacio-temporales, en donde es necesario considerar vecindarios de mayor tamaño. Esto indica que todavía hay oportunidad para mejorar el procesamiento de datos en los métodos propuestos.

Desde el punto de vista teórico, parece interesante hacer un estudio de las funciones de base radial espacio-temporales en modelos espacio-temporales; producto, suma y suma producto, comprobar su bondad de ajuste con simu-

laciones y datos reales, para valorar sus ventajas o desventajas con respecto a los métodos kriging espaciales y espacio-temporales.

# Referencias

- Abramowitz, M. & Stegun, I. A. (1965), *Handbook of Mathematical Functions*, Dover, New York.
- Akaike, H. (1973), *Information theory and the maximum likelihood principle*, In International Symposium on Information Theory, Akademiai Kiado, Budapest.
- Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**, 716–723.
- Arenas, C. & Cuadras, C. (2002), 'Recent statistical methods based on distances', *Contributions to Science, Institut d'Estudis Catalans Barcelona* **2**(2), 183–191.
- Armstrong, M., C. G. & Hubert, P. (1993), 'Kriging the rainfall in lesotho. In geostatistics tróia '92', *Kluwer Academic Publishers, Dordrecht* **2**, 661–672.
- Armstrong, M. & Diamond, P. (1984), 'Testing variograms for positive-definiteness', *Mathematical Geology* **16**, 407–421.
- Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data.*, Chapman & Hall/CRC, Boca Raton-Florida.
- Bárdossy, A. (2001), *Introduction to Geostatistics*, Institut für Wasserbau der Universität Stuttgart.
- Bardossy, A., Haberlandt, U. & Grimm-Strele, J. (1997), Interpolation of groundwater quality parameters using additional information, Technical

report, In GeoENV I (Geostatistics for Environmental Applications), 189-200, Kluwer Academic Publ., Dordrecht.

Baxter, B. (1992), The interpolation theory of radial basis functions, PhD thesis, Department of Applied Mathematics and Theoretical Physics, University of Cambridge.

Bellier, E., Planque, B. & Petitgas, P. (2007), 'Historical fluctuations in spawning location of anchovy (*engraulis encrasicolus*) and sardine (*sardina pilchardus*) in the bay of biscay during 1967-73 and 2000-2004', *Fisheries Oceanography* **16**(1), 1-15.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford Press, Oxford.

Bivand, R., Pebesma, E. & Rubio, V. (2008), *Applied Spatial Data Analysis with R*, Springer, New York.

Braud, I. (1990), Etude methodologique del analyze en composantes principales de processus bidimensionels. Effets des approximations numeriques et de l'Echantillonnage et utilisation pour la simulation de champs aleatoires. Application au traitement des temperatures mensuelles de surface de la mer sur l'Atlantique Intertropi, PhD thesis, l'Institut National Polytechnique de Grenoble, France.

Breiman, L. (1995), 'Better subset regression using the nonnegative garrote', *Technometrics* **37**, 373-384.

Brent, R. (1973), *Algorithms for minimization without derivatives*, Prentice-Hall, Englewood Cliffs.

Capeche, C. L., Macedo, J. R., Manzatto, H. R. H. & Silva, E. F. (1997), Caracterização pedológica da fazenda angra - pesagro/rio, Technical report, Technical Report: Estação experimental de Campos (RJ). In Informação, globalização, uso do solo, Rio de Janeiro.

- Chen, C. (2007), Model selection for curve and surface fitting using generalized degrees of freedom, PhD thesis, Institute of Statistical, National Central University, Jhongli-Taiwan.
- Chen, L., Fuentes, M. & Davis, J. M. (2006), ‘Spatial temporal statistical modelling and prediction of environmental processes’, *Hierarchical Modelling for the Environmental Sciences*, Oxford Univ. Press pp. 121–144.
- Chica, J., Cano, R. & Chica, M. (2007), ‘Modelo hedónico espacio-temporal y análisis variográfico del precio de la vivienda’, *GeoFocus* **7**, 56–72.
- Chilès, J. P. & Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, New York.
- Christakos, G. (2000), *Modern Spatiotemporal Geostatistics*, Oxford University Press, Oxford.
- Cleveland, W. (1979), ‘Robust locally weighted regression and smoothing scatterplots’, *Journal of the American Statistical Association* **74**, 829–836.
- Craven, P. & Wahba, G. (1979), ‘Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation’, *Numerische Mathematik* **31**, 377–403.
- Cressie, N. (1985), ‘Fitting variogram models by weighted least squares’, *Journal of the International Association for Mathematical Geology* **17**, 563–586.
- Cressie, N. (1989), ‘Geostatistics.’, *The American Statistician* **43**, 197–202.
- Cressie, N. (1993), *Statistics for Spatial Data*, Revised Edition. John Wiley & Sons Inc., New York.
- Cressie, N. & Hawkins, D. M. (1980), ‘Robust estimation of the variogram’, *Mathematical Geology* **12**, 115–125.
- Cressie, N. & Huang, H. C. (1999), ‘Classes of nonseparable, spatio-temporal stationary covariance functions’, *Journal of the American Statistical Association* **94**, 1330–1340.

- Cressie, N. & Majure, J. (1995), Non-point source pollution of surface waters over a watershed, in 'Programme Abstracts of the third SPRUCE International Conference. Merida, Mexico'.
- Cressie, N. & Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, John Wiley and Sons.
- Şen, Z., Altunkaynak, A. & Özger, M. (2006), 'Space-time interpolation by combining air pollution and meteorologic variables', *Pure and Applied Geophysics* **163**, 1435–1451.
- Cuadras, C. (1989), 'Distance analysis in discrimination and classification using both continuous and categorical variables', *Recent Developments in Statistical Data Analysis and Inference*. pp. 459–474.
- Cuadras, C. (1993), 'Interpreting an inequality in multiple regression', *The American Statistician* **47(4)**, 256–258.
- Cuadras, C. (2007), *Métodos Multivariados Basados en Distancias*, Curso de doctorado, Universidad de Barcelona, Barcelona.
- Cuadras, C. & Arenas, C. (1990), 'A distance based regression model for prediction with mixed data', *Communications in Statistics A - Theory and Methods* **19**, 2261–2279.
- Cuadras, C., Arenas, C. & Fortiana, J. (1996), 'Some computational aspects of a distance-based model for prediction', *Communications in Statistics - Simulation and Computation* **25(3)**, 593–609.
- Cuadras, C. M. (2009), 'Distance-based multivariate regression', *Frontieres of Interfaces Between Statistics and Sciences, Hyderabad* pp. 65–70.
- Cuadras, C. M. & Fortiana, J. (1993), 'Aplicaciones de las distancias en estadística', *Questió* **17(1)**, 39–74.
- De Cesare, L., Myers, D. E. & Posa, D. (1997), 'Spatial temporal modeling of SO<sub>2</sub> in the milan district', *Geostatistics Wollongong'96, Kluwer Academic Publishers* **2**, 1031–1042.

- 
- De Cesare, L., Myers, D. E. & Posa, D. (2001*a*), ‘Estimating and modeling space-time correlation structures’, *Statistics and Probability Letters* **51**, 9–14.
- De Cesare, L., Myers, D. E. & Posa, D. (2001*b*), ‘Estimating and modeling space-time correlation structures’, *Statistics and Probability Letters* **51**, 9–14.
- Diggle, P., Harper, L. & Simon, S. (1995), Geostatistical analysis of residual contamination from nuclear weapons testing, *in* ‘Abstracts of the third SPRUCE international conference, Mérida, México’.
- Dimitrakopoulos, R. & Luo, X. (1994), Spatiotemporal modeling: covariances and ordinary kriging systems, Technical report, Geostatistics for the next century, Kluwer Academic Publishers.
- Donoho, D. & Johnstone, I. (1994), ‘Ideal spatial adaptation by wavelet shrinkage’, *Biometrika* **81**, 425–456.
- Duchon, J. (1976), ‘Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces’, *Rairo Analyse Numerique* **10**, 5–12.
- Efron, B., Johnstone, I., Hastie, T. & Tibshirani, R. (2004), ‘Least angle regression’, *Annals of Statistics* **32**, 407–499.
- Emery, X. & Cornejo, J. (2010), ‘Truncated gaussian simulation of discrete-valued, ordinal coregionalized variables’, *Computers & Geosciences* **36**, 1325–1338.
- Emery, X. & Silva, D. (2009), ‘Conditional co-simulation of continuous and categorical variables for geostatistical applications’, *Computers & Geosciences* **35**, 1234 – 1246.
- Esteve, A., Boj, E. & Fortiana, J. (2009), ‘Interaction terms in distance-based regression’, *Communications in Statistics A. Theory and Methods* **38**, 3498–3509.



- Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.
- Faraway, J. (2005), *Linear Models with R*, Chapman and Hall, London.
- Finkenstädt, B., Held, L. & Isham, V. (2006), *Statistical Methods for Spatio-Temporal Systems*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability.
- Franke, R. (1982), ‘Smooth interpolation of scattered data by local thin plate splines’, *Computer and Mathematics with Applications* **8**, 273–281.
- Fuentes, M. (2001), ‘A high frequency kriging approach for nonstationary environmental processes’, *Environmetrics* **12**, 469–483.
- Fuentes, M. (2002a), ‘Modeling and prediction of nonstationary spatial processes’, *Statistical Modelling: An International Journal* **2**, 281–298.
- Fuentes, M. (2002b), ‘Spectral methods for nonstationary spatial processes’, *Biometrika* **89**, 197–210.
- Fuentes, M. & Smith, R. (2001), A new class of nonstationary models, Technical report, Technical report no. 2534, North Carolina State University, North Carolina State.
- Gaetan, C. & Guyon, X. (2010), *Spatial Statistics and Modeling*, Series in Statistics, Springer.
- García-Soidán, P., Menezes, R. & Rubiños-López, O. (2012), ‘An approach for valid covariance estimation via the fourier series’, *Enviromental Earth Sciences* **66**(2), 615–624.
- George, E. I. & McCulloch, R. (1993), ‘Variable selection via gibbs sampling’, *Journal American Statistician Asociation* **88**, 881–889.
- Gneiting, T. (2002), ‘Nonseparable, stationary covariance functions for space-time data’, *Journal of the American Statistical Association* **97**, 590–600.

- Gneiting, T. (2003), Should weather and climate models be deterministic?, Technical report, En Joint Statistical Meeting, San Francisco, California.
- Gneiting, T., Genton, M. G. & Guttorp, P. (2005), Geoestatistical space-time models, stationarity, separability, and full simmetry, Technical report, Technical Report no. 475, University of Washington, Washington.
- Gneiting, T., Genton, M. G. & Guttorp, P. (2007), *Statistical Methods for Spatio-Temporal Systems*, Chapman & Hall/CRC, Boca Raton-Florida, chapter Geostatistical space-time models, stationarity, separability and full symmetry, pp. 151–175.
- Gomez, M. & Hazen, K. (1970), Evaluating sulfur and ash distribution in coal seems by statistical response surface regression analysis, Technical report, U.S. Bureau of Mines Report RI 7377.
- Gower, J. (1968), ‘Adding a point to vector diagrams in multivariate analysis’, *Biometrika* **55**, 582–585.
- Gower, J. (1971), ‘A general coefficient of similarity and some of its properties’, *Biometrics* **27**, 857–871.
- Haas, T. C. (1995), ‘Local prediction of a spatio-temporal process with an application to wet sulfate deposition’, *Journal of the American Statistical Association* **90**, 1189–1199.
- Haining, R. (2004), *Spatial Data Analysis: Theory and Practice*, Cambridge University Press, Cambridge.
- Hall, P. & Patil, P. (1994), ‘Properties of nonparametric estimators of autocovariance for stationary random fields’, *Probability Theory Related Fields* **99**(3), 399–424.
- Hardy, R. & Gopfert, W. (1975), ‘Least squares prediction of gravity anomalies, geodial undulations, and detections of the vertical with multiquadric harmonic functions’, *Geophysical Research Letters* **2**, 423–426.
- Hardy, R. L. (1971), ‘Multiquadric equations of topography and other irregular surfaces’, *Journal of Geophysical Research* **76**, 1905–1915.

- Hardy, R. L. (1990), ‘Theory and applications of the multiquadric-biharmonic method. 20 years of discovery 1968-1988’, *Computers & Mathematics with Applications* **19**, 163–208.
- Hengl, T. (2009), *A Practical Guide to Geostatistical Mapping*, 2nd edn, University of Amsterdam, Amsterdam.
- Hengl, T., Heuvelink Gerard, B. M., Perčec Tadić, M. & Pebesma, E. J. (2012), ‘Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images’, *Theoretical and Applied Climatology* **107**, 1-2, 265–277.
- Heryudono, A. & Driscoll, T. (2010), ‘Radial basis function interpolation on irregular domain through conformal transplantation’, *Journal of Scientific Computing* **44**(3), 286–300.
- Hiebl, J., Auer, I., Böhm, R., Schöner, W., Maugeri, M., Lentini, G., Spinoni, J., Brunetti, M., Nanni, T. Perčec Tadić, M., Bihari, Z., Dolinar, M. & Müller-Westermeier, G. (2009), ‘A high-resolution 1961-1990 monthly temperature climatology for the greater alpine region’, *Meteorologische Zeitschrift* **18**, 507–530.
- Higdon, D., Swall, J. & Kern, J. (1999), Non-stationary spatial modeling, Technical report, In *Bayesian Statistics*, 761-768, Oxford University Press, Oxford.
- Holland, D., Saltzman, N., Cox, L. H. & Nychka, D. (1999), Spatial prediction of sulfur dioxide in the eastern United States, Technical report, In *GeoENV II - Geostatistics for Environmental Applications*, 65-76, Kluwer Academic Publ., Dordrecht.
- Huang, H. C. & Cressie, N. (1996), ‘Spatio-temporal prediction of snow water equivalent using the Kalman filter’, *Computational Statistics & Data Analysis* **22**, 159–175.
- Isaaks, E. & Srisvastava, R. (1989), *An Introduction to Applied Geostatistics*, Oxford Univ. Press, New York.

- Jin, R., Chen, W. & Simpson, T. (2001), ‘Comparative studies of metamodelling techniques under multiple modeling criteria’, *Journal of Structural & Multidisciplinary Optimization* **23**, 1–13.
- Johnston, K., Ver, J., Krivoruchko, K. & Lucas, N. (2001), *Using ArcGIS Geostatistical Analysis*, ESRI.
- Joseph, V. R., Hung, Y. & Sudjianto, A. (2008), ‘Blind kriging: A new method for developing metamodels.’, *Journal of Mechanical Design* **3**, 31–102.
- Journel, A. G. & Huijbregts, C. J. (1978), *Mining Geoestistics*, Academic Press, New York.
- Kolovos, A., Christakos, G., Hristopulos, D. T. & Serre, M. L. (2004), ‘Methods for generating non-separable spatiotemporal covariance models with potential environmental applications’, *Advances in Water Resources* **27**, 815–830.
- Kondoh, H., Koizumi, T. & Ikeda, K. (2011), ‘A geostatistical approach to spatial density distributions of sika deer (*Cervus nippon*)’, *Journal of Forest Research* pp. 1–8.
- Krige, D. G. (1951), ‘A statistical approach to some basic mine valuation problems on the witwatersrand’, *Journal of the Chemical, Metallurgical and Mining Society of South Africa* **52**, 119–139.
- Kyriakidis, P. C. & Journel, A. G. (1999), ‘Geostatistical space-time models: a review’, *Mathematical Geology* **31**, 651–684.
- Le, N. & Zidek, J. (2006), *Statistical Analysis of Environmental Space-Time Processes*, Springer.
- Lloyd, C. D. (2010), *Local Models for Spatial Analysis*, second edn, Taylor & Francis Group, Boca Raton-Florida.
- Mardia, K. V., Kent, J. T. & Bibby, J. M. (2002), *Multivariate Analysis*, Academic Press, Inc, London.

- Martínez, F. (2008), Modelización de la función de covarianza en procesos espacio-temporales: análisis y aplicaciones., PhD thesis, Universidad de Valencia-España.
- Mateu, J., Montes, F. & Fuentes, M. (2003), ‘Recent advances in space-time statistics with applications to atmospheric data: An overview’, *Journal of Geophysical Research* **108**, (D24).
- Matheron, G. (1962), *Traité de Géostatistique Appliquée*, Editions Technip, Paris.
- Matheron, G. (1971), *The theory of regionalized variables and its applications*, Cahiers du Centre de Morphologie Mathématique, 5, Fontainebleau, Paris.
- Mitáš, L. & Mitášová, H. (1988), ‘General variational approach to the interpolation problem’, *Computers and Mathematics with Applications* **16**, 983–992.
- Mitášová, H. & Hofierka, J. (1993), ‘Interpolation by regularized spline with tension: II. Application to terrain modeling and surface geometry analysis’, *Mathematical Geology* **25**, 657–669.
- Mitášová, H. & Mitáš, L. (1993), ‘Interpolation by regularized spline with tension: I. Theory and implementation’, *Mathematical Geology* **25**, 641–655.
- Mittelhammer, R., Judge, G. & Miller, D. (2000), *Econometric Foundations*, New York.
- Myers, D. (1992), ‘Kriging, cokriging, radial basic functions and the role of positive definiteness’, *Computers Mathematical Application* **24**, 139–148.
- Myers, D., De Iaco, S., Posa, D. & De Cesare, L. (2002), ‘Space-time radial basis functions’, *Computers and Mathematics with Applications* **43**, 539–549.
- Myers, D. & Journel, A. (1990), ‘Variograms with zonal anisotropies and non-invertible kriging systems’, *Mathematical Geology* **22**, 779–785.

- Nychka, D. & Saltzman, N. (1998), Design of air quality networks, Technical report, In Case Studies in Environmental Statistics, 51-76, Springer Verlag, New York.
- Ortega-Pérez, M. (2009), Método de registro no rígido basado en funciones de base radial. Aplicación a neurocirugía utilizando atlas cerebrales, PhD thesis, Universitat Politècnica de València. Departamento de Sistemas Informáticos y Computación.
- Paez, M., G. D. & De Oliveira, V. (2005), 'Interpolacion performance of a spatio temporal model with spatially varying coefficients: Application to pm10 concentration in rio de janeiro', *Environmental and Ecological Statistics* **12**, 169–193.
- Perčec Tadić, M. (2010), 'Gridded croatian climatology for 1961-1990', *Theoretical and Applied Climatology* pp. 1434–4483.
- Planque, B., Bellier, E. & Lazure, P. (2007), 'Modelling potential spawning habitat of sardine (*sardina pilchardus*) and anchovy (*engraulis encrasicolus*) in the bay of biscay', *Fisheries Oceanography* **16**(1), 16–30.
- R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL <http://www.R-project.org/>.
- Robertson, G. (1987), 'Geostatistics in ecology. interpolating with known variance', *Ecology* **68**(3), 744–748.
- Rodriguez-Iturbe, I. & Mejia, J. (1974), 'The design of rainfall networks in time and space', *Water Resources Research* **10**, 713–728.
- Rouhani, S. & Hall, T. J. (1989), Space-time kriging of groundwater data, Technical report, En Geostatistics, 2, 639-650, Kluwer Academic Publ., Dordrecht.
- Rouhani, S. & Myers, D. E. (1990), 'Problems in space-time kriging of hydrogeological data', *Mathematical Geology* **22**, 611–623.

- Roustant, O., Dupuy, D. & Helbert, C. (2007), Robust estimation of the variogram in computer experiments, Technical report, Ecole des Mines, Département 3MI, 158 Cours Fauriel, 42023 Saint-Etienne, France.
- Sacks, J., Welch, W., Mitchell, T. J. & Wynn, H. P. (1989), ‘Design and analysis of computer experiments’, *Statistical Science* **4**, 409–423.
- Sahu, S. K. & Mardia, K. V. (2005), Recent trends in modeling spatio-temporal data, *in* ‘In Proceedings of the special meeting on Statistics and Environment, Messina, Italy’.
- Samper, F. & Carrera, J. (1993), *Geoestadística. Aplicaciones a la Hidrogeología Subterránea*, Centro Internacional de Métodos Numéricos en Ingeniería.
- Sampson, P. D. & Guttorp, P. (1992), ‘Nonparametric estimation of non-stationary spatial covariance structure’, *Journal of American Statistical Association* **87**, 108–119.
- Santner, T., Williams, B. & Notz, W. (2003), *The Design and Analysis of Computer Experiments*, Springer-Verlag, New York.
- Schagen, I. P. (1979), ‘Interpolation in two dimensions: a new technique’, *Journal of the Institute of Mathematics and its Applications* **23**, 53–59.
- Schwarz, G. (1978), ‘Estimating the dimension of a model’, *Annals of Statistics* **6**, 461–464.
- Snepvangers, J.J.J.C., H. G. & Huisman, J. (2003), ‘Soil water content interpolation using spatio-temporal kriging with external drift’, *Geoderma* **112**, 253–271.
- Späh, H. (1969), ‘Exponential spline interpolation’, *Computing* **4**, 225–233.
- Thiébaux, H. & Pedder, M. (1987), *Spatial Objective Analysis: With Applications in Atmospheric Science.*, Academic Press. London.
- Tibshirani, R. (1996), ‘Regression shrinkage and selection via the lasso’, *Royal Statistics Society* **58**, 267–288.

- 
- van de Kasstele, J., Stein, A., Dekkers, A. & Velders, G. (2009), ‘External drift kriging of NO<sub>x</sub> concentrations with dispersion model output in a reduced air quality monitoring network’, *Environmental and Ecological Statistics* **16**, 321–339.
- Velleman, P. F. & Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis.*, Duxbury, Boston, MA.
- Wackernagel, H. (2003), *Multivariate Geostatistics: An Introduction with applications*, Third Completely Revised Edition. Springer-Verlag, New York.
- Wand, M. (2000), ‘A comparison of regression spline smoothing procedures’, *Computational Statistics* **15**, 443–462.
- Wand, M. & Jones, M. (1995), *Kernel Smoothing*, Chapman and Hall, New York.
- Yavuz, H. & Erdoğan, S. (2012), ‘Spatial analysis of monthly and annual precipitation trends in Turkey’, *Water Resources Management* **26**, 609–621.
- Ye, J. (2008), Geostatistical methods for spatio-temporal analysis of fMRI data, PhD thesis, Department of Statistics, University of Georgia.
- Ye, J., Lazar, N. & Li, Y. (2011), ‘Sparse geostatistical analysis in clustering fMRI time series’, *Journal of Neuroscience Methods* **199(2)**, 336–345.
- Zaninovic, K., Gajic-Capka, M. & Percec-Tadic, M. (2008), Klimatski atlas Hrvatske, climate atlas of Croatia 1961, 1990, 1971 2000., Technical report, Meteorological and Hydrological Service Republic of Croatia, Zagreb.
- Zhang, G. (2011), ‘Smoothing splines using compactly supported, positive definite, radial basis functions’, *Computational Statistics* pp. 1–12.





**Anexos A**

**Programación en R**

En este apéndice, se presentan los códigos que se han utilizado en la elaboración de esta investigación. La idea es mostrar los procedimientos prácticos desarrollados en el programa R, así como las funciones implementadas para su correcto funcionamiento en lo práctico y en las simulaciones realizadas en este trabajo.

## A.1 Funciones implementadas y utilizadas en el Capítulo 3

Se implementaron las funciones *est.variograms*, ya explicada en el Capítulo 6, *krige.u.db* y *x.new* la cual esta en la programación de los Capítulos 3, 4 y 5. A continuación se describe la función *krige.u.db*:

*krige.u.db*: Calcula la predicción de la variable regionalizada  $Z$  y su correspondiente varianza del error para un punto, para ello requiere la variable regionalizada  $z$ , la tendencia asociada con las coordenadas principales de la muestra *trend*, los valores propios asociados a las coordenadas principales mas relevantes *ValoresPropios*, es decir asociados a *trend*, las coordenadas de la muestra *coordinates*, las coordenadas espaciales del nuevo individuo *newdata*, el número de vecinos más cercanos *n.neigh*, el modelo de covarianza *modelo.cov*, y los parámetros asociados con el modelo de covarianza *cov.pars*

```
assign("krige.u.db",
function(z, trend, ValoresPropios, coordinates, newdata, n.neigh,
        modelo.cov, cov.pars){
  So <- newdata
  So <- as.data.frame(cbind(x=So[1],y=So[2]))
  coordinates(So) <- c("x", "y")
  s <- coordinates
  dist.So <- spDists(So,s)
  remove("newdata")
  vec.orden <- order(dist.So)
  dist.vec.cerca <- dist.So[vec.orden[1:n.neigh]]
  m.dist.vec <- spDists(s)[vec.orden[1:n.neigh], vec.orden[1:n.neigh]]
  b <- diag(tcrossprod(trend))
```

```

x.new <- (1/2) * (diag(length(ValoresPropios[1:ncol(trend)])) *
  (ValoresPropios[1:ncol(trend)]^(-1))) %*% crossprod(trend,b -
  as.numeric(dist.So)^2)
one = rep(1,length(dist.vec.cerca))
tend <- as.matrix(trend[vec.orden[1:n.neigh],])
cov.0 <- cov.spatial(0, cov.model= modelo.cov, cov.pars =cov.pars)
m.cov<-cov.spatial(m.dist.vec, cov.model= modelo.cov, cov.pars=cov.pars)
v <- cov.spatial(dist.vec.cerca, cov.model= modelo.cov, cov.pars = cov.pars)
m.cov.k <- rbind(as.matrix(cbind(m.cov, one, tend)),as.matrix(cbind(rbind(
  one,t(tend)),matrix(0,ncol=(ncol(tend)+1),nrow=(ncol(tend)+1))))))
m.cov.k.i <- solve(m.cov.k)
v.ko <- c(v,1,x.new)
Pesos.ko <- m.cov.k.i%*%v.ko
KUpr <- z[vec.orden[1:n.neigh]]%*%Pesos.ko[1:length(dist.vec.cerca)]
KUvr <- cov.0-v.ko%*%Pesos.ko
remove("dist.So","vec.orden","dist.vec.cerca","m.dist.vec","b","x.new","one",
  "tend","cov.0","m.cov","v","m.cov.k","m.cov.k.i","v.ko","Pesos.ko",
  "ValoresPropios","trend","coordinates","n.neigh","modelo.cov",
  "cov.pars","s","So")
data.frame(KUpr,KUvr)
}

```

## A.2 Funciones implementadas y utilizadas en los Capítulos 4 y 5

### A.2.1 Predicción espacial basada en distancias con funciones de base radial

1. *rbf.t*: Calcula la predicción de la variable regionalizada  $Z$  y su correspondiente varianza del error para un punto o un conjunto de puntos, para ello requiere el parámetro de suavizamiento *eta*, la variable regionalizada  $z$ , las coordenadas de la muestra *coordinates*, la tendencia asociada con las coordenadas principales de la muestra *trend*, las coordenadas espaciales junto con la tendencia asociada con las coordenadas principales de los nuevos individuos *nd.trend*, el número de vecinos más cercanos *n.neigh*, y la función de base

radial *func*.

```

assign("rbf.t",
  function(eta, z, coordinates, trend, nd.trend, n.neigh, func){
    if(func=="TPS") library(limSolve)
    rbf.pred <- as.data.frame(matrix(NA,nrow= nrow(nd.trend), ncol=4))
    colnames(rbf.pred) <- c("x","y","var1.pred","var1.var")

    rbf.t0 <- function(eta, z, coordinates, trend, nd.trend, n.neigh, func){
    newdata <- nd.trend[1:2]
    t.newdata <- nd.trend[-c(1,2)]
    dist.newdata <- as.numeric(Dis(coordinates,newdata))
    neigh.orden <- order(dist.newdata)
    dist.vec.cerca <- dist.newdata[neigh.orden[1:n.neigh]]
    trend <- trend[neigh.orden[1:n.neigh],]
    m.dist.vec <- as.matrix(dist(coordinates[neigh.orden[1:n.neigh],]))
    phi <- RBF.phi(m.dist.vec,eta,func)
    PHI.Matriz<-rbind(as.matrix(cbind(phi, 1, trend)),as.matrix(cbind(rbind(1,
      t(trend)),matrix(0,ncol=(ncol(trend)+1),nrow=(ncol(trend)+1))))))
    b <- RBF.phi(dist.vec.cerca,eta,func)
    PHI.Vector <- as.matrix(c(b,1,t.newdata))
    W.fbr <- if(func=="TPS") Solve(PHI.Matriz, PHI.Vector) else solve(PHI.Matriz,
    PHI.Vector)
    RBF.pred <- W.fbr[1:n.neigh]%%z[as.numeric(colnames(phi))]
    A <- (as.matrix(cbind(phi,1,trend))%%solve(PHI.Matriz))[1:n.neigh]
    GCV <-(t((diag(nrow(A))-A)%%z[as.numeric(colnames(phi))])%%((diag(nrow(A))
      -A)%%z[as.numeric(colnames(phi))])/n.neigh)/((sum(diag(diag(nrow(A))/
      n.neigh)))^2)
    RBF.var <-(t((diag(nrow(A))-A)%%z[as.numeric(colnames(phi))])%%((diag(nrow
      (A))-A)%%z[as.numeric(colnames(phi))]))/(sum(diag(diag(nrow(A))))))
    res <- as.matrix(cbind(RBF.pred,RBF.var))
    res[1,]
  }
  rbf.pred[,3:4] <- apply(nd.trend, 1, rbf.t0, eta=eta, z=z, trend=trend,
    coordinates=coordinates, n.neigh=n.neigh, func=func)
  rbf.pred[,1:2] <- nd.trend[,1:2]
  rbf.pred
}
)

```

2. *rbf.tr*: Esta función trabaja de manera similar a la anterior, lo adicional es el parámetro de robustez *rho*. Se utiliza para la optimización de los parámetros *eta* y *rho* en la función *bobyqa* de la librería **minqa**.

```
assign("rbf.tr",
  function(eta, z, coordinates, trend, nd.trend, rho, n.neigh, func){
    rbf.pred <- matrix(NA,nrow= nrow(newdata), ncol=4)
    rbf.t0 <-function(eta, z, coordinates, trend, nd.trend, rho, n.neigh, func){
      newdata <- nd.trend[1:2]
      t.newdata <- nd.trend[-c(1,2)]
      dist.newdata <- as.numeric(Dis(coordinates,newdata))
      neigh.orden <- order(dist.newdata)
      dist.vec.cerca <- dist.newdata[neigh.orden[1:n.neigh]]
      trend <- trend[neigh.orden[1:n.neigh],]
      m.dist.vec <- as.matrix(dist(coordinates[neigh.orden[1:n.neigh],]))
      phi <- RBF.phi(m.dist.vec,eta,func)
      PHI.Matriz <- rbind(as.matrix(cbind((phi+rho*diag(n.neigh)), 1, trend)),
        as.matrix(cbind(rbind(1,t(trend)),matrix(0,ncol=(ncol(trend)+1),
          nrow=(ncol(trend)+1))))))
      b <- RBF.phi(dist.vec.cerca,eta,func)
      PHI.Vector <- as.matrix(c(b,1,t.newdata))
      W.fbr <- solve(PHI.Matriz, PHI.Vector)
      RBF.pred <- W.fbr[1:n.neigh]%%z[as.numeric(colnames(phi))]
      A <- (as.matrix(cbind((phi+rho*diag(n.neigh)),1,trend))%%solve(PHI.Matriz))
      [,1:n.neigh]
      #GCV<-(t((diag(nrow(A))-A)%%z[as.numeric(colnames(phi))])%%((diag(nrow(A))
      #-A)%%z[as.numeric(colnames(phi))])/n.neigh)/((sum(diag(diag(nrow(A))/
      #n.neigh)))^2)
      RBF.var <- (t((diag(nrow(A))-A)%%z[as.numeric(colnames(phi))])%%
        ((diag(nrow(A))-A)%%z[as.numeric(colnames(phi))]))/(sum(diag(
        diag(nrow(A))))))
      cbind(newdata[1],newdata[2],RBF.pred,RBF.var)
    }
    rbf.pred <- apply(nd.trend, 1, rbf.t0, eta=eta, z=z, trend=trend,
      coordinates=coordinates,rho=rho, n.neigh=n.neigh, func=func)
    rbf.pred <- as.data.frame(t(rbf.pred))
    names(rbf.pred) <- c("x","y","var1.pred","var1.var")
    rbf.pred
  }
)
```

3. *rbf.t.cvop*: Esta función genera el *RMSPE* a partir de validación cruzada “leave-one-out” utilizando la función *rbf.tr*, y para su correcto funcionamiento requiere; los parámetros de suavizamiento *eta* y *rho* en *param*, la variable regionalizada *z*, las coordenadas de la muestra *coordinates*, la tendencia asociada con las coordenadas principales de la muestra *trend*, el número de vecinos más cercanos *n.neigh*, y la función de base radial *func*.

```
assign("rbf.t.cvop",
  function(param, z, coordinates, trend, n.neigh, func){
    eta <- param[1]
    rho <- param[2]
    nt <- data.frame(coordinates,trend)
    rbf.pred <- as.data.frame(matrix(NA,nrow= length(z), ncol=4))
    colnames(rbf.pred) <- c("x","y","var1.pred","var1.var")
    for(i in 1:(length(z))){
      rbf.pred[i,3] <- rbf.tr(eta, z, coordinates=coordinates[-i,], trend=trend[-i,],
        nd.trend=nt[i,], rho, n.neigh, func)[,3]
    }
    RMSPE <- sqrt(sum((rbf.pred$var1.pred-z)^2)/length(z))
    RMSPE
  }
)
```

4. *rbf.t.tcv*: Genera un data.frame con las coordenadas de los datos, la predicción, los valores observados, los residuos, la varianza de predicción, y el zscore (residual dividido por el error estándar) de la variable analizada, los cuales están asociados a la LOOCV. Para obtenerla es necesario incorporar el parámetro de suavizamiento *eta*, la variable regionalizada *z*, las coordenadas de la muestra *coordinates*, la tendencia asociada con las coordenadas principales de la muestra *trend*, el número de vecinos más cercanos *n.neigh*, y la función de base radial *func*.

```
assign("rbf.t.tcv",
  function(eta, z, coordinates, trend, n.neigh, func){
    nt <- data.frame(coordinates,trend)
    rbf.pred <- as.data.frame(matrix(NA,nrow= length(z), ncol=8))
    colnames(rbf.pred) <- c("var1.pred","var1.var","observed","residual",
```

```

        "zscore", "fold", "x", "y")
  for(i in 1:(length(z))){
    rbf.pred[i,1] <- rbf.t(eta=eta, z, coordinates=coordinates[-i,],
                          trend=trend[-i,], nd.trend=nt[i,], n.neigh, func)[,3]
    rbf.pred[i,6] <- i
  }
  rbf.pred[,3]<- z
  rbf.pred[,7:8]<-coordinates
  rbf.pred[,4]<- rbf.pred[,3]-rbf.pred[,1]
  rbf.pred
}
)

```

## A.2.2 Predicción espacio-temporal basada en distancias usando funciones de base radial

1. *rbf.trst*: Calcula la predicción de la variable regionalizada  $Z(s, t)$  para un punto o un conjunto de puntos espacio-tiempo, para ello requiere los parámetros de suavizamiento *eta* y de robustez *rho*, la variable regionalizada *z*, las coordenadas de la muestra espacio-temporales *coordinates*, la tendencia asociada con las coordenadas principales de la muestra *trend*, las coordenadas espacio-temporales junto con la tendencia asociada con las coordenadas principales de los nuevos individuos *nd.trend*, el número de vecinos espacio-temporales más cercanos *n.neigh*, y la función de base radial *func*.

```

assign("rbf.trst",
  function(eta, z, coordinates, trend, nd.trend, rho, n.neigh, func){
  library(limSolve)
    rbf.pred <- matrix(NA, nrow= nrow(trend), ncol=5)
  rbf.t0 <-function(eta, z, coordinates, trend, nd.trend, rho, n.neigh, func){
  newdata <- nd.trend[1:3]
  t.newdata <- nd.trend[-c(1:3)]
  dist.newdata <- as.numeric(Dis(coordinates, newdata))
  vec.orden <- order(dist.newdata)
  dist.vec.cerca <- dist.newdata[vec.orden[1:n.neigh]]
  trend <- trend[vec.orden[1:n.neigh],]
  m.dist.vec <- as.matrix(dist(coordinates[vec.orden[1:n.neigh],] ))
  }
  }

```



```

phi <- RBF.phi(m.dist.vec,eta,func)
PHI.Matriz <- rbind(as.matrix(cbind((phi+rho*diag(n.neigh)), 1, trend)),
as.matrix(cbind(rbind(1,t(trend)),matrix(0,ncol=(ncol(trend)+1),
nrow=(ncol(trend)+1))))))
b <- RBF.phi(dist.vec.cerca,eta,func)
PHI.Vector <- as.matrix(c(b,1,t.newdata))
W.fbr <- Solve(PHI.Matriz, as.numeric(PHI.Vector))
RBF.pred <- W.fbr[1:n.neigh]%*%z[vec.orden[1:n.neigh]]
cbind(newdata[1],newdata[2],newdata[3],RBF.pred,NA) #RBF.var)
}
  rbf.pred <- apply(nd.trend, 1, rbf.t0, eta=eta, z=z, trend=trend,
  coordinates=coordinates,rho=rho, n.neigh=n.neigh, func=func)
  rbf.pred <- as.data.frame(t(rbf.pred))
  names(rbf.pred) <- c("x","y","t","var1.pred","var1.var")
  rbf.pred
}
)

```

2. *rbf.st.cvop*: Esta función genera el *RMSPE* a partir de validación cruzada “leave-one-out” espacio-temporal utilizando la función *rbf.tr*, y para su correcto funcionamiento requiere; los parámetros de suavizamiento *eta* y *rho* en *param*, la variable regionalizada *z*, las coordenadas espacio-tiempo de la muestra *coordinates*, la tendencia asociada con las coordenadas principales de la muestra *trend*, el número de vecinos mas cercanos espacio-temporales *n.neigh*, y la función de base radial *func*.

```

assign("rbf.st.cvop",
  function(param, z, coordinates, trend, n.neigh, func){
    eta <- param[1]
    rho <- param[2]
    nt <- data.frame(coordinates,trend)
    rbf.pred <- as.data.frame(matrix(NA,nrow= length(z), ncol=5))
    colnames(rbf.pred) <- c("x","y","t","var1.pred","var1.var")
    for(i in 1:(length(z))){
      rbf.pred[i,4] <- rbf.trst(eta, z, coordinates=coordinates[-i,],
      trend=trend[-i,], nd.trend=nt[i,], rho, n.neigh, func)[,4]
    }
    RMSPE <- sqrt(sum((rbf.pred$var1.pred-z)^2)/length(z))
    RMSPE
  }
)

```

```
}
)
```

3. *rbf.st.tcv*: Genera un data.frame con las coordenadas de los datos y el tiempo, la predicción, los valores observados, los residuos, la varianza de predicción, y el zscore (residual dividido por el error estándar) de la variable analizada, los cuales están asociados a la LOOCV. Para obtenerla es necesario incorporar los parámetros; de suavizamiento *eta* y de robustez *rho*, la variable regionalizada *z*, las coordenadas espacio-temporales de la muestra *coordinates*, la tendencia asociada con las coordenadas principales de la muestra *trend*, el número de vecinos espacio-temporales más cercanos *n.neigh*, y la función de base radial *func*.

```
assign("rbf.st.tcv",
      function(eta, z, coordinates, trend, rho, n.neigh, func){
        nt <- data.frame(coordinates,trend)
        rbf.pred <- as.data.frame(matrix(NA,nrow= length(z), ncol=9))
        colnames(rbf.pred) <- c("var1.pred","var1.var","observed","residual",
                               "zscore","fold","x","y","t")
        for(i in 1:(length(z))){
          rbf.pred[i,1:2] <- rbf.trst(eta=eta, z, coordinates=coordinates[-i,],
                                     trend=trend[-i,], nd.trend=nt[i,], rho, n.neigh, func)[,4:5]
          rbf.pred[i,6] <- i
        }
        rbf.pred[,3]<- z
        rbf.pred[,5]<- (z-mean(z))/sqrt(rbf.pred[,2])
        rbf.pred[,7:9]<-coordinates
        rbf.pred[,4]<- rbf.pred[,3]-rbf.pred[,1]
        rbf.pred
      }
)
```

## A.3 Programación capítulo 3

```
-----
#####  CAPITULO 3: SIMULACIÓN KRIGING ESPACIAL  #####
-----
```

```

library(gstat)
library(geoR)
library(ade4)
library(cluster)
library(rgdal)
library(splancs)
library("sp")
library("maptools")

#-----
#### SIMULACIÓN 1: GRF SIN OMISIÓN DE VARIABLE, UNIVERSAL KRIGING DB ####
#-----

p1 <- cbind( c(0,1,1,0), c(1,1,0.5,1))
p2 <- cbind( c(0,1,0,0), c(1,0.5,0.3,1))
p3 <- cbind( c(0,0,1,1,0), c(0,0.3,0.5,0,0))

poly1 <- Polygons(list(Polygon(p1)), "R1")
poly2 <- Polygons(list(Polygon(p2)), "R2")
poly3 <- Polygons(list(Polygon(p3)), "R3")
sppo <- SpatialPolygons(list(poly1,poly2,poly3))

#####
###      CONSTRUCCION GRAFICO REGIONES      ###
#####

par(mfrow=c(1,1), mar=c(5,5,4,4))
plot(sppo, ylim=c(0,1), xlab="x", ylab="y")
marcas <- seq(0,1,0.1)
axis(side=2, at=marcas)
axis(side=1, at=marcas, xlab = "x")
names <- unlist(lapply(slot(sppo,"polygons"), function(x) slot(x,"ID")))
text(coordinates(sppo), labels=names)
text(0.5,-0.1, "X", cex=1)
title(xlab="X")
title(ylab="Y")
# points(sim$coords, cex= 0.4)  # Correr esto una vez se tenga las
# coordenadas de una simulación

```

```

#####
###          CALCULO AREAS POLIGONOS          ###
#####

areapl(p1)
areapl(p2)
areapl(p3)

getClass("Polygon")
areas <- sapply(slot(sppo,"polygons"),function(x) sapply(slot(x,"Polygons"),
slot, "area"))
str(areas)
porc.area1 <- areas[1]/sum(areas)      # Para el caso es la misma area
porc.area2 <- areas[2]/sum(areas)
porc.area3 <- areas[3]/sum(areas)

100*areas[1]
50*areas[2]
50*areas[3]

Muestra1 <- c(13,17,20) # Regiones 1,2 y 3 respectivamente # 50 datos
Muestra2 <- c(25,35,40) # 100 datos
Muestra3 <- c(28,52,60) # 150 datos

#####
#####          SIMULACIÓN SIN OMISIÓN DE VARIABLE GRF          #####
#####

#_-----
# Valores parámetros

B0 <- 10; B1 <- -4; B2 <- 2; B3 <- -4; p <- 0.4

UKDB.sim <- function(N,nsim,nugget,range1,sill,ang, anis,kappa1,lambda,
                    modelo.cov, cov.model,p){

psill <- c(nugget,sill)
range <- c(0,range1)

```

```

ang1 <- c(0,ang)
anis1 <- c(1,anis)
kappa <- c(0,kappa1)
cov.pars0 <- as.matrix(cbind(psil1,range,kappa,ang1,anis1))

set.seed(127)
sim <- grf(N,nsim=105,grid= "irreg",cov.model=cov.model,xlims=c(0, 1),
          lambda=lambda, ylims = c(0, 1), cov.pars=cov.pars0)

set.seed(127)
V1 <- rbinom(N, size=1, prob=p)

#build nominal covariate vector for prediction locations indicating subarea#
ind.reg.sim <- numeric(nrow(sim$coords))
ind.reg.sim[.geoR_inout(sim$coords,p1)] <- 1
ind.reg.sim[.geoR_inout(sim$coords,p2)] <- 2
ind.reg.sim[.geoR_inout(sim$coords,p3)] <- 3
ind.reg.sim <- as.factor(ind.reg.sim)

V2 <- ifelse(ind.reg.sim %in% 2,1,0)
V3 <- ifelse(ind.reg.sim %in% 3,1,0)

sim$data<-B0+B1*V1+B2*V2*sim$coords[,1]+B3*V3*sim$coords[,2]+sim$data

tabla.sim <- as.data.frame(matrix(NA,nrow= nsim, ncol=20))
colnames(tabla.sim)<-c("sim","C1","a","Co","Kappa","MPE","ASEPE","RMSPE",
"R2","C1db","adb","Codb","Kappadb","MPEdb","ASEPEdb","RMSPEdb","R2db",
"AICc","AICdb","No.CP")

Data.sim <- data.frame(sim$coords[,1],sim$coords[,2],V1,V2,V3)
names(Data.sim) <- c("x","y","V1","V2","V3")

#####
#####      CONSTRUCCION COORDENADAS PRINCIPALES Y KRIGEADO DB      #####
#####

Delt<-daisy(Data.sim,type=list(asymm=c("V1","V2","V3")),metric="gower")
class(Delt)
is.euclid(Delt^(1/2))

```

```

mds <- cmdscale(Delt^(1/2), k = nrow(Data.sim)-1, eig = TRUE)
names(mds)
round(mds$points[,1],4)

m <- sum(mds$eig > 0.007)
mds <- cmdscale(Delt^0.5, k = m, eig = TRUE)
X <- mds$points

for(i in 1:100){

ValoresPropios <- mds$eig
CorrCuadrado <- as.vector(cor(sim$data[,i],X)^2)
Porc.Inercia <- ValoresPropios/length(CorrCuadrado)
o<-data.frame(1:length(CorrCuadrado),round(ValoresPropios[1:length
  (CorrCuadrado)],10),round(CorrCuadrado,10),round(Porc.Inercia[1:length
  (CorrCuadrado)],10))
names(o)<-c("ID", "ValoresProp", "CorrCuad", "Porc.Inercia")

names(o)
o1<-o[o$CorrCuad>0.007,]
Xr.sim <- X[,o1$ID]
rdb.sim <- lm(sim$data[,i] ~ Xr.sim)
model.db.sim <- summary(rdb.sim)

x <- sim$coords[,1]
y <- sim$coords[,2]
z <- sim$data[,i]
dXr.sim <- data.frame(x,y,z,Xr.sim)
sim.gd <- as.geodata(dXr.sim, coords.col = 1:2, data.col = 3,
  covar.col=4:(ncol(Xr.sim)+3))

## Create a formula for a model with a large number of variables:
xnam <- paste("X", 1:ncol(Xr.sim), sep="")
fmla <- as.formula(paste(" ~ ", paste(xnam, collapse= "+")))
try({m4.db.sim <- likfit(sim.gd, trend=fmla, cov.model= cov.model,
  ini = c(sill,range1), kappa=kappa1, nug=nugget, messages=F)})
ls()
f <- ifelse(ls() %in% "m4.db.sim",1,0)
w <- sum(f)
phid <- ifelse(w==0,range1,m4.db.sim$phi)

```

```

sigmasqd <- ifelse(w==0,sill,m4.db.sim$sigmasq)
tausqd <- ifelse(w==0,nugget,m4.db.sim$tausq)
kappad <- ifelse(w==0,kappa,m4.db.sim$kappa)
m4.db.sim0 <- likfit(s100, ini=c(0.5, 0.5), fix.nug = TRUE)
m4.db.sim0$phi <- phid
m4.db.sim0$sigmasq <- sigmasqd
m4.db.sim0$tausq <- tausqd
m4.db.sim0$kappa <- kappad

m4.db.sim0$phi <- ifelse(m4.db.sim0$phi==0,range1,m4.db.sim0$phi)
m4.db.sim0$sigmasq <- ifelse(m4.db.sim0$sigmasq==0,sill,m4.db.sim0$sigmasq)

vuk.db.sim <- vgm(m4.db.sim0$sigmasq,modelo.cov,m4.db.sim0$phi,
                 m4.db.sim0$tausq,kappa=m4.db.sim0$kappa)
g1.db.sim <- gstat(id = "z", formula = z~X1, locations = ~x+y,
                  model=vuk.db.sim, data = dXr.sim)

## Create a formula for a model with a large number of variables:
g1.db.sim$data$z$formula<-as.formula(paste("z~",paste(xnam,collapse="+")))
kcv.db.sim <- gstat.cv(g1.db.sim,nfold=N, nmax=40)
MPEdb <- mean(kcv.db.sim$residual)
ASEPEdb <- mean(sqrt(kcv.db.sim$z.var))
RMSPEdb <- sqrt(sum(kcv.db.sim$residual^2)/length(kcv.db.sim$residual))
resid.mean.sim <- dXr.sim$z- mean(dXr.sim$z)
R2db <- 1- sum((kcv.db.sim$residual)^2)/sum(resid.mean.sim^2)

#####
#####          CONSTRUCCION KRIGEADO UNIVERSAL CLASICO          #####
#####

dataX.sim <- data.frame(x,y,z,Data.sim[,-(1:2)])
sim.gdc <- as.geodata(dataX.sim, coords.col = 1:2, data.col = 3,
                     covar.col=4:ncol(dataX.sim))

try({m5.sim <- likfit(sim.gdc, trend=~coords+covariate$V1+covariate$V2+
                     covariate$V3, cov.model= cov.model,
                     kappa=kappa1, ini = c(sill,range1), nug=nugget, messages=F)})
ls()
f1 <- ifelse(ls() %in% "m5.sim",1,0)
w1 <- sum(f1)

```

```

phic <- ifelse(w1==0,range1,m5.sim$phi)
sigmasqc <- ifelse(w1==0,sill,m5.sim$sigmasq)
tausqc <- ifelse(w1==0,nugget,m5.sim$tausq)
kappac <- ifelse(w1==0,kappa,m5.sim$kappa)
m5.sim1 <- likfit(s100, ini=c(0.5, 0.5), fix.nug = TRUE)
m5.sim1$phi <- phic
m5.sim1$sigmasq <- sigmasqc
m5.sim1$tausq <- tausqc
m5.sim1$kappa <- kappac

m5.sim1$phi <- ifelse(m5.sim1$phi==0,range1,m5.sim1$phi)
m5.sim1$sigmasq <- ifelse(m5.sim1$sigmasq==0,sill,m5.sim1$sigmasq)

vuk.sim <- vgm(m5.sim1$sigmasq,modelo.cov,m5.sim1$phi,m5.sim1$tausq,
              kappa=m5.sim1$kappa)
g1.sim <- gstat(id = "z", formula = z~x+y+V1+V2+V3, locations = ~x+y,
              model=vuk.sim, data = dataX.sim)
kcv.sim <- gstat.cv(g1.sim,nfold=N, nmax=40)
MPE <- mean(kcv.sim$residual)
ASEPE <- mean(sqrt(kcv.sim$z.var))
RMSPE <- sqrt(sum(kcv.sim$residual^2)/length(kcv.sim$residual))
R2 <- 1- sum((kcv.sim$residual)^2)/sum(resid.mean.sim^2)

tabla.sim[i,]<- c(i,m5.sim1$sigmasq,m5.sim1$phi,m5.sim1$tausq,m5.sim1$kappa,
  MPE,ASEPE,RMSPE,R2,m4.db.sim0$sigmasq,m4.db.sim0$phi,m4.db.sim0$tausq,
  kappa=m4.db.sim0$kappa,MPedb,ASEPEdb,RMSPEdb,R2db,AIC(m5.sim),
  AIC(m4.db.sim),ncol(Xr.sim))
}
tabla.sim
}

tabla1.s.o<-UKDB.sim(N=50,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=0.5,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
tabla2.s.o<-UKDB.sim(N=100,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=0.5,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
tabla3.s.o<-UKDB.sim(N=150,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=0.5,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)

tabla4.s.o<-UKDB.sim(N=50,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=3/2,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)

```



```

tabla5.s.o<-UKDB.sim(N=100,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=3/2,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
tabla6.s.o<-UKDB.sim(N=150,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=3/2,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
.
.
.
tabla94.sim<-UKDB.sim(N=50,nsim=100,nugget=1,range1=0.6,sill=2,ang=0,anis=1,
  kappa1=0,lambda=1,modelo.cov="Sph",cov.model="spherical",p=0.4)
tabla95.sim<-UKDB.sim(N=100,nsim=100,nugget=1,range1=0.6,sill=2,ang=0,anis=1,
  kappa1=0,lambda=1,modelo.cov="Sph",cov.model="spherical",p=0.4)
tabla96.sim<-UKDB.sim(N=150,nsim=100,nugget=1,range1=0.6,sill=2,ang=0,anis=1,
  kappa1=0,lambda=1,modelo.cov="Sph",cov.model="spherical",p=0.4)

#-----
### SIMULACIÓN 2: GRF CON OMISIÓN DE VARIABLE, UNIVERSAL KRIGING DB ###
#-----

UKDB.sim <- function(N,nsim,nugget,range1,sill,ang,anis,kappa1,lambda,
  modelo.cov, cov.model,p){
  psill <- c(nugget,sill)
  range <- c(0,range1)
  ang1 <- c(0,ang)
  anis1 <- c(1,anis)
  kappa <- c(0,kappa1)
  cov.pars0 <- as.matrix(cbind(psill,range,kappa,ang1,anis1))

  set.seed(127)
  sim <- grf(N,nsim=105,grid="irreg",cov.model=cov.model,xlims=c(0, 1),
    lambda=lambda, ylims = c(0, 1), cov.pars=cov.pars0)
  set.seed(127)
  V1 <- rbinom(N, size=1, prob=p)

#build nominal covariate vector for prediction locations indicating subarea#
ind.reg.sim <- numeric(nrow(sim$coords))
ind.reg.sim[.geoR_inout(sim$coords,p1)] <- 1
ind.reg.sim[.geoR_inout(sim$coords,p2)] <- 2
ind.reg.sim[.geoR_inout(sim$coords,p3)] <- 3

```

```

ind.reg.sim <- as.factor(ind.reg.sim)

V2 <- ifelse(ind.reg.sim %in% 2,1,0)
V3 <- ifelse(ind.reg.sim %in% 3,1,0)

sim$data <- B0+B1*V1+B2*V2*sim$coords[,1]+B3*V3*sim$coords[,2]+sim$data

tabla.sim <- as.data.frame(matrix(NA,nrow= nsim, ncol=20))
colnames(tabla.sim)<-c("sim","C1","a","Co","Kappa","MPE","ASEPE","RMSPE",
"R2","C1db","adb","Codb","Kappadb","MPEdb","ASEPEdb","RMSPEdb","R2db",
"AICc","AICdb","No.CP")

Data.sim <- data.frame(sim$coords[,1],sim$coords[,2],V2,V3)
names(Data.sim) <- c("x","y","V2","V3")

#####
##### CONSTRUCCION COORDENADAS PRINCIPALES Y KRIGEADO DB #####
#####

Delt <- daisy(Data.sim, type=list(asymm=c("V2","V3")), metric ="gower")
class(Delt)
library(ade4)
is.euclid(Delt^(1/2))

mds <- cmdscale(Delt^(1/2), k = nrow(Data.sim)-1, eig = TRUE)
names(mds)
round(mds$points[,1],4)

m <- sum(mds$eig > 0.007)
mds <- cmdscale(Delt^0.5, k = m, eig = TRUE)
X <- mds$points

for(i in 1:100){
ValoresPropios <- mds$eig
CorrCuadrado <- as.vector(cor(sim$data[,i],X)^2)
Porc.Inercia <- ValoresPropios/length(CorrCuadrado)
o<-data.frame(1:length(CorrCuadrado),round(ValoresPropios[1:length
(CorrCuadrado)],10),round(CorrCuadrado,10),round(Porc.Inercia[1:
length(CorrCuadrado)],10))
names(o)<-c("ID", "ValoresProp", "CorrCuad", "Porc.Inercia")

```

```

names(o)
o1<-o[o$CorrCuad>0.007,]
Xr.sim <- X[,o1$ID]
rdb.sim <- lm(sim$data[,i] ~ Xr.sim)
model.db.sim <- summary(rdb.sim)

x <- sim$coords[,1]
y <- sim$coords[,2]
z <- sim$data[,i]
dXr.sim <- data.frame(x,y,z,Xr.sim)
sim.gd <- as.geodata(dXr.sim, coords.col = 1:2, data.col = 3,
                    covar.col=4:(ncol(Xr.sim)+3))
## Create a formula for a model with a large number of variables:
xnam <- paste("X", 1:ncol(Xr.sim), sep="")
fmla <- as.formula(paste(" ~ ", paste(xnam, collapse="+")))
try({m4.db.sim <- likfit(sim.gd, trend=fmla, cov.model= cov.model,
    ini = c(sill,range1), kappa=kappa1, nug=nugget, messages=F)})
ls()
f <- ifelse(ls() %in% "m4.db.sim",1,0)
w <- sum(f)
phid <- ifelse(w==0,range1,m4.db.sim$phi)
sigmasqd <- ifelse(w==0,sill,m4.db.sim$sigmasq)
tausqd <- ifelse(w==0,nugget,m4.db.sim$tausq)
kappad <- ifelse(w==0,kappa,m4.db.sim$kappa)
m4.db.sim0 <- likfit(s100, ini=c(0.5, 0.5), fix.nug = TRUE)
m4.db.sim0$phi <- phid
m4.db.sim0$sigmasq <- sigmasqd
m4.db.sim0$tausq <- tausqd
m4.db.sim0$kappa <- kappad
AICdb <- ifelse(f==1, "NA", AIC(m4.db.sim))

m4.db.sim0$phi <- ifelse(m4.db.sim0$phi==0,range1,m4.db.sim0$phi)
m4.db.sim0$sigmasq<-ifelse(m4.db.sim0$sigmasq==0,sill,m4.db.sim0$sigmasq)

vuk.db.sim<-vgm(m4.db.sim0$sigmasq,modelo.cov,m4.db.sim0$phi,
               m4.db.sim0$tausq, kappa=m4.db.sim0$kappa)
g1.db.sim <- gstat(id = "z", formula = z~X1, locations = ~x+y,
                 model=vuk.db.sim, data = dXr.sim)
## Create a formula for a model with a large number of variables:

```

```

g1.db.sim$data$z$formula<-as.formula(paste("z~",paste(xnam,collapse="+")))
kcv.db.sim <- gstat.cv(g1.db.sim,nfold=N)
MPEdb      <- mean(kcv.db.sim$residual)
ASEPEdb    <- mean(sqrt(kcv.db.sim$z.var))
RMSPEdb    <- sqrt(sum(kcv.db.sim$residual^2)/length(kcv.db.sim$residual))
resid.mean.sim <- dXr.sim$z- mean(dXr.sim$z)
R2db <- 1- sum((kcv.db.sim$residual)^2)/sum(resid.mean.sim^2)

#####
#####          CONSTRUCCION KRIGEADO UNIVERSAL CLASICO          #####
#####

dataX.sim <- data.frame(x,y,z,Data.sim[-(1:2)])
sim.gdc <- as.geodata(dataX.sim, coords.col = 1:2, data.col = 3,
                      covar.col=4:ncol(dataX.sim))

try({m5.sim <- likfit(sim.gdc, trend=~coords+covariate$V2+covariate$V3,
cov.model=cov.model,ini=c(sill,range1),kappa=kappa1,nug=nugget,messages=F)})
ls()
f1 <- ifelse(ls() %in% "m5.sim",1,0)
w1 <- sum(f1)
phic <- ifelse(w1==0,range1,m5.sim$phi)
sigmasqc <- ifelse(w1==0,sill,m5.sim$sigmasq)
tausqc <- ifelse(w1==0,nugget,m5.sim$tausq)
kappac <- ifelse(w1==0,kappa,m5.sim$kappa)
m5.sim1 <- likfit(s100, ini=c(0.5, 0.5), fix.nug = TRUE)
m5.sim1$phi <- phic
m5.sim1$sigmasq <- sigmasqc
m5.sim1$tausq <- tausqc
m5.sim1$kappa <- kappac

m5.sim1$phi <- ifelse(m5.sim1$phi==0,range1,m5.sim1$phi)
m5.sim1$sigmasq <- ifelse(m5.sim1$sigmasq==0,sill,m5.sim1$sigmasq)

vuk.sim <- vgm(m5.sim1$sigmasq,modelo.cov,m5.sim1$phi,m5.sim1$tausq,
              kappa=m5.sim1$kappa)
g1.sim <- gstat(id= "z", formula = z~x+y+V2+V3, locations= ~x+y,
               model=vuk.sim, data = dataX.sim)
kcv.sim <- gstat.cv(g1.sim,nfold=N)

```

```

MPE      <- mean(kcv.sim$residual)
ASEPE    <- mean(sqrt(kcv.sim$z.var))
RMSPE    <- sqrt(sum(kcv.sim$residual^2)/length(kcv.sim$residual))
R2       <- 1- sum((kcv.sim$residual)^2)/sum(resid.mean.sim^2)

tabla.sim[i,]<-c(i,m5.sim1$sigmasq,m5.sim1$phi,m5.sim1$tausq,m5.sim1$kappa,
  MPE,ASEPE,RMSPE,R2,m4.db.sim0$sigmasq,m4.db.sim0$phi,m4.db.sim0$tausq,
  kappa=m4.db.sim0$kappa,MPEdb,ASEPEdb,RMSPEdb,R2db,AIC(m5.sim),
  AIC(m4.db.sim), ncol(Xr.sim))
}
tabla.sim
}

tabla1.c.o<-UKDB.sim(N=50,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=0.5,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
tabla2.c.o<-UKDB.sim(N=100,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=0.5,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
tabla3.c.o<-UKDB.sim(N=150,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=0.5,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)

tabla4.c.o<-UKDB.sim(N=50,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=3/2,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
tabla5.c.o<-UKDB.sim(N=100,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=3/2,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
tabla6.c.o<-UKDB.sim(N=150,nsim=100,nugget=0,range1=0.15,sill=1,ang=0,anis=1,
  kappa1=3/2,lambda=1,modelo.cov="Mat",cov.model="matern",p=0.4)
.
.
.
tabla94.c.o<-UKDB.sim(N=50,nsim=100,nugget=1,range1=0.6,sill=2,ang=0,anis=1,
  kappa1=0,lambda=1,modelo.cov="Sph",cov.model="spherical",p=0.4)
tabla95.c.o<-UKDB.sim(N=100,nsim=100,nugget=1,range1=0.6,sill=2,ang=0,anis=1,
  kappa1=0,lambda=1,modelo.cov="Sph",cov.model="spherical",p=0.4)
tabla96.c.o<-UKDB.sim(N=150,nsim=100,nugget=1,range1=0.6,sill=2,ang=0,anis=1,
  kappa1=0,lambda=1,modelo.cov="Sph",cov.model="spherical",p=0.4)

#-----
####  CAPITULO 3: KRIGING UNIVERSAL DB VS KRIGING UNIVERSAL CLASICO  ####
#-----

```

```

#-----
## A. APLICACIÓN TEMPERATURA CROATIA 01-12-2008 ##
#-----

library(sp)
library(maptools)
library(gstat)
library(fields)
library(rgdal)
library(lattice)
library(spatstat)
library(RSAGA)
library(cluster)
utm33 <- "+proj=utm +zone=33 +ellps=WGS84 +datum=WGS84 +units=m +no_defs"

# Download auxiliary maps and station measurements:
download.file("http://spatial-analyst.net/book/sites/default/files/
HRclim2008.zip",destfile=paste(getwd(), "HRclim2008.zip", sep="/"))
unzip(zipfile="HRclim2008.zip", exdir=getwd())
unlink("HRclim2008.zip")

# Download MODIS LST images:
download.file("http://spatial-analyst.net/book/sites/default/files/
LST2008HR.zip",destfile=paste(getwd(), "LST2008HR.zip", sep="/"))
unzip(zipfile="LST2008HR.zip", exdir=paste(getwd(), "/LST", sep="/"))
unlink("LST2008HR.zip")

# -----
# STEP 1: Importación de los datos y formateo (estandarización);
# -----

# Import location of stations:
IDSTA<-read.table("stations_temp_xy_2008.csv",header=TRUE,sep="," ,quote="\")
coordinates(IDSTA) <- ~LON+LAT
proj4string(IDSTA)<-CRS("+proj=longlat+ellps=bessel+towgs84=550.499,164.116,
475.142,5.80967,2.07902,-11.62386,0.99999445824")
# HRG2000 geoid [http://spatial-analyst.net/wiki/index.php?title=MGI/_
_Balkans_coordinate_systems]
IDSTA.ll <- spTransform(IDSTA, CRS("+proj=longlat+ellps=WGS84+datum=WGS84"))

```

```

writeOGR(IDSTA.ll, "gl_stations.kml", "IDSTA", "KML")
IDSTA.utm <- spTransform(IDSTA, CRS(utm33))
writeOGR(IDSTA.utm, "gl_stations.shp", "IDSTA", "ESRI Shapefile")
locs <- as.data.frame(IDSTA.utm)
names(locs)[c(5,6)] <- c("X", "Y")
str(locs)
# ST_ID      : Identification code of a station
# NAME       : Station name
# SP_CODE    : Station type "gl"-main, "kl"-climatological, "ks"-precipitation
# ELEV       : Station elevation in official database

# convert to lines:
rsaga.geoprocessor(lib="shapes_lines",module=0,param=list(LINES=
                  "borders.shp", POLYGONS="countries_s.shp"))
# import country borders:
borders <- readOGR(".", "borders")

# Croatia.shp fue generado en ArcGIS 10.0, allí se eliminaron los bordes de
# otros países del archivo "countries_s.shp"
croatia.shp <- readShapePoly("D:/...../croatia.shp")

# Import temperature at stations:
HRtemp2008 <- read.delim("temp_2008.csv", header=TRUE, sep=",", quote="\")
# NA values:
HRtemp2008$T07[HRtemp2008$T07== -99.9] <- NA; HRtemp2008$T14[HRtemp2008$T14==
-99.9] <- NA; HRtemp2008$T21[HRtemp2008$T21== -99.9] <- NA
HRtemp2008$TEMP <- (HRtemp2008$T07+HRtemp2008$T14+2*HRtemp2008$T21)/4
str(HRtemp2008)# Mean daily temperature for 365 days (2008) at 152 locations;
summary(HRtemp2008$TEMP) # 712 NA's
HRtemp2008$ST_ID <- as.factor(HRtemp2008$ST_ID)
# format the DATES (convert to number of cumulative days since 1970-01-01):
HRtemp2008$DATE <- as.Date(as.character(HRtemp2008$DATE))
HRtemp2008$cdays <- floor(unclass(as.POSIXct(HRtemp2008$DATE))/86400)
floor(unclass(as.POSIXct("2008-01-30"))/86400)[[1]]

# stations without measurements:
dif.IDSTA <- merge(locs, IDSTA.utm, by.x="ST_ID", all.x=TRUE)
dif.IDSTA1 <- merge(dif.IDSTA, HRtemp2008, by.x="ST_ID", all.x=TRUE)
dif.IDSTA2 <- merge(dif.IDSTA1, IDSTAPREC, by.x=c("ST_ID", "DATE"), all.x=TRUE)
dif.IDSTA3 <- dif.IDSTA2[(dif.IDSTA2$DATE %in% "2008-12-01"),]

```

```

names(dif.IDSTA3)

# temperature the day before:
tmp <- data.frame(ST_ID=HRtemp2008$ST_ID, DATE=HRtemp2008$DATE+1,
                 TEMP1M=HRtemp2008$TEMP)
HRtemp2008.f <- merge(HRtemp2008, tmp, by=c("DATE", "ST_ID"), all.x=TRUE)
# check numbers:
HRtemp2008.f[(0:5)*(159-1)+1,c("ST_ID", "DATE", "TEMP", "TEMP1M")]

# Import grids:
grids <- readGDAL("HRdem.asc")
names(grids@data)[1] <- "HRdem"
for(j in c("HRdsea", "HRtwi")){
  grids@data[,j] <- readGDAL(paste(j, ".asc", sep=""))$band1
}
proj4string(grids) <- CRS(utm33)
# create dummy grids (Lat/Lon):
grids.ll <- spTransform(grids[1], CRS("+proj=longlat +datum=WGS84"))
grids$Lat <- grids.ll@coords[,2]
grids$Lon <- grids.ll@coords[,1]
str(grids@data)
# spplot(grids["Lat"])
lat.c <- mean(grids.ll@coords[,2])

# -----
# STEP 2: importación de las grillas y derivación de INSOL;
# -----

# derive total solar insolation for 365 days:
writeGDAL(grids["HRdem"], "HRdem.sdat", "SAGA")
writeGDAL(grids["Lat"], "HRlat.sdat", "SAGA")
writeGDAL(grids["Lon"], "HRlon.sdat", "SAGA")

Sys.chmod(getwd(), mode="0777"); dir.create("INSOL")
for(j in 1:365){
  rsaga.geoprocessor(lib="ta_lighting",module=3,param=list(GRD_DEM="HRdem.sgrd"
,GRD_LAT="HRlat.sgrd",GRD_LON="HRlon.sgrd", GRD_TOTAL=paste("INSOL/", "INSOL"
,j,".sgrd",sep=""),PERIOD=1,DHOUR=4,DAY_A=j),show.output.on.console=FALSE)
# read to R:
grids@data[,paste("INSOL", j, sep="")]<-readGDAL(paste("INSOL/", "INSOL", j,

```



```

".sdat", sep=""), silent=TRUE)$band1
} # takes cca 30 mins!!

# Import LST MODIS images
# List of images:
LST.listday <- dir(path=paste(getwd(), "LST", sep="/"),
  pattern=glob2rx("LST2008_**_**.LST_Day_1km.tif"), full.names=TRUE)
for(i in 1:length(LST.listday)){
LSTname<-strsplit(strsplit(LST.listday[i],"LST/")[1][2],".LST_")[1][1]
tmp1 <- readGDAL(LST.listday[i], silent=TRUE)
# convert to celsius:
grids@data[,LSTname]<-ifelse(tmp1$band1<=7500, NA, tmp1$band1*0.02-273.15)
}

# -----
# STEP 3: Generación mapa de ubicación estaciones y geodata;
# -----

## Construccion Mapa Ubicacion Estaciones Croacia
par(mfrow=c(1,1), mar=c(5,5,4,4))
plot(croatia.shp, col="gray",border="blue", axes=F,)
at.y <- (47:51)*100000
axis(2, at = at.y, labels = formatC(at.y, format="fg"))
my.at <- (4:8)*100000
axis(1, at = my.at, labels = formatC(my.at, format="fg"))
box(lty = 'solid', )
points(dif.IDSTA3$X,dif.IDSTA3$Y,cex=0.4)
title(xlab="X")
title(ylab="Y")

# Se convierten a un formato de clase "as.geodata"

IDSTA.geoR <- as.geodata(dif.IDSTA3, coords.col = 3:4, data.col = 13)
plot(IDSTA.geoR)
points.geodata(IDSTA.geoR, x.leg=3,y.leg=5,main=c("Gráfico de Intensidades
de Temperatura", "(Celsius) 2008-12-01"), col.main=3, pt.div="quintile")
# Graficar los datos (puntos) y ver simbolos graduados, además la opción
# "add.to.plot" es útil para adicionar opciones de la instrucción "plot",
# como en este caso la grilla.
points.geodata(IDSTA.geoR,x.leg=3,y.leg=5,main=c("Gráfico de Intensidades",

```

```

"de Temperatura"), col.main=3,pt.div="quintile", add.to.plot = TRUE,
panel.first = grid())
#Valores de precipitación, desplazados para no superponerlos con los puntos
text(dif.IDSTA3$X-15000,dif.IDSTA3$Y,round(dif.IDSTA3$TEMP,1),col=1,cex=0.6)

#-----
#STEP 4: Calculo de coordenadas principales y regresión basada en distancias
#-----

Delta <- daisy(cbind(dif.IDSTA3$X, dif.IDSTA3$Y), metric ="euclidean")
class(Delta)
library(ade4)
is.euclid(Delta)

mds <- cmdscale(Delta, k = n-1, eig = TRUE)
names(mds)
round(mds$points[,1],4)

m <- sum(mds$eig > 1.0e-15)
mds <- cmdscale(Delta, k = m, eig = TRUE)
X <- mds$points

ValoresPropios <- mds$eig
CorrCuadrado <- as.vector(cor(y,X)^2)
Porc.Inercia <- ValoresPropios/sum(ValoresPropios)
o<-data.frame(1:77,round(ValoresPropios[1:77],7),
              round(CorrCuadrado[1:77],7),round(Porc.Inercia[1:77],7))
names(o)<-c("ID", "ValoresProp", "CorrCuad", "Porc.Inercia")
fix(o)
length(ValoresPropios)

aux <- CorrCuadrado[1:83]*ValoresPropios[1:83]/sum(CorrCuadrado[1:83]*
          ValoresPropios[1:83])
c.pred <- c(0,cumsum(aux))

plot(0:m,1-c.pred,xlab="Número de coordenadas principales",ylab="1 -
      Predictibilidad",main=c("Gráfico de no predictibilidad para las",
      "coordenadas principales"),cex.main=1.2, col.main=4, ylim=c(0,0.35),
      xlim=c(0,10), type="l")

```

```

abline(v=2,lty=2,col="blue")

names(o)
o1<-o[o$CorrCuad>0.007,]
Xr <- X[,o1$ID]
rdb <- lm(y ~ Xr)
summary(rdb)

rdb1<-lm(y~dif.IDSTA3$X+dif.IDSTA3$Y+I(dif.IDSTA3$X^2)+I(dif.IDSTA3$Y^2)+
      dif.IDSTA3$Y*dif.IDSTA3$X)
summary(rdb1)
dif.IDSTA3$residuos <- rdb$residuals

rdb <- lm(y ~ X)
plot(density(rdb1$residuals), main="Gráfico de densidad de los residuos\n
      de la Regresión Clásica y DB")
lines(density(rdb$residuals), col=2)
exc <-list(paste("Clásico Orden 2 "),paste("Basado en Distancias"))
legend(locator(1),c(exc[[1]],exc[[2]]),title="Modelo Tendencia",text.col=
c(1,1),lty=c(1,1), pch=c(-1,-1), box.col=F, col=c(1,2),lwd=c(2,2),cex=0.8)

midataframe3<-data.frame(Xr,midataframe[,3])
names(midataframe3)<-c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10",
      "X11","X12","X13","X14","X15","TEMP")

# -----
# STEP 5: Ajuste Variogramas;
# -----

midataframe <- data.frame(dif.IDSTA3[,c(3,4,17)])
names(midataframe) <- c("x", "y", "r")

library(sgeostat)
# library(geospt) # No seria necesario cargar "sgeostat"
midataframe <- data.frame(dif.IDSTA3[,c(3,4,17)])
names(midataframe) <- c("x", "y", "r")
P.L.point <- point(midataframe)
P.L.pair <- pair(P.L.point,num.lags=40,maxdist=200000)
#El parametro trim se utiliza para calcular la media recortada
P.L.v <- est.variograms(P.L.point,P.L.pair,"r",trim=0.1)

```

```

#A continuación se grafica los semivariogramas experimentales
plot(P.L.v$bins,P.L.v$robust,lty=1, col =1,main = c("Ajuste de Modelos de
  Semivarianza", "Deriva DB"),xlab="Distancia",ylab="Semivarianza",type="l")
lines(P.L.v$bins,P.L.v$med, col=2)
lines(P.L.v$bins,P.L.v$classic, col=3)
lines(P.L.v$bins,P.L.v$med.trim, col=4)
legend(locator(1), c("Clásico", "Robusto", "Mediana", "Media Recortada"),
  col=c(1,2,3,4), lty=c(1,1,1,1))
# Se inactiva sgeostat, dado que genera conflicto con geoR en algunas
# funciones
detach("package:sgeostat")

# A continuación se ajustan algunos semivariogramas teóricos al
# semivariograma experimental
dir.hor <- seq(0, 0, length.out=40)
dir.ver <- seq(0, 0, length.out=40)
id <- seq (length.out=40)
id <- rep("var1",40)

#####
##ESFERICO CLASICO
#####

library(gstat)
y <- data.frame (P.L.v$n, P.L.v$bins, P.L.v$classic, dir.hor, dir.ver, id)
names(y) <- c("np", "dist", "gamma", "dir.hor","dir.ver","id")
class(y) <- c("variogram","gstatVariogram","data.frame")

Sph.wls <- fit.variogram(y, vgm(7.5,"Sph",100000,1.5,anis=c(p=45, s=0.5)),
  fit.method = 2) # metodo 2 MCP Nj/(G(hj))^2
Sph.reml <-fit.variogram.reml(TEMP~1,~x+y,midataframe2,model=vgm(7.5,"Sph",
  100000, 1.5, anis=c(p=45, s=0.5)))
Sph.ols <- fit.variogram(y,vgm(7.5,"Sph",100000,1.5,anis=c(p=45, s=0.5)),
  fit.method = 6) # metodo 6 MCO
Sph.wls1 <- fit.variogram(y, vgm(7.5,"Sph",100000,1.5,anis=c(p=45,s=0.5)),
  fit.method = 7) # metodo 7 MCP Nj/(hj)^2
print(list(Sph.wls,Sph.reml,Sph.ols, Sph.wls1))

dist.s <- P.L.v$bins
Sph.WLS <- variogramLine(vgm(Sph.wls$psill[2], "Sph", Sph.wls$range[2],

```

```

      Sph.wls$psill[1],anis=c(p=45, s=0.5)), min=0, dist_vector=dist.s)
Sph.RML <- variogramLine(vgm(Sph.reml$psill[2], "Sph", Sph.reml$range[2],
      Sph.reml$psill[1],anis=c(p=45, s=0.5)), min=0, dist_vector=dist.s)
Sph.WLS1 <- variogramLine(vgm(Sph.wls1$psill[2], "Sph", Sph.wls1$range[2],
      Sph.wls1$psill[1],anis=c(p=45, s=0.5)), min=0, dist_vector=dist.s)
Sph.OLS <- variogramLine(vgm(Sph.ols$psill[2], "Sph", Sph.ols$range[2],
      Sph.ols$psill[1],anis=c(p=45, s=0.5)), min=0, dist_vector=dist.s)
CME.Sph.WLS <- mean((P.L.v$classic-Sph.WLS$gamma)^2)
CME.Sph.RML <- mean((P.L.v$classic-Sph.RML$gamma)^2)
CME.Sph.OLS <- mean((P.L.v$classic-Sph.OLS$gamma)^2)
CME.Sph.WLS1 <- mean((P.L.v$classic-Sph.WLS1$gamma)^2)
print(data.frame(CME.Sph.WLS,CME.Sph.RML,CME.Sph.OLS,CME.Sph.WLS1))
plot(P.L.v$bins,P.L.v$classic,lty=2,pch=1,lwd=2,bg="yellow",type="p",col=1,
      font.main=3,main=("AJUSTE DE MODELO ESFERICO CLÁSICO DB"),
      xlab="Distancia",ylab="Semivarianza")
lines(Sph.WLS, col =2,lty=6,lwd=2)
lines(Sph.RML, col =3,lty=6,lwd=2)
lines(Sph.WLS1, col =4,lty=6,lwd=2)
lines(Sph.OLS, col =5,lty=6,lwd=2)
exc<-list(paste("CLÁSICO      "),paste("WLS =", round(CME.Sph.WLS,4)),
paste("RML =",round(CME.Sph.RML,4)), paste("WLS1=",round(CME.Sph.WLS1,4)),
paste("OLS      =",round(CME.Sph.OLS,4)))
legend(locator(1),c(exc[[1]],exc[[2]],exc[[3]],exc[[4]],exc[[5]]),title =
"Valores de CME",text.col=c(1,1,1,1,1),lty=c(-1,6,6,6,6),
pch=c(1,-1,-1,-1,-1),box.col=F,col=c(1,2,3,4,5),lwd=c(2,2,2,2,2),cex=0.8)

# Para los demás variogramas (Exponencial, Gaussiano y Matern) se trabajo de
# manera similar para el ajuste a los variogramas experimentales; clásico,
# robusto, mediana y media recortada

#####
# Mejores Modelos Exponencial Robusto por OLS (CME=0.7951) #
#           Matern Media recortada WLS (CME=0.7968) #
#####

# -----
# STEP 6: Calculo de las coordenadas principales de nuevos individuos;
# -----

# Grilla generada para un recuadro sin cortar al borde de Croacia

```

```

x <- dif.IDSTA3$X
y <- dif.IDSTA3$Y
puntos <- expand.grid(x=seq(min(x),max(x),1500), y=seq(min(y),max(y),1500))
plot(puntos)
xy = puntos[c("x", "y")]
coordinates(puntos) = xy
coordinates(puntos) <- c("x","y")
gridded(puntos) <- TRUE
tendencia <- Xr[,1:2]

x.new <- (1/2)*(diag(length(ValoresPropios[1:ncol(tendencia)])))*
        (ValoresPropios[1:ncol(tendencia)]^(-1))
%*%crossprod(tendencia,b - as.numeric(dist.So)^2)
d <- as.matrix(dist(c(x,y)))
d <- daisy(rbind(puntos[1:10,],preci[,2:3][1,]), metric ="euclidean")
dis <- function(x){
d1 <-as.matrix(dist(rbind(x,dif.IDSTA3[,3:4]))) [2:nrow(dif.IDSTA3[,3:4]),1]
d1
}
a <- dis(puntos[2,])
d <- apply(puntos,1,dis)

x.news <- function(ValoresPropios,coordenadas,tendencia,newdata){
So <- newdata
So <- as.data.frame(cbind(x=So[1],y=So[2]))
coordinates(So) <- c("x", "y")
s <- coordenadas
dist.So <- spDists(So,s)
b <- diag(tcrossprod(tendencia))
x.new<-(1/2)*(diag(length(ValoresPropios[1:ncol(tendencia)])))*(ValoresPropios
[1:ncol(tendencia)]^(-1))%*%crossprod(tendencia,b-as.numeric(dist.So)^2)
x.new
}

tabla1 <- function(newdata){
x.new<-x.news(ValoresPropios=ValoresPropios,coordenadas=
SpatialPoints(coordenadas), tendencia,newdata)
x.new
}

```

```

tabla2<-x.news(ValoresPropios,SpatialPoints(coordenadas),
              tendencia,df.pts[1,])
tabla <- apply(df.pts,1,tarla1)
tablaa <- data.frame(t(tabla))
names(tablaa)<-c("x","y")

library(maptools)
croatia.shp <- readShapePoly("D:/...../croatia.shp")
# Grilla generada dentro del borde de Croacia
pts <- spsample(croatia.shp, n=70000, type="regular")
df.pts <- as.matrix(pts)
names(df.pts) <- c("x", "y")

#tendencia1 <- tendencia
#names(tendencia1) <- c("x", "y")
#coordinates(tendencia1) <- c("x", "y")
pts<- data.frame(tendencia,dif.IDSTA3$TEMP)
names(pts)<-c("x","y","TEMP")
coordinates(pts) <- c("x", "y")
pts1<- data.frame(tendencia,dif.IDSTA3$residuos)
names(pts1)<-c("x","y","res")
coordinates(pts1) <- c("x", "y")
coordinates(tablaa) = c("x", "y")

# Los parámetros establecidos fueron los de mejor ajuste del variograma
cov.pars <- as.matrix(cbind(Exp.ols$psill,Exp.ols$range,Exp.ols$ang1,
                          Exp.ols$anis1))

# -----
# STEP 7: Generación de mapas de interpolación;
# -----

coordenadas <- dif.IDSTA3[,3:4]
names(coordenadas) <- c("x", "y")
coordinates(coordenadas) <- c("x", "y")

krige.u.bd <- function(var.reg,tendencia,ValoresPropios,coordenadas,newdata,
                      n.vec, modelo.cov,cov.pars){
z <- var.reg

```

```

So <- newdata
So <- as.data.frame(cbind(x=So[1],y=So[2]))
coordinates(So) <- c("x", "y")
s <- coordenadas
dist.So <- spDists(So,s)
remove("newdata","var.reg")
vec.orden <- order(dist.So)
dist.vec.cerca <- dist.So[vec.orden[1:n.vec]]
m.dist.vec <- spDists(s)[vec.orden[1:n.vec], vec.orden[1:n.vec]]
b <- diag(tcrossprod(tendencia))
x.new<-(1/2)*(diag(length(ValoresPropios[1:ncol(tendencia)]))*(ValoresPropios
  [1:ncol(tendencia)]^(-1)))%% crossprod(tendencia,b-as.numeric(dist.So)^2)
one = rep(1,length(dist.vec.cerca))
tend <- as.matrix(tendencia[vec.orden[1:n.vec],])
cov.0 <- cov.spatial(0, cov.model= modelo.cov, cov.pars =cov.pars)
m.cov<-cov.spatial(m.dist.vec, cov.model= modelo.cov, cov.pars=cov.pars)
v <- cov.spatial(dist.vec.cerca, cov.model= modelo.cov, cov.pars = cov.pars)
m.cov.k <- rbind(as.matrix(cbind(m.cov,one,tend)),as.matrix(cbind(rbind(one,
  t(tend)),matrix(0,ncol=(ncol(tend)+1),nrow=(ncol(tend)+1))))))
m.cov.k.i <- solve(m.cov.k)
v.ko <- c(v,1,x.new)
Pesos.ko <- m.cov.k.i%%v.ko
KUpr <- z[vec.orden[1:n.vec]]%%Pesos.ko[1:length(dist.vec.cerca)]
KUvr <- cov.0-v.ko%%Pesos.ko
remove("dist.So","vec.orden","dist.vec.cerca","m.dist.vec","b","x.new",
  "one","tend","cov.0","m.cov","v","m.cov.k","m.cov.k.i","v.ko","Pesos.ko",
  "ValoresPropios","tendencia","coordenadas","n.vec","modelo.cov","cov.pars",
  "s","So")
data.frame(KUpr,KUvr)
}
krige.u.bd1 <- function(newdata){
krige.u.bd(var.reg=dif.IDSTA3$TEMP, tendencia, ValoresPropios, coordenadas,
newdata=newdata, n.vec=10, modelo.cov="exponential", cov.pars)
}

So <- c(671155.2,5083499)
So = as.data.frame(cbind(x=So[1],y=So[2]))
#coordinates(So) <- c("x", "y")
krige.u.bd0 <- krige.u.bd1(newdata=puntos[1,]) # prueba funcionamiento

```



```

start <- Sys.time()
krige.u.bd2 <- apply(puntos,1,krige.u.bd1)
Sys.time() - start
uno <- as.data.frame(do.call("rbind",krige.u.bd2))
krige.u.bd3 <- data.frame(puntos, uno)
coordinates(krige.u.bd3) = c("x", "y")
names(krige.u.bd3)<-c("var1.pred","var1.var")

splot(krige.u.bd3["var1.pred"], cuts=10, col.regions=bpy.colors(100),
      main = "Interpolaciones de Kriging Universal\n Basado en Distancias de las
      Temperaturas", scales = list(draw =T), xlab="Este (m)",ylab = "Norte (m)",
      key.space=list(space="right", cex=0.6))

splot(krige.u.bd3["var1.var"], cuts=10, col.regions=bpy.colors(100),
      main "Interpolaciones de Kriging Universal Basado en\n Distancias de las
      Varianzas de la Temperatura", scales = list(draw =T), xlab="Este (m)",
      ylab = "Norte (m)", key.space=list(space="right", cex=0.6))

coordenadas <- Xr[,1:2]
pts<- data.frame(coordenadas,dif.IDSTA3$TEMP)
names(pts) <- c("x","y","TEMP")
coordinates(pts) <- c("x", "y")
# Kriging Universal Orden 1
krige.u1<-krige(TEMP~x+y,pts,SpatialPoints(puntos),vgm(cov.pars[2,1],"Exp",
  cov.pars[2,2],cov.pars[1,1],anis=c(cov.pars[2,3],cov.pars[2,4])),nmax=100)
# Kriging Universal Orden 2
krige.u2 <- krige(TEMP~x + y+ x*y +I(x^2)+I(y^2),pts,SpatialPoints(puntos),
  vgm(cov.pars[2,1],"Exp",cov.pars[2,2],cov.pars[1,1],anis=c(cov.pars[2,3],
  cov.pars[2,4])), nmax=100)
# Kriging Universal Orden 3
krige.u3 <- krige(z~x + y + x*y + I(x^2)+I(y^2), pts,SpatialPoints(puntos),
  vgm(cov.pars[2,1],"Exp",cov.pars[2,2],cov.pars[1,1],anis=c(cov.pars[2,3],
  cov.pars[2,4])), nmax=100)

gridded(krige.u2)<-TRUE
splot(krige.u1["var1.pred"], cuts=10, col.regions=bpy.colors(100), main =
  "Interpolaciones de Kriging Universal\n Orden 1 de la Temperatura",scales=
  list(draw=T),xlab="Este (m)",ylab="Norte (m)",key.space=list(space="right",
  cex=0.6))

```

```

spplot(krige.u1["var1.var"], cuts=10, col.regions=bpy.colors(100), main =
  "Interpolaciones de Kriging Universal\n Orden 1 de la Temperatura",scales=
  list(draw=T),xlab="Este (m)",ylab="Norte (m)",key.space=list(space="right",
  cex=0.6))

```

```

spplot(krige.u2["var1.pred"], cuts=20, col.regions=bpy.colors(100),main =
  "Interpolaciones de Kriging Universal\n Orden 2 de la Temperatura",scales=
  list(draw=T),xlab="Este (m)",ylab="Norte (m)",key.space=list(space="right",
  cex=0.6))

```

```

spplot(krige.u2["var1.var"], cuts=20, col.regions=bpy.colors(100), main =
  "Interpolaciones de Kriging Universal\n Orden 2 de la Temperatura",scales=
  list(draw=T),xlab="Este (m)",ylab="Norte (m)",key.space=list(space="right",
  cex=0.6))

```

```

# Igualmente se estima un modelo de variograma para la variable TEMP y luego
# con los parametros se construye cov.pars para asi generar con las
# coordenadas los mapas de interpolación de la variable estimada TEMP y su
# correspondiente varianza.

```

```

gridded(tablaa) <- TRUE
kud.MDTEMP<-krige(TEMP~x+y,pts,tablaa,vgm(cov.pars[2,1],"Exp",cov.pars[2,2],
  cov.pars[1,1],anis=c(cov.pars[2,3],cov.pars[2,4])),nmax=100)
kud.MDTEMP@coords <- coordinates(puntos)
gridded(kud.MDTEMP) <- TRUE
spplot(kud.MDTEMP["var1.pred"], cuts=20, col.regions=bpy.colors(100),main=
  "Interpolaciones de Kriging Universal \n DB de la Temperatura 2008-12-01",
  scales=list(draw=T),xlab="Este (m)",ylab="Norte (m)",key.space=list(space=
  "right", cex=0.6))

```

```

spplot(kud.MDTEMP["var1.var"],cuts=20,col.regions=bpy.colors(100), main =
  "Interpolaciones de Varianza Kriging Universal\n DB de la Temperatura
  2008-12-01",scales=list(draw=T), xlab="Este (m)", ylab="Norte (m)",
  key.space=list(space= "right", cex=0.6))

```

```

#-----
## B. Aplicación "ca20" ##
#-----

```

```
# -----  
# STEP 1: Preparación de la muestra, pixelado de puntos  
#         para los mapas y generación shapefile;  
# -----  
  
library(geoR)  
library(fields)  
library(cluster)  
library(minqa)  
data(ca20)  
  
### define 5 m grid, and select points within study area ###  
# pixelado de puntos:  
gr <- pred_grid(ca20$borders,by=5)  
gr0 <- polygrid(gr,borders=ca20$border,bound=T)  
  
# Construcción de vector de covariables para las ubicaciones de predicción  
# que indica la subárea:  
ind.reg.pts <- numeric(nrow(gr0))  
ind.reg.pts[.geoR_inout(gr0,ca20$reg1)] <- 1  
ind.reg.pts[.geoR_inout(gr0,ca20$reg2)] <- 2  
ind.reg.pts[.geoR_inout(gr0,ca20$reg3)] <- 3  
ind.reg.pts <- as.factor(ind.reg.pts)  
  
v2.pts <- ifelse(ind.reg.pts %in% 2,1,0)  
v3.pts <- ifelse(ind.reg.pts %in% 3,1,0)  
  
df.pts <- data.frame(gr0[,1],gr0[,2],v2.pts,v3.pts)  
names(df.pts) <- c("x","y","v2","v3")  
  
### construcción de vector de covariables para la muestra###  
ind.reg <- numeric(nrow(gr0))  
ind.reg[.geoR_inout(gr0,ca20$reg1)] <- 1  
ind.reg[.geoR_inout(gr0,ca20$reg2)] <- 2  
ind.reg[.geoR_inout(gr0,ca20$reg3)] <- 3  
ind.reg <- as.factor(ind.reg)  
  
v2 <- ifelse(ca20$covariate$area %in% 2,1,0)
```

```

v3 <- ifelse(ca20$covariate$area %in% 3,1,0)

Data.f <- data.frame(ca20$coords[,1],ca20$coords[,2],v2,v3,
                    ca20$covariate$altitud)
names(Data.f) <- c("x","y","v2","v3","altitud")
Data.f1 <- Data.f[,-5]
#-----
# Generación shapefile
library(shapefiles)
poligonos <- data.frame(c(rep(1,nrow(ca20$reg1)),rep(2,nrow(ca20$reg2)),
                          rep(3,nrow(ca20$reg3))),rbind(ca20$reg1,ca20$reg2,ca20$reg3))
names(poligonos) <- c("Id","x","y")
ddTable <- data.frame(Id=c(1,2,3),Name=c("reg1","reg2","reg3"))
ddShapefile <- convert.to.shapefile(poligonos, ddTable, "Id", 5)
write.shapefile(ddShapefile, "D:/.../ca20", arcgis=T)

#-----
#STEP 2: Calculo de coordenadas principales y regresión basada en distancias
#-----
# No se considera altitud, dado que no se dispone de esta variable en los
# puntos no muestreados

Delta1 <- daisy(Data.f[,-5], type=list(asymm=c("v2","v3")),metric ="gower")
class(Delta1)
library(ade4)
is.euclid(Delta1^(1/2))

mds0 <- cmdscale(Delta1^(1/2), k = nrow(Data.f[,-5])-1, eig = TRUE)
names(mds0)
round(mds0$points[,1],4)

m1 <- sum(mds0$eig > 0.007)
mds1 <- cmdscale(Delta1^0.5, k = m1, eig = TRUE)
Xa <- mds1$points

ValoresPropios1 <- mds1$eig

CorrCuadrado1 <- as.vector(cor(ca20$data,Xa)^2)
Porc.Inercia1 <- ValoresPropios1/length(CorrCuadrado1)
o1<-data.frame(1:length(CorrCuadrado1),round(ValoresPropios1[1:length

```

```

      (CorrCuadrado1]],10),round(CorrCuadrado1,10),round(Porc.Inercial[1:
      length(CorrCuadrado1]],10))
names(o1)<-c("ID", "ValoresProp1", "CorrCuad1", "Porc.Inercial")

names(o1)
o2<-o1[o1$CorrCuad1>0.007,]
Xr1 <- Xa[,o2$ID]
rdb1 <- lm(ca20$data ~ Xr1[,1:17])
model.db1 <- summary(rdb1)

aux1 <- CorrCuadrado1[1:17]*ValoresPropios1[1:17]/sum(CorrCuadrado1[1:17]*
      ValoresPropios1[1:17])
c.pred1 <- c(0,cumsum(aux))

plot(0:18,1-c.pred1,
      xlab="Number Principal Coordinates",
      ylab="1 - Predictability",
#   main=c("No Predictability Principal Coordinates"),
      cex.main=1.2, col.main=4, ylim=c(0,1), xlim=c(0,18), type="l")
abline(v=4,lty=2,col="blue")

df.ca20 <- data.frame(ca20)
names(df.ca20) <- c("x","y","ca20","altitude","area")
dXr1 <- data.frame(df.ca20[,1:3],Xr1)
ca20g1 <- as.geodata(dXr1, coords.col = 1:2, data.col = 3, covar.col=4:20)

# -----
# STEP 3: Ajuste del modelo de variograma;
# -----

# Ajuste variograma para variables sin transformar (caso clásico)

### Box-Cox transformation? ###
boxcox.fit(ca20$data)
hist(ca20$data)
#no transformation (lambda=1) seems appropriate
### semivariogram ###
plot(variog(ca20,max.dist=510)) # no covariates
t.all <- trend.spatial(ca20,~area + altitude,add="2nd")
t.all1 <- trend.spatial("2nd",ca20,add=~altitude)

```

```

ca20$area <- ca20$covariates$area
t.all1 <- trend.spatial(~coords+area,ca20)
# adjusting for area, altitude and quadratic spatial trend
plot(variogram(ca20,max.dist=510,trend=~t.all))
### linear geostatistical models: deciding on kappa ###
likfit(ca20, ini = c(10,200), nug=50,kappa=0.5) # largest likelihood
likfit(ca20, ini = c(10,200), nug=50,kappa=1.5)
likfit(ca20, ini = c(10,200), nug=50,kappa=2.5)
### linear geostatistical models ###
m1 <- likfit(ca20, cov.model="spherical", ini = c(10,200), nug=50)
m2 <- likfit(ca20, trend=~covariate$area, cov.model="spherical",ini=c(60,100),
            nug=40) # chosen model
m3 <- likfit(ca20,trend=~covariate$area+covariate$altitude,cov.model=
            "spherical", ini = c(60,100), nug=40)
m4 <- likfit(ca20, trend=~covariate$area + coords, cov.model="spherical",
            ini =c(60,100), nug=40)
m5 <- likfit(ca20, trend=~covariate$area + covariate$altitude + coords,
            cov.model="spherical", ini = c(60,100), nug=40)

m11.2logL <- 2*m1$loglik
m21.2logL <- 2*m2$loglik
m31.2logL <- 2*m3$loglik
m41.2logL <- 2*m4$loglik
m51.2logL <- 2*m5$loglik
print(c(m11.2logL,m21.2logL,m31.2logL,m41.2logL,m51.2logL))

# Ajuste variograma para variables transformadas (es decir, coordenadas
# principales) caso DB

dXr1 <- data.frame(df.ca20[,1:3],Xr1)
ca20g1 <- as.geodata(dXr1, coords.col = 1:2, data.col = 3, covar.col=4:20)

m1.db1 <- likfit(ca20g1,ini = c(10,200), cov.model= "spherical", nug=50)
m2.db1 <- likfit(ca20g1,trend=~V1+V2,cov.model="spherical",ini=c(60,100),
            nug=40)
m3.db1<-likfit(ca20g1,trend=~V1+V2+V3+V4,cov.model="spherical",ini=c(60,100),
            nug=40)
m4.db1 <-likfit(ca20g1,trend=~V1+V2+V3+V4+V5+V6+V7+V8,cov.model="spherical",
            ini = c(60,100), nug=40)
m5.db1 <-likfit(ca20g1,trend=~V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13+

```

```

V14+V15+V16+V17, cov.model= "spherical", ini = c(60,100), nug=40)

m1db1.2logL <- 2*m1.db1$loglik
m2db1.2logL <- 2*m2.db1$loglik
m3db1.2logL <- 2*m3.db1$loglik
m4db1.2logL <- 2*m4.db1$loglik
m5db1.2logL <- 2*m5.db1$loglik
print(c(m1db1.2logL,m2db1.2logL,m3db1.2logL,m4db1.2logL,m5db1.2logL))

# -----
# STEP 4: Calculo de las coordenadas principales de nuevos individuos;
# -----
# Nuevos individuos ubicados sobre el pixelado del mapa

x.news <- function(ValoresPropios1,Data,tendencia,newdata,id){
Data[length(ValoresPropios1)+1, ] <- newdata
d <- as.matrix(daisy(Data, type=list(asymm=c("v2","v3")),metric="gower"))
  [-(length(ValoresPropios1)+1), (length(ValoresPropios1)+1)]
b <- diag(tcrossprod(tendencia))
x.new <- (1/2)*diag(ValoresPropios1[1:ncol(tendencia)]^(-1))%*% t(tendencia)
  %*%(b-d)
x.new[id,]
}

tabla <- function(newdata){
x.new0<-x.news(ValoresPropios1=ValoresPropios1,Data=Data.f1,tendencia=Xa,
  newdata,id=(o1[o1$CorrCuad1>0.007,][,1]))
x.new0
}

tabla1 <- apply(df.pts,1,tblaa)
tblaa <- data.frame(t(tabla1))
tblaa1 <- tblaa
names(tblaa1) <- c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10",
  "X11","X12","X13","X14","X15","X16","X17")

coord.cp <- data.frame(df.pts[,1:2], tblaa1)
newdata <- coord.cp

# -----

```





```

spplot(mpred1["ca20.var"], cuts=60, col.regions=terrain.colors(100), scales=
list(draw=T), xlab="East(m)", ylab="North(m)", key.space=list(space="right",
cex=0.8))
par(mfrow=c(1,1), mar=c(5,5,4,4))
image(data.frame(df.pts[,1:2]), mpred1@data$ca20.var), zlim=
  c(min(mpred1@data$ca20.var), max(mpred1@data$ca20.var)), col=
  terrain.colors(100), axes=T, xlab="East (m)", ylab="North (m)", cex=0.6)
polygon(ca20$reg1)
polygon(ca20$reg2)
polygon(ca20$reg3)
image.plot(legend.only=TRUE, legend.width=1.3, legend.mar=3.5, legend.shrink=1,
  col=terrain.colors(100), zlim=c(-0.7, 11.7))

# -----
# Predicciones con el variograma obtenido en el caso DB

vuk.db1<-vgm(m5.db1$sigmasq, "Sph", m5.db1$phi, m5.db1$tausq, kappa=m5.db1$kappa)
g1.db1 <- gstat(id="ca20", formula= ca20~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+
  X12+X13+X14+X15+X16+X17, locations=~x+y, model=vgm(m5.db1$sigmasq, "Sph",
  m5.db1$phi, m5.db1$tausq, kappa=m5.db1$kappa), data = dXr1)
tablaa1 <- tablaa
names(tablaa1) <- c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9", "X10", "X11",
  "X12", "X13", "X14", "X15", "X16", "X17")
trendb1 <- data.frame(tablaa1, df.pts[,1:2])
mpredb1 <- predict.gstat(g1.db1, trendb1)
kcv.ca20db1 <- gstat.cv(g1.db1, nfold=178)
RMSPE.UKdb1<-sqrt(sum(kcv.ca20db1$residual^2)/length(kcv.ca20db1$residual))
resid.mean1 <- dXr1$ca20- mean(dXr1$ca20)
# Valoración método:
R2db1 <- 1- sum((kcv.ca20db1$residual)^2)/sum(resid.mean1^2)

# Mapas de predicciones de ca20 y varianzas de errores
str(mpredb1)
gridded(mpredb1) = ~x+y
par(mfrow=c(1,1), mar=c(5,5,4,4))
image(data.frame(df.pts[,1:2]), mpredb1@data$ca20.pred), zlim=
  c(min(mpredb1@data$ca20.pred), max(mpredb1@data$ca20.pred)),
  col=terrain.colors(100), axes=T, xlab="East (m)", ylab="North (m)", cex=0.6)
#, xlim=c(4900, 6000), ylim = c(4800, 5800))

```

```

polygon(ca20$reg1)
polygon(ca20$reg2)
polygon(ca20$reg3)
image.plot(legend.only=TRUE,legend.width=1.3,legend.mar=3.5,legend.shrink=1,
           col=terrain.colors(100), zlim=c(15,81))

par(mfrow=c(1,1), mar=c(5,5,4,4))
image(data.frame(df.pts[,1:2],mpredb1@data$ca20.var), zlim=
c(min(mpredb2@data$ca20.var),max(mpredb2@data$ca20.var)), col=
terrain.colors(100),axes=T, xlab="East (m)", ylab = "North (m)", cex=0.6)
polygon(ca20$reg1)
polygon(ca20$reg2)
polygon(ca20$reg3)
library(fields)
image.plot(legend.only=TRUE,legend.width=1.3,legend.mar=3.5,legend.shrink=1,
           col=terrain.colors(100), zlim=c(-0.3,9.5))

```

## A.4 Programación Capítulo 4

```

-----
#####      CAPITULO 4: SIMULACIÓN FBR ESPACIAL      #####
-----

```

```

library(gstat)
library(geoR)
library(cluster)
library(maptools)
library(rgdal)
library(outliers)

p1 <- cbind( c(0,1,1,0), c(1,1,0.5,1))
p2 <- cbind( c(0,1,0,0), c(1,0.5,0.3,1))
p3 <- cbind( c(0,0,1,1,0), c(0,0.3,0.5,0,0))

poly1 <- Polygons(list(Polygon(p1)), "R1")
poly2 <- Polygons(list(Polygon(p2)), "R2")
poly3 <- Polygons(list(Polygon(p3)), "R3")

```

```

sppo <- SpatialPolygons(list(poly1,poly2,poly3))

#####
###      CONSTRUCCION GRAFICO REGIONES      ###
#####

par(mfrow=c(1,1), mar=c(5,5,4,4))
plot(sppo, ylim=c(0,1), xlab="x", ylab="y")
marcas <- seq(0,1,0.1)
#box(lty = 'solid', col = 'blue', lwd=2, mar=c(5,5,5,5))
axis(side=2, at=marcas)
axis(side=1, at=marcas, xlab = "x")
names <- unlist(lapply(slot(sppo,"polygons"), function(x) slot(x,"ID")))
text(coordinates(sppo), labels=names)#, cex=0.6)
text(0.5,-0.1, "X", cex=1)
title(xlab="X")
title(ylab="Y")

# Tamaños muestrales
Muestra1 <- c(13,17,20)      # Regiones 1,2 y 3 respectivamente # 50 datos
Muestra2 <- c(25,35,40)      # 100 datos
Muestra3 <- c(28,52,60)      # 150 datos

#####
###      PROGRAMA SIMULACION EN LAS REGIONES Y VARIABLE BINOMIAL GRF      ###
#####

# Funciones escenarios

phi.u <- (1/sqrt(2*pi))*exp((-2^2)/2)
# dnorm(2, mean = 0, sd = 1, log = FALSE)

fx <- function(x){
fx <- 1.5*dnorm((x-0.35)/0.15)-dnorm((x-0.8)/0.04)
fx
}

fxy <- function(x,y){
fxy <- (1.5*dnorm((x-0.35)/0.15)-dnorm((x-0.8)/0.04))*(1.5*
dnorm((y-0.35)/0.15)-dnorm((y-0.8)/0.04))

```

```

fxy
}

sigma.j <- function(j){
s.j <- 0.02 + 0.04*(j-1)^2
s.j
}

fjx <- function(j,x){
fjx <- sqrt(x*(1-x))*sin((2*pi*(1+2^((9-4*j)/5)))/(x+2^((9-4*j)/5)))
fjx
}

fjxy <- function(j,x,y){
fjxy <- (sqrt(x*(1-x))*sin((2*pi*(1+2^((9-4*j)/5)))/(x+2^((9-4*j)/5))))*
(sqrt(y*(1-y))*sin((2*pi*(1+2^((9-4*j)/5)))/(y+2^((9-4*j)/5))))
fjxy
}

vjx <- function(j,x){
vjx <- (0.15*(1+0.4*(2*j-7)*(x-0.5)))^2
vjx
}

vjxy <- function(j,x,y){
vjxy <- ((0.15*(1+0.4*(2*j-7)*(x-0.5)))^2)*((0.15*(1+0.4*(2*j-7)*(y-0.5)))^2)
vjxy
}

bjx <- function(j,x){
bjx <- dbeta(x,(j+4)/5,(11-j)/5)
bjx
}

-----
##### VALORES PARAMETROS #####

B0 <- 10; B1<- -4; B2 <- 2; B3 <- -4; p <- 0.4

nuggeto <- 0.1

```

```

rango1o <- 0.5
sillo <- 1
ango <- 0
aniso <- 1
kappa1o <- 0

psillo <- c(nuggeto,sillo)
rangeo <- c(0,rango1o)
ang1o <- c(0,ango)
anis1o <- c(1,aniso)
kappao <- c(0,kappa1o)
cov.pars0o <- as.matrix(cbind(psillo,rangeo,kappao,ang1o,anis1o))
cov.modelo <- "matern"
modelo.covo <- "Mat"
lambdao <- 1

# Se cargan las funciones a utilizar:
source("D:/.../function/rbf.t.r")
source("D:/.../function/RBF.phi.r")
source("D:/.../function/rbf.t.tcv.r")

-----
#####          Factor: Generic Form          #####

# N:          Tamaños de muestra 50,100 y 150 datos
# nsim:       número de simulaciones, en general 100
# eta:        parámetro de suavizamiento en la optimización de la función de
#             base radial (FBR), 0.01 y 0.1
# n.neigh:    número de vecinos para la validacion cruzada, en este caso el
#             número de nodos en el spline 8 y 32
# func:       tipo de FBR (ST, M, TPS, EXP, y CRS)
# j:          parámetro asociado a la variacion espacial, 1 y 3

rbf.sim <- function(N, nsim, sigma, n.neigh, func, j){

set.seed(127)
sim <- grf(N,nsim=nsim,grid="irreg",cov.model=cov.modelo,xlims=c(0, 1),
          lambda=lambdao,

```

```

ylims = c(0, 1), cov.pars=cov.pars0o)
# aniso.pars=c(pi/4, 2). Esto es lo mismo que dejar cov.pars0 ó
# cov.pars1 y esta línea
set.seed(127)
V1 <- rbinom(N, size=1, prob=p)

#build nominal covariate vector for prediction locations indicating subarea#
ind.reg.sim <- numeric(nrow(sim$coords))
ind.reg.sim[.geoR_inout(sim$coords,p1)] <- 1
ind.reg.sim[.geoR_inout(sim$coords,p2)] <- 2
ind.reg.sim[.geoR_inout(sim$coords,p3)] <- 3
ind.reg.sim <- as.factor(ind.reg.sim)

V2 <- ifelse(ind.reg.sim %in% 2,1,0)
V3 <- ifelse(ind.reg.sim %in% 3,1,0)

# Noise Level:
# sim$data <- B0 + B1*V1 + B2*V2*sim$coords[,1] + B3*V3*sim$coords[,2] +
# fx(sim$coords[,1])+fx(sim$coords[,2])+fxy(sim$coords[,1],sim$coords[,2])+
# sigma.j(j)*sim$data
# Design density:
# sim$data <- B0 + B1*V1 + B2*V2*sim$coords[,1] + B3*V3*sim$coords[,2] +
# fx(bjx(j,sim$coords[,1]))+fx(bjx(j,sim$coords[,2]))+fxy(bjx(j,
# sim$coords[,1]), bjx(j,sim$coords[,2])) + 0.1*sim$data
# Spatial variation:
sim$data <- B0 + B1*V1 + B2*V2*sim$coords[,1] + B3*V3*sim$coords[,2] +
  fjx(j,sim$coords[,1]) + fjx(j,sim$coords[,2]) + fjxy(j,sim$coords[,1],
  sim$coords[,2]) + 0.2*sim$data

# Variance function:
# sim$data <- B0 + B1*V1 + B2*V2*sim$coords[,1] + B3*V3*sim$coords[,2] +
# fx(sim$coords[,1])+fx(sim$coords[,2]) +fxy(sim$coords[,1],sim$coords[,2])+
# (sqrt(vjx(j,sim$coords[,1])+vjx(j,sim$coords[,2])+vjxy(j,sim$coords[,1],
# sim$coords[,2])))*sim$data

tabla.sim <- as.data.frame(matrix(NA,nrow=nsim,ncol=5))
colnames(tabla.sim) <- c("sim", "MPE", "RMSPE", "R2", "No.CP")

Data.sim <- data.frame(sim$coords[,1],sim$coords[,2],V1,V2,V3)
names(Data.sim) <- c("x","y","V1","V2","V3")

```

```
#####
##### CONSTRUCCION COORDENADAS PRINCIPALES Y FBR DB #####
#####

Delt <- daisy(Data.sim,type=list(asymm=c("V1","V2","V3")),metric="gower")
class(Delt)
library(ade4)
is.euclid(Delt^(1/2))

mds <- cmdscale(Delt^(1/2), k = nrow(Data.sim)-1, eig = TRUE)
names(mds)
round(mds$points[,1],4)

m <- sum(mds$eig > 0.007)
mds <- cmdscale(Delt^0.5, k = m, eig = TRUE)
X <- mds$points

for(i in 1:100){

ValoresPropios <- mds$eig
CorrCuadrado <- as.vector(cor(sim$data[,i],X)^2)
Porc.Inercia <- ValoresPropios/length(CorrCuadrado)

o<-data.frame(1:length(CorrCuadrado),round(ValoresPropios[1:
length(CorrCuadrado)],10),round(CorrCuadrado,10),round(Porc.Inercia
[1:length(CorrCuadrado)],10))
names(o)<-c("ID", "ValoresProp", "CorrCuad", "Porc.Inercia")

names(o)
o1<-o[o$CorrCuad>0.007,]
Xr.sim <- X[,o1$ID]
rdb.sim <- lm(sim$data[,i] ~ Xr.sim)
model.db.sim <- summary(rdb.sim)

# x <- sim$coords[,1]
# y <- sim$coords[,2]

try({spdb.sim<-rbf.t.tcv(sigma=sigma,z=sim$data[,i], coordinates=sim$coords,
trend=Xr.sim, n.neigh=n.neigh, func=func)})
```

```

MPE          <- mean(spdb.sim$residual)
RMSPE        <- sqrt(sum(spdb.sim$residual^2)/length(spdb.sim$residual))
resid.mean.sim <- spdb.sim$observed - mean(spdb.sim$observed)
R2           <- 1- sum((spdb.sim$residual)^2)/sum(resid.mean.sim^2)

tabla.sim[i,] <- c(i,MPE,RMSPE,R2,ncol(Xr.sim))
}
as.matrix(tabla.sim)
}

# Se debe activar previamente en la simulación el escenario de interés

#####
#####      Factor: Noise Level      #####
#####

Resultado.simn1<-array(data=NA,c(100,5,5,2,2,3,2),dimnames=list(1:100,c("sim"
, "MPE", "RMSPE", "R2", "No. CP"),c("M", "TPS", "CRS", "ST", "EXP"),c("1", "3"),
c("0.01", "0.1"), c("50", "100", "150"), c("8", "32")))
funcion <- c("M", "TPS", "CRS", "ST", "EXPON")
for (func in funcion)
  for (j in c(1,3))      # 4 si, 6 no
    for (sigma in c(0.01,0.1))
      for (N in c(50,100,150))
        for (n.neigh in c(8,32))
Resultado.simn1[, ,func,paste(j,sep=""),paste(sigma,sep=""),paste(N,sep=""),
paste(n.neigh,sep="")] <- rbf.sim(N=N, nsim=100, sigma=sigma, n.neigh=
n.neigh,func=func, j=j)

#####
#####      Factor: Design density      #####
#####

Resultado.simbd<-array(data=NA,c(100,5,5,2,2,3,2),dimnames=list(1:100,c("sim"
, "MPE", "RMSPE", "R2", "No. CP"),c("M", "TPS", "CRS", "ST", "EXP"),c("1", "3"),
c("0.01", "0.1"), c("50", "100", "150"), c("8", "32")))
funcion <- c("M", "TPS", "CRS", "ST", "EXPON")
for (func in funcion)
  for (j in c(1,3))      # 4 si, 6 no

```



```

        for (sigma in c(0.01,0.1))
          for (N in c(50,100,150))
            for (n.neigh in c(8,32))
Resultado.simbd[,func,paste(j,sep=""),paste(sigma,sep=""),paste(N,sep=""),
  paste(n.neigh,sep="")] <- rbf.sim(N=N, nsim=100, sigma=sigma, n.neigh=
  n.neigh,func=func, j=j)

#####
#####      Factor: Spatial variation      #####
#####

Resultado.simsv<-array(data=NA,c(100,5,5,2,2,3,2),dimnames=list(1:100,c("sim"
  ,"MPE","RMSPE","R2","No.CP"),c("M","TPS","CRS","ST","EXP"),c("1","3"),
  c("0.01","0.1"), c("50","100","150"), c("8","32")))
funci <- c("M","TPS","CRS","ST","EXPON")
for (func in funci)
  for (j in c(1,3))
    for (sigma in c(0.01,0.1))
      for (N in c(50,100,150))
        for (n.neigh in c(8,32))
Resultado.simsv[,func,paste(j,sep=""),paste(sigma,sep=""),paste(N,sep=""),
  paste(n.neigh,sep="")] <- rbf.sim(N=N, nsim=100, sigma=sigma, n.neigh=
  n.neigh,func=func, j=j)

#####
#####      Factor: Variance function      #####
#####

Resultado.simvf<-array(data=NA,c(100,5,5,2,2,3,2),dimnames=list(1:100,c("sim"
  ,"MPE","RMSPE","R2","No.CP"),c("M","TPS","CRS","ST","EXP"),c("1","3"),
  c("0.01","0.1"),c("50","100","150"), c("8","32")))
funci <- c("M","TPS","CRS","ST","EXPON")
for (func in funci)
  for (j in c(1,3))
    for (sigma in c(0.01,0.1))
      for (N in c(50,100,150))
        for (n.neigh in c(8,32))
Resultado.simvf[,func,paste(j,sep=""),paste(sigma,sep=""),paste(N,sep=""),
  paste(n.neigh,sep="")] <- rbf.sim(N=N, nsim=100, sigma=sigma, n.neigh=
  n.neigh,func=func, j=j)

```

```

-----
##### CAPITULO 4: APLICACIÓN FBR ESPACIAL "ca20" #####
-----

library(geoR)
library(fields)
library(cluster)
library(minqa)
data(ca20)

# Aquí, los pasos 1, 2 y 4 tienen el mismo código presentado anteriormente
# en la segunda aplicación del Capítulo 3, para la base de datos "ca20"

# -----
# STEP 3: Optimización parámetros ETA y RHO en las diferentes FBR;
# -----
# FBR: Funciones de base radial

#-----
# A: Optimización parámetro ETA

# Optimización "ETA" Multicuadrática Inversa
IM.o <- optimize(rbf.t.cv,c(0,30),z=ca20g1$data, coordinates=ca20g1$coords,
                trend=dXr1[,4:20], n.neigh=177, func="IM")
cat("Parámetro Óptimo FBR: ", "\n", "ETA = ", IM.o$minimum, "\n", "RMSPE= ",
    IM.o$objective, "\n")
# Parámetro Óptimo FBR:          Con 177 vecinos
# ETA      = 27.35744
# RMSPE    = 7.297001

# Optimización "ETA" Multicuadrática
M.o <- optimize(rbf.t.cv, c(0,1), z=ca20g1$data, coordinates=ca20g1$coords,
                trend=dXr1[,4:20], n.neigh=177, func="M")
cat("Parámetro Óptimo FBR: ", "\n", "ETA = ", M.o$minimum, "\n", "RMSPE=",
    M.o$objective, "\n")
# Parámetro Óptimo FBR:          Con 177 vecinos
# ETA      = 5.575865e-05
# RMSPE    = 7.375285

```

```
# Parámetro Óptimo FBR:          Con 40 vecinos
# P          = 5.575865e-05
# RMSPE     = 8.931973

# Optimization "ETA": Completely regularized spline function,
CRS.o<-optimize(rbf.t.cv,c(0,1),z=ca20g1$data,coordinates=ca20g1$coords,
               trend=dXr1[,4:20], n.neigh=177, func="CRS")
cat("Parámetro Óptimo FBR: ", "\n", "ETA =",CRS.o$minimum, "\n","RMSPE=",
    CRS.o$objective, "\n")
# Parámetro Óptimo FBR:
# ETA       = 0.1012015
# RMSPE     = 7.333849

# Optimization "ETA": Spline with Tension function
ST.o <- optimize(rbf.t.cv, c(0,1), z=ca20g1$data, coordinates=ca20g1$coords,
                trend=dXr1[,4:20], n.neigh=177, func="ST")
cat("Parámetro Óptimo FBR: ", "\n", "ETA= ", ST.o$minimum, "\n", "RMSPE=",
    ST.o$objective, "\n")
# Parámetro Óptimo FBR:
# P         = 0.08303821
# RMSPE     = 7.344243

# Optimizacion "ETA" Thin Plate Spline
TPS.o <- optimize(rbf.t.cv, c(0,0.81), z=ca20g1$data, coordinates=
                 ca20g1$coords,trend=dXr1[,4:20],n.neigh=177,func="TPS")
cat("Parámetro Óptimo FBR: ", "\n", "ETA= ",TPS.o$minimum, "\n", "RMSPE=",
    TPS.o$objective, "\n")
# Parámetro Óptimo FBR:
# ETA      = 0.8099541
# RMSPE    = 7.619761

# Optimizacion "ETA" Exponential RBF
EXP.o <- optimize(rbf.t.cv,c(0,1),z=ca20g1$data,coordinates=ca20g1$coords,
                 trend=dXr1[,4:20], n.neigh=177, func="EXPON")
cat("Parámetro Óptimo FBR: ", "\n", "ETA =", EXP.o$minimum, "\n","RMSPE=",
    EXP.o$objective, "\n")
# Parámetro Óptimo FBR:
# ETA      = 0.02450879
# RMSPE    = 7.317443
```

```

# Optimizacion "ETA" Gaussian RBF, da lo mismo con cualquier eta
GAU.o <-optimize(rbf.t.cv,c(0,0.2),z=ca20g1$data, coordinates=ca20g1$coords,
                trend=dXr1[,4:20],n.neigh=177,func="GAU")
cat("Parámetro Óptimo FBR: ", "\n","ETA= ",GAU.o$minimum, "\n", "RMSPE=",
    GAU.o$objective, "\n")
# Parámetro Óptimo FBR:
# ETA      = 0.2
# RMSPE    = 8.01272

#-----
# B: Optimización parámetro RHO

P.opt.crs <- bobyqa(c(1, 2), rbf.t.cvop, lower = c(0, 0), upper = c(2, 4),
                  z=ca20g1$data,coordinates=ca20g1$coords,trend=dXr1[,4:20],n.neigh=177,
                  func="CRS")
#parameter estimates: 0.329865648310229, 2.38278578402578
#objective: 7.3335085291925
#number of function evaluations: 31

P.opt.st <- bobyqa(c(1, 2), rbf.t.cvop, lower = c(0, 0), upper = c(2, 4),
                  z=ca20g1$data,coordinates=ca20g1$coords,trend=dXr1[,4:20],n.neigh=177,
                  func="ST")
#parameter estimates: 0.578017806545431, 2.04092020273964
#objective: 7.33350852919374
#number of function evaluations: 34

P.opt.exp <- bobyqa(c(0, 1), rbf.t.cvop, lower = c(0, 0), upper = c(1, 4),
                  z=ca20g1$data,coordinates=ca20g1$coords,trend=dXr1[,4:20],n.neigh=177,
                  func="EXP")
#parameter estimates: 0.0245091649898321, 0
#objective: 7.3174428056804
#number of function evaluations: 234

P.opt.gau <- bobyqa(c(0, 1), rbf.t.cvop, lower = c(0, 0), upper = c(1, 4),
                  z=ca20g1$data,coordinates=ca20g1$coords,trend=dXr1[,4:20],n.neigh=177,
                  func="GAU")
#parameter estimates: 0.2, 1
#objective: 8.01272045781122
#number of function evaluations: 28

```

```

P.opt.m <- bobyqa(c(0, 1), rbf.t.cvop, lower = c(0, 0), upper = c(4, 4),
  z=ca20g1$data, coordinates=ca20g1$coords, trend=dXr1[,4:20], n.neigh=177,
  func="M")
#parameter estimates: 0, 0
#objective: 7.37528496811618
#number of function evaluations: 19

P.opt.mi <- bobyqa(c(10, 1), rbf.t.cvop, lower = c(0, 0), upper = c(200, 4),
  z=ca20g1$data, coordinates=ca20g1$coords, trend=dXr1[,4:20], n.neigh=177,
  func="IM")
#parameter estimates: 27.3574363334684, 0
#objective: 7.29700064336846
#number of function evaluations: 45

P.opt.tps <- bobyqa(c(0.01, 1), rbf.t.cvop, lower = c(0, 0), upper = c(4, 4),
  z=ca20g1$data, coordinates=ca20g1$coords, trend=dXr1[,4:20], n.neigh=177,
  func="TPS")
#parameter estimates: 0.198730051128008, 1.24431018941423
#objective: 7.6104081199563
#number of function evaluations: 23

# -----
# STEP 5: Predicciones y generación de mapas de interpolación;
# -----

# Generacion Predicciones IM
IM.p <- rbf.t(eta=27.357, z=ca20g1$data, coordinates=ca20g1$coords,
  trend=dXr1[,4:20], nd.trend=newdata, n.neigh=177, func="IM")
gridded(IM.p) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(IM.p["var1.pred"], col.regions=terrain.colors(100), cuts=60, scales=
  list(draw =T), xlab="East (m)", ylab = "North (m)",
  key.space=list(space="right", cex=0.8), panel = function(...) {
  panel.gridplot(...)
  sp.polygons(ca20shp1,lwd=1.5)
  sp.polygons(ca20shp2,lwd=1.5)
  sp.polygons(ca20shp3,lwd=1.5)
  })

```

```
# Generacion Predicciones M
M.p <- rbf.t(eta=5.575865e-05, z=ca20g1$data, coordinates=ca20g1$coords,
            trend=dXr1[,4:20], nd.trend=newdata, n.neigh=177, func="M")
gridded(M.p) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(M.p["var1.pred"], col.regions=terrain.colors(100), cuts=60,scales=
       list(draw =T), xlab="East (m)", ylab = "North (m)",
       key.space=list(space="right", cex=0.8), panel = function(...) {
         panel.gridplot(...)
         sp.polygons(ca20shp1,lwd=1.5)
         sp.polygons(ca20shp2,lwd=1.5)
         sp.polygons(ca20shp3,lwd=1.5)
       })

# Generacion Predicciones CRS
CRS.p <- rbf.t(eta=0.1012015, z=ca20g1$data, coordinates=ca20g1$coords,
              trend=dXr1[,4:20], nd.trend=newdata, n.neigh=177,func="CRS")
gridded(CRS.p) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(CRS.p["var1.pred"], col.regions=terrain.colors(100), cuts=60,
       scales = list(draw =T), xlab="East (m)", ylab = "North (m)",
       key.space=list(space="right", cex=0.8), panel = function(...) {
         panel.gridplot(...)
         sp.polygons(ca20shp1,lwd=1.5)
         sp.polygons(ca20shp2,lwd=1.5)
         sp.polygons(ca20shp3,lwd=1.5)
       })

# Generacion Predicciones TPS
TPS.p <- rbf.t(eta=0.8099541, z=ca20g1$data, coordinates=ca20g1$coords,
              trend=dXr1[,4:20], nd.trend=newdata, n.neigh=177,func="TPS")
gridded(TPS.p) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(TPS.p["var1.pred"], col.regions=terrain.colors(100), cuts=60,
       scales = list(draw =T), xlab="East (m)", ylab = "North (m)",
       key.space=list(space="right", cex=0.8), panel = function(...) {
         panel.gridplot(...)
         sp.polygons(ca20shp1,lwd=1.5)
         sp.polygons(ca20shp2,lwd=1.5)
         sp.polygons(ca20shp3,lwd=1.5)
       })
```

```

    })

# Generacion Predicciones ST
ST.p <- rbf.t(eta=0.08303821, z=ca20g1$data, coordinates=ca20g1$coords,
            trend=dXr1[,4:20], nd.trend=newdata, n.neigh=177, func="ST")
gridded(ST.p) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(ST.p["var1.pred"], col.regions=terrain.colors(100), cuts=60,
       scales = list(draw =T), xlab="East (m)", ylab = "North (m)",
       key.space=list(space="right", cex=0.8), panel = function(...) {
         panel.gridplot(...)
         sp.polygons(ca20shp1,lwd=1.5)
         sp.polygons(ca20shp2,lwd=1.5)
         sp.polygons(ca20shp3,lwd=1.5)
       })

# Generacion Predicciones EXP
EXP.p <- rbf.t(eta=0.02450879, z=ca20g1$data, coordinates=ca20g1$coords,
            trend=dXr1[,4:20], nd.trend=newdata, n.neigh=177, func="EXP")
gridded(EXP.p) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(EXP.p["var1.pred"], col.regions=terrain.colors(100), cuts=60,
       scales = list(draw =T), xlab="East (m)", ylab = "North (m)",
       key.space=list(space="right", cex=0.8), panel = function(...) {
         panel.gridplot(...)
         sp.polygons(ca20shp1,lwd=1.5)
         sp.polygons(ca20shp2,lwd=1.5)
         sp.polygons(ca20shp3,lwd=1.5)
       })

# Generacion Predicciones GAU
GAU.p <- rbf.t(eta=0.2, z=ca20g1$data, coordinates=ca20g1$coords,
            trend=dXr1[,4:20], nd.trend=newdata, n.neigh=177,func="GAU")
gridded(GAU.p) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(GAU.p["var1.pred"], col.regions=terrain.colors(100), cuts=60,
       scales = list(draw =T), xlab="East (m)", ylab = "North (m)",
       key.space=list(space="right", cex=0.8), panel = function(...) {
         panel.gridplot(...)
         sp.polygons(ca20shp1,lwd=1.5)

```

```

sp.polygons(ca20shp2,lwd=1.5)
sp.polygons(ca20shp3,lwd=1.5)
})

```

## A.5 Programación Capítulo 5

```

-----
##### CAPITULO 5: SIMULACIÓN SPACE-TIME FBR #####
-----

```

```

library(gstat)
library(geoR)
library(RandomFields)
library(cluster)
#library(FD)
library(maptools)
library(rgdal)
library(outliers)

```

```

#####
### CONSTRUCCION GRAFICO REGIONES ###
#####

```

```

p1 <- cbind( c(0,1,1,0), c(1,1,0.5,1))
p2 <- cbind( c(0,1,0,0), c(1,0.5,0.3,1))
p3 <- cbind( c(0,0,1,1,0), c(0,0.3,0.5,0,0))

poly1 <- Polygons(list(Polygon(p1)), "R1")
poly2 <- Polygons(list(Polygon(p2)), "R2")
poly3 <- Polygons(list(Polygon(p3)), "R3")
sppo <- SpatialPolygons(list(poly1,poly2,poly3))

x1 <- rep(1:5, c(5,5,5,5,5))/5 -0.1
y1 <- rep(y,5)
plot(x1,y1, cex=0.5)

```



```

t <- (1:10)/10

par(mfrow=c(1,1), mar=c(5,5,4,4))
plot(sppo, ylim=c(0,1), xlab="x", ylab="y")
marcas <- seq(0,1,0.1)
axis(side=2, at=marcas)
axis(side=1, at=marcas, xlab = "x")
names <- unlist(lapply(slot(sppo,"polygons"), function(x) slot(x,"ID")))
text(coordinates(sppo), labels=names)#, cex=0.6)
text(0.5,-0.1, "X", cex=1)
title(xlab="X")
title(ylab="Y")
points(x1,y1,cex=0.5)

# Tamaños muestrales
# Regiones 1,2 y 3 respectivamente y t.
Muestral <- c(6,9,10,6)      # 25 datos en el espacio y 6 tiempos
Muestra2 <- c(6,9,10,10)    # 25 datos en el espacio y 10 tiempos

#####
##### PROGRAMA SIMULACION EN LAS REGIONES Y VARIABLE BINOMIAL GRF #####
#####

# Funciones escenarios:

phi.u <- (1/sqrt(2*pi))*exp((-2^2)/2)
# dnorm(2, mean = 0, sd = 1, log = FALSE)

fx <- function(x){
fx <- 1.5*dnorm((x-0.35)/0.15)-dnorm((x-0.8)/0.04)
fx
}

fxy <- function(x,y){
fxy <- (1.5*dnorm((x-0.35)/0.15)-dnorm((x-0.8)/0.04))*
      (1.5*dnorm((y-0.35)/0.15)-dnorm((y-0.8)/0.04))
fxy
}

fxyt <- function(x,y,t){

```

```

fxyt <- (1.5*dnorm((x-0.35)/0.15)-dnorm((x-0.8)/0.04))*(1.5*dnorm((y-0.35)/
0.15)-dnorm((y-0.8)/0.04))*(1.5*dnorm((t-0.35)/0.15)-dnorm((t-0.8)/
0.04))

fxyt
}

sigma.j <- function(j){
s.j <- 0.02 + 0.04*(j-1)^2
s.j
}

fjx <- function(j,x){
fjx <- sqrt(x*(1-x))*sin((2*pi*(1+2^((9-4*j)/5)))/(x+2^((9-4*j)/5)))
fjx
}

fjxy <- function(j,x,y){
fjxy <- (sqrt(x*(1-x))*sin((2*pi*(1+2^((9-4*j)/5)))/(x+2^((9-4*j)/5))))*
(sqrt(y*(1-y))*sin((2*pi*(1+2^((9-4*j)/5)))/(y+2^((9-4*j)/5))))
fjxy
}

fjxyt <- function(j,x,y,t){
fjxyt <- (sqrt(x*(1-x))*sin((2*pi*(1+2^((9-4*j)/5)))/(x+2^((9-4*j)/5))))*
(sqrt(y*(1-y))*sin((2*pi*(1+2^((9-4*j)/5)))/(y+2^((9-4*j)/5))))*
(sqrt(t*(1-t))*sin((2*pi*(1+2^((9-4*j)/5)))/(t+2^((9-4*j)/5))))
fjxyt
}

vjx <- function(j,x){
vjx <- (0.15*(1+0.4*(2*j-7)*(x-0.5)))^2
vjx
}

vjxy <- function(j,x,y){
vjxy <- ((0.15*(1+0.4*(2*j-7)*(x-0.5)))^2)*((0.15*(1+0.4*(2*j-7)*(y-0.5)))^2)
vjxy
}

vjxyt <- function(j,x,y,t){

```

```
vjxyt <- ((0.15*(1+0.4*(2*j-7)*(x-0.5)))^2)*((0.15*(1+0.4*(2*j-7)*
(y-0.5)))^2)*((0.15*(1+0.4*(2*j-7)*(t-0.5)))^2)
```

```
vjxyt
}
```

```
bjx <- function(j,x){
bjx <- dbeta(x,(j+4)/5,(11-j)/5)
bjx
}
```

```
-----
##### VALORES PARAMETROS #####
```

```
B0 <- 10; B1<- -4; B2 <- 2; B3 <- -4; p <- 0.4
# Se cargan las funciones a utilizar:
source("D:/...../function/rbf.trst.r")
source("D:/...../function/RBF.phi.r")
source("D:/...../function//rbf.st.tcv.r")
```

```
-----
##### Factor: Generic Form #####
```

```
# N:      Tamaños de muestra para T=6 tiempos 150 datos y cuando T=10,
#         250 datos
# nsim:   número de simulaciones, en general 100
# eta:    parámetro de suavizamiento en la optimización de la función de
#         base radial (FBR), 0.01 y 0.1
# n.neigh: número de vecinos para la validacion cruzada, en este caso el
#         número de nodos en el spline 8 y 32
# func:   tipo de FBR (ST, M, TPS, EXP, y CRS)
# j:      parámetro asociado a la variacion espacial, 1 y 3
```

```
rbf.sim <- function(N, nsim, eta, rho, n.neigh, t, func, j){
```

```
x <- seq(1, 5, 1)/5-0.1
y <- seq(1, 5, 1)/5-0.1
x1 <- rep(1:5, c(5,5,5,5,5))/5 -0.1
y1 <- rep(y,5)
coord <- data.frame(x1,y1)
names(coord) <- c("x","y")
```

```

B0 <- 10; B1<- -4; B2 <- 2; B3 <- -4; B4 <- 3; p <- 0.4
T <- c(1,t,1)      ## note necessarily gridtriple definition
ti <- (1:t)/t

ma1=as.matrix(rbind(c(5,0),c(0,0.5)))
model <- list("$", var=1, aniso=ma1, list("gauss"))

set.seed(127)
z <- GaussRF(n=nsim, x=x, y=y, T=T, grid=TRUE, model=model, method="ci",
             CE.strategy=1,CE.trials=if (interactive()) 4 else 1)
Z <- matrix(NA,ncol=nsim,nrow=N*t)
for (i in 1:nsim)
  Z[,i] <- as.vector(z[,,,i])

set.seed(127)
V1 <- rbinom(N, size=1, prob=p)

#build nominal covariate vector for prediction locations indicating subarea#
ind.reg.sim <- numeric(nrow(coord))
ind.reg.sim[.geoR_inout(coord,p1)] <- 1
ind.reg.sim[.geoR_inout(coord,p2)] <- 2
ind.reg.sim[.geoR_inout(coord,p3)] <- 3
ind.reg.sim <- as.factor(ind.reg.sim)

V2 <- ifelse(ind.reg.sim %in% 2,1,0)
V3 <- ifelse(ind.reg.sim %in% 3,1,0)

grilla <- data.frame(rep(x1,t),rep(y1,t),as.numeric(kronecker(ti,rep(1,25))),
                    rep(V1,t),rep(V2,t),rep(V3,t)) # i=1,...,100 (simulacion_i)
names(grilla) <- c("x","y","t","V1","V2","V3")

# Noise Level:
# Z1 <- B0 + B1*V1 +B2*V2*grilla$x +B3*V3*grilla$y+B4*grilla$t+fx(grilla$x)
# +fx(grilla$y) +fx(grilla$t)+fxy(grilla$x,grilla$y) +fxy(grilla$x,grilla$t)
# +fxy(grilla$y,grilla$t) + fxyt(grilla$x,grilla$y,grilla$t) + sigma.j(j)*Z

# Design density:
# Z1 <- B0 + B1*V1 + B2*V2*grilla$x + B3*V3*grilla$y + B4*grilla$t +
# fx(bjx(j,grilla$x))+fx(bjx(j,grilla$x))+fx(bjx(j,grilla$t))+

```

```

# fxy(bjx(j,grilla$x),bjx(j,grilla$y))+fxy(bjx(j,grilla$x),bjx(j,grilla$t))
# +fxy(bjx(j,grilla$y),bjx(j,grilla$t))+fxyt(bjx(j,grilla$x),bjx(j,grilla$y),
# bjx(j,grilla$t))+0.1*Z
# Spatial variation:
# Z1 <-B0+B1*V1+B2*V2*grilla$x+B3*V3*grilla$y +B4*grilla$t+fjx(j,grilla$x)+
# fjx(j,grilla$y) + fjx(j,grilla$t) + fjxy(j,grilla$x,grilla$y) +
# fjxy(j,grilla$x,grilla$t)+fjxy(j,grilla$y,grilla$t) +
# fjxyt(j,grilla$x,grilla$y,grilla$t)+0.2*Z

# Variance function:
Z1 <-B0+B1*V1+B2*V2*grilla$x+B3*V3*grilla$y+B4*grilla$t+fx(grilla$x) +
fx(grilla$y)+fx(grilla$t)+fxy(grilla$x,grilla$y)+(sqrt(vjx(j,grilla$x) +
vjx(j,grilla$y)+vjx(j,grilla$t)+vjxy(j,grilla$x,grilla$y)+vjxy(j,grilla$x,
grilla$t)+vjxy(j,grilla$y,grilla$t)+vjxyt(j,grilla$x,grilla$y,grilla$t)))*Z

tabla.sim <- as.data.frame(matrix(NA,nrow= nsim, ncol=5))
colnames(tabla.sim) <- c("sim", "MPE", "RMSPE", "R2", "No.CP")

#####
##### CONSTRUCCION COORDENADAS PRINCIPALES Y FBR DB #####
#####

# Delt <- daisy(grilla,type=list(asymm=c("V1","V2","V3")),metric="gower")
Delt <- gowdis(grilla, asym.bin = 4:6)
class(Delt)
library(ade4)
is.euclid(Delt^(1/2))

mds <- cmdscale(Delt^(1/2), k = nrow(grilla)-1, eig = TRUE)
names(mds)
round(mds$points[,1],4)

m <- sum(mds$eig > 0.007)
mds <- cmdscale(Delt^0.5, k = m, eig = TRUE)
X <- mds$points

for(i in 1:nsim){

ValoresPropios <- mds$eig
CorrCuadrado <- as.vector(cor(Z1[,i],X)^2) # 1 luego i

```

```

Porc.Inercia <- ValoresPropios/length(CorrCuadrado)
o<-data.frame(1:length(CorrCuadrado),round(ValoresPropios[1:
  length(CorrCuadrado)],10),round(CorrCuadrado,10),round(Porc.Inercia[1:
  length(CorrCuadrado)],10))names(o)<-c("ID", "ValoresProp", "CorrCuad",
  "Porc.Inercia")
names(o)
o1<-o[o$CorrCuad>0.007,] # Así quedan las significativas al 5\%
Xr.sim <- X[,o1$ID]
rdb.sim <- lm(Z1[,i] ~ Xr.sim)
model.db.sim <- summary(rdb.sim)

# x <- sim$coords[,1]
# y <- sim$coords[,2]
# system.time(rbf.st.tcv(eta=0.1,z=Z1[,i],coordinates=grilla[,1:3],trend=
  Xr.sim,rho=0, n.neigh=8, func="ST"))
try({spdb.sim <- rbf.st.tcv(eta=eta, z=Z1[,i], coordinates=grilla[,1:3],
  trend=Xr.sim, rho=rho, n.neigh=n.neigh, func=func)})
MPE <- mean(spdb.sim$residual)
RMSPE <- sqrt(sum(spdb.sim$residual^2)/length(spdb.sim$residual))
resid.mean.sim <- spdb.sim$observed - mean(spdb.sim$observed)
R2 <- 1- sum((spdb.sim$residual)^2)/sum(resid.mean.sim^2)

tabla.sim[i,] <- c(i,MPE,RMSPE,R2,ncol(Xr.sim))
}
as.matrix(tabla.sim)
}

-----
##### Factor: Noise level #####

Resultado.simn1 <- array(data=NA,c(100,5,5,2,2,2,2), dimnames=list(1:100,
  c("sim","MPE", "RMSPE", "R2","No.CP"),c("M","TPS","CRS","ST","EXP"),
  c("1","3"),c("0.01","0.1"),c("6","10"),c("8","32")))
funci <- c("M","TPS","CRS","ST","EXP")
for (func in funci)
for (j in c(1,3))
for (eta in c(0.01,0.1))
for (N in c(6,10))
for (n.neigh in c(8,32))
Resultado.simn1[, ,func,paste(j,sep=""),paste(sigma,sep=""),paste(N,sep=""),

```

```

paste(n.neigh,sep="")] <- rbf.sim(N=N, nsim=100, sigma=sigma, n.neigh=
n.neigh,func=func, j=j)

-----
#####          Factor: Design density          #####

Resultado.simbd <- array(data=NA,c(100,5,5,2,2,3,2), dimnames=list(1:100,
  c("sim","MPE","RMSPE","R2","No.CP"),c("M","TPS","CRS","ST","EXP"),
  c("1","3"),c("0.01","0.1"),c("6","10"),c("8","32")))
funci <- c("M","TPS","CRS","ST","EXP")
for (func in funci)
for (j in c(1,3))
for (eta in c(0.01,0.1))
for (N in c(6,10))
for (n.neigh in c(8,32))
Resultado.simbd[,func,paste(j,sep=""),paste(sigma,sep=""),paste(N,sep=""),
  paste(n.neigh,sep="")] <- rbf.sim(N=N, nsim=100, sigma=sigma, n.neigh=
  n.neigh,func=func, j=j)

-----
#####          Factor: Spatial variation          #####

Resultado.simsv <- array(data=NA,c(100,5,5,2,2,2,2), dimnames=list(1:100,
  c("sim","MPE","RMSPE","R2","No.CP"),c("M","TPS","CRS","ST","EXP"),
  c("1","3"),c("0.01","0.1"),c("6","10"),c("8","32")))
funci <- c("M","TPS","CRS","ST","EXP")
for (func in funci)
  for (j in c(1,3))
    for (eta in c(0.01,0.1))
      for (t in c(6,10))
        for (n.neigh in c(8,32))
Resultado.simsv[,func,paste(j,sep=""),paste(eta,sep=""),paste(t,sep=""),
  paste(n.neigh,sep="")] <- rbf.sim(N=N, nsim=100, eta=eta, t=t, n.neigh=
  n.neigh,func=func, j=j)

-----
#####          Factor: Variance function          #####

```

```

Resultado.simvf <- array(data=NA,c(100,5,5,2,2,2,2), dimnames=list(1:100,
  c("sim","MPE","RMSPE","R2","No.CP"),c("M","TPS","CRS","ST","EXP"),
  c("1","3"),c("0.01","0.1"),c("6","10"),c("8","32")))
funci <- c("M","TPS","CRS","ST","EXP")
for (func in funci)
  for (j in c(1,3))
    for (eta in c(0.01,0.1))
      for (t in c(6,10))
        for (n.neigh in c(8,32))
Resultado.simvf[, ,func,paste(j,sep=""),paste(eta,sep=""),paste(t,sep=""),
  paste(n.neigh,sep="")] <- rbf.sim(N=N, nsim=100, eta=eta, t=t, n.neigh=
  n.neigh,func=func, j=j)

```

```

-----
##### CAPITULO 5: APLICACIÓN SPACE-TIME FBR CROATIA #####
-----

```

```

# -----
# STEP 0: Configuración inicial y descarga de los datos:
# -----

library(maptools)
library(gstat)
library(fields)
library(rgdal)
library(lattice)
library(spatstat)
library(RSAGA)
utm33 <- "+proj=utm +zone=33 +ellps=WGS84 +datum=WGS84 +units=m +no_defs"

# Download auxiliary maps and station measurements:
download.file("http://spatial-analyst.net/book/sites/default/files/
HRclim2008.zip",destfile=paste(getwd(), "HRclim2008.zip", sep="/"))
unzip(zipfile="HRclim2008.zip", exdir=getwd())
unlink("HRclim2008.zip")

# Download MODIS LST images:
download.file("http://spatial-analyst.net/book/sites/default/files/
LST2008HR.zip",destfile=paste(getwd(), "LST2008HR.zip", sep="/"))

```



```

unzip(zipfile="LST2008HR.zip", exdir=paste(getwd(), "/LST", sep=""))
unlink("LST2008HR.zip")

# -----
# STEP 1: Importación de los datos y formateo (estandarización);
# -----

# Import location of stations:
IDSTA<-read.table("stations_temp_xy_2008.csv",header=TRUE,sep="," ,quote="\")
coordinates(IDSTA) <- ~LON+LAT
proj4string(IDSTA)<-CRS("+proj=longlat+ellps=bessel+towgs84=550.499,164.116,
                      475.142,5.80967,2.07902,-11.62386,0.99999445824")
IDSTA.ll <- spTransform(IDSTA, CRS("+proj=longlat+ellps=WGS84+datum=WGS84"))
writeOGR(IDSTA.ll, "gl_stations.kml", "IDSTA", "KML")
IDSTA.utm <- spTransform(IDSTA, CRS(utm33))
writeOGR(IDSTA.utm, "gl_stations.shp", "IDSTA", "ESRI Shapefile")
locs <- as.data.frame(IDSTA.utm)
names(locs)[c(5,6)] <- c("X", "Y")
str(locs)
# ST_ID      : Identification code of a station
# NAME       : Station name
# SP_CODE    : Station type "gl"-main,"kl"-climatological,"ks"-precipitation
# ELEV       : Station elevation in official database

# convert to lines:
rsaga.geoprocessor(lib="shapes_lines", module=0, param=list(LINES=
                  "borders.shp", POLYGONS="countries_s.shp"))
# import country borders:
borders <- readOGR(".", "borders")

# Croatia.shp fue generado en ArcGIS 10.0, allí se eliminaron los bordes
# de otros países del archivo "countries_s.shp"

# Import temperature at stations:
HRtemp2008 <- read.delim("temp_2008.csv", header=TRUE, sep="," ,quote="\")
# NA values:
HRtemp2008$T07[HRtemp2008$T07== -99.9] <- NA; HRtemp2008$T14[HRtemp2008$T14==
-99.9] <- NA; HRtemp2008$T21[HRtemp2008$T21== -99.9] <- NA
HRtemp2008$TEMP <- (HRtemp2008$T07+HRtemp2008$T14+2*HRtemp2008$T21)/4
str(HRtemp2008)# Mean daily temperature for 365 days (2008) at 152 locations

```

```

summary(HRtemp2008$TEMP) # 712 NA's
HRtemp2008$ST_ID <- as.factor(HRtemp2008$ST_ID)
# format the DATES (convert to number of cumulative days since 1970-01-01):
HRtemp2008$DATE <- as.Date(as.character(HRtemp2008$DATE))
HRtemp2008$cday <- floor(unclass(as.POSIXct(HRtemp2008$DATE))/86400)
# floor(unclass(as.POSIXct("2008-01-30"))/86400)[[1]]
# 13907
# inverse transformation:
# as.POSIXct(13907*86400, origin="1970-01-01")

# temperature the day before:
tmp <- data.frame(ST_ID=HRtemp2008$ST_ID, DATE=HRtemp2008$DATE+1,
                  TEMP1M=HRtemp2008$TEMP)
HRtemp2008.f <- merge(HRtemp2008, tmp, by=c("DATE", "ST_ID"), all.x=TRUE)
# check numbers:
HRtemp2008.f[(0:5)*(159-1)+1,c("ST_ID", "DATE", "TEMP", "TEMP1M")]

# Import grids:
grids <- readGDAL("HRdem.asc")
names(grids@data)[1] <- "HRdem"
for(j in c("HRdsea", "HRtwi")){
  grids@data[,j] <- readGDAL(paste(j, ".asc", sep=""))$band1
}
proj4string(grids) <- CRS(utm33)
# create dummy grids (Lat/Lon):
grids.ll <- spTransform(grids[1], CRS("+proj=longlat +datum=WGS84"))
grids$Lat <- grids.ll@coords[,2]
grids$Lon <- grids.ll@coords[,1]
str(grids@data)
# spplot(grids["Lat"])
lat.c <- mean(grids.ll@coords[,2])

# -----
# STEP 2: importación de las grillas y derivación de INSOL;
# -----

# derive total solar insolation for 365 days:
writeGDAL(grids["HRdem"], "HRdem.sdat", "SAGA")
writeGDAL(grids["Lat"], "HRlat.sdat", "SAGA")
writeGDAL(grids["Lon"], "HRlon.sdat", "SAGA")

```

```

Sys.chmod(getwd(), mode="0777"); dir.create("INSOL")
for(j in 1:365){
rsaga.geoprocessor(lib="ta_lighting",module=3,param=list(GRD_DEM="HRdem.sgrd"
,GRD_LAT="HRlat.sgrd", GRD_LON="HRlon.sgrd", GRD_TOTAL=paste("INSOL/",
"INSOL",j,".sgrd",sep=""), PERIOD=1, D HOUR=4, DAY_A=j),
show.output.on.console=FALSE)
# read to R:
grids@data[,paste("INSOL", j, sep="")] <- readGDAL(paste("INSOL/", "INSOL",
j,".sdat", sep=""), silent=TRUE)$band1
} # takes time!!

# Import LST MODIS images
# List of images:
LST.listday <- dir(path=paste(getwd(), "LST",sep="/"),pattern=
glob2rx("LST2008_**_**.LST_Day_1km.tif"), full.names=TRUE)
for(i in 1:length(LST.listday)){
LSTname<-strsplit(strsplit(LST.listday[i],"LST/")[[1]][2],".LST_")[[1]][1]
tmp1 <- readGDAL(LST.listday[i], silent=TRUE)
# convert to celsius:
grids@data[,LSTname] <- ifelse(tmp1$band1<=7500, NA, tmp1$band1*0.02-273.15)
}

croatia.shp <- readShapePoly("D:/...../croatia.shp")

#-----
#STEP 3: Ordenación de los datos (matriz espacio-tiempo) y gráficos
#-----

library(maptools)
pts <- spsample(croatia.shp, n=25000, type="regular")
df.pts <- as.data.frame(pts)
names(df.pts) <- c("x", "y")
coordinates(df.pts) <- c("x", "y")
IDSTA.OV <- overlay(grids, df.pts)
plot(IDSTA.OV@coords,cex=0.1)
spplot(IDSTA.OV["HRdem"],cex=0.1, at=seq(-15,30,by=45/50), col.regions=
grey(seq(0,1,by=0.02)), sp.layout=list("sp.lines", col="black",
croatia.shp))
IDSTA.OV1 <- data.frame(IDSTA.OV@coords,IDSTA.OV$HRdem,IDSTA.OV$HRdsea,

```

```

IDSTA.OV$HRtwi,
apply(cbind(IDSTA.OV$LST2008_01_01, IDSTA.OV$LST2008_01_09,
IDSTA.OV$LST2008_01_17, IDSTA.OV$LST2008_01_25), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_02_02, IDSTA.OV$LST2008_02_10,
IDSTA.OV$LST2008_02_18, IDSTA.OV$LST2008_02_26), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_03_05, IDSTA.OV$LST2008_03_13,
IDSTA.OV$LST2008_03_21, IDSTA.OV$LST2008_03_29), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_04_06, IDSTA.OV$LST2008_04_14,
IDSTA.OV$LST2008_04_22, IDSTA.OV$LST2008_04_30), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_05_08, IDSTA.OV$LST2008_05_16,
IDSTA.OV$LST2008_05_24), 1, mean, na.rm = TRUE),
apply(cbind(IDSTA.OV$LST2008_06_01, IDSTA.OV$LST2008_06_09,
IDSTA.OV$LST2008_06_17, IDSTA.OV$LST2008_06_25), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_07_03, IDSTA.OV$LST2008_07_11,
IDSTA.OV$LST2008_07_19, IDSTA.OV$LST2008_07_27), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_08_04, IDSTA.OV$LST2008_08_12,
IDSTA.OV$LST2008_08_20, IDSTA.OV$LST2008_08_28), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_09_05, IDSTA.OV$LST2008_09_13,
IDSTA.OV$LST2008_09_21, IDSTA.OV$LST2008_09_29), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_10_07, IDSTA.OV$LST2008_10_15,
IDSTA.OV$LST2008_10_23, IDSTA.OV$LST2008_10_31), 1, mean, na.rm=TRUE),
apply(cbind(IDSTA.OV$LST2008_11_08, IDSTA.OV$LST2008_11_16,
IDSTA.OV$LST2008_11_24), 1, mean, na.rm = TRUE),
apply(cbind(IDSTA.OV$LST2008_12_02, IDSTA.OV$LST2008_12_10,
IDSTA.OV$LST2008_12_18, IDSTA.OV$LST2008_12_26), 1, mean, na.rm=TRUE))

names(IDSTA.OV1) <- c("x", "y", "dem", "dsea", "twi", "MT1", "MT2", "MT3", "MT4",
                    "MT5", "MT6", "MT7", "MT8", "MT9", "MT10", "MT11", "MT12")
Grilla <- IDSTA.OV1

# overlay IDSTA and grids:
IDSTA.ov <- overlay(grids, IDSTA.utm)
locs.ov <- cbind(IDSTA.ov@data[c("HRdem", "HRdsea", "HRtwi", "Lat", "Lon")], locs)

# Merge the locations and constant predictors:
HRtemp2008locs <- merge(HRtemp2008.f[c("ST_ID", "TEMP", "TEMP1M", "DATE",
                                       "cday")], locs.ov, by="ST_ID")

str(HRtemp2008locs)
HRtemp2008locs$cday <- HRtemp2008locs$cday-13878
HRtemp2008locs[(HRtemp2008locs$ST_ID %in% 663),]

```

```

HRtemp2008locs$EST<- ifelse(HRtemp2008locs$cday %in% c(1:80,355:367),1,
  ifelse(HRtemp2008locs$cday %in% c(81:172),2, ifelse(HRtemp2008locs$cday
    %in% c(173:266),3,4)))
HRtemp2008locs$MONTH<- ifelse(HRtemp2008locs$cday %in% 1:31,1,
  ifelse(HRtemp2008locs$cday %in% 32:60,2,ifelse(HRtemp2008locs$cday %in%
    61:91,3, ifelse(HRtemp2008locs$cday %in% 92:121,4,ifelse(
    HRtemp2008locs$cday %in% 122:152,5,ifelse(HRtemp2008locs$cday %in%
    153:182,6, ifelse(HRtemp2008locs$cday %in% 183:213,7,ifelse(H
    Rtemp2008locs$cday %in% 214:244,8,ifelse(HRtemp2008locs$cday %in%
    245:274,9,ifelse(HRtemp2008locs$cday %in% 275:305,10,ifelse(
    HRtemp2008locs$cday %in% 306:335,11,12)))))))))))))

HRtemp2008locs[(HRtemp2008locs$TEMP %in% NA),]$ST_ID
HRtemp2008locs[(HRtemp2008locs$ST_ID ==1),]
HRtemp2008locs$EST<- ifelse(HRtemp2008locs$cday %in% c(1:80,355:367),1,
  ifelse(HRtemp2008locs$cday %in% c(81:172),2, ifelse(HRtemp2008locs$cday
    %in% c(173:266),3,4)))
write.table(HRtemp2008locs, "D:/...../Aplicación/Muestra.txt", sep=" ",
  col.names=TRUE, row.names=TRUE, quote=TRUE, na="NA")

## Construcción Mapa Ubicacion Estaciones Croacia
par(mfrow=c(1,1), mar=c(4,6,3,2))
plot(croatia.shp, las=2, col="cornsilk3",border="black", axes=F,
xlim=c(min(IDSTA.OV1$x),max(Grilla$x)), ylim=c(min(Grilla$y),max(Grilla$y)))
at.y <- (48:51)*100000
axis(2, at = at.y, labels = formatC(at.y, format="fg"), las=2)
my.at <- (4:8)*100000
axis(1, at = my.at, labels = my.at)
box(lty = 'solid', )
#main="Ubicación Estaciones 2008-12-01")
points(Muestra$X,Muestra$Y,pch=20,cex=0.9)
title(xlab="East(m)", line=3)
title(ylab="North(m)", line=5)

## Construcción mapas variables explicativas continuas Croacia
pacol <- colorRampPalette(c("white","black"))
GrillaDDT <- Grilla[,1:5]
names(GrillaDDT) <- c("x","y","dem","dsea","twi")
addCr<-list("sp.polygons",croatia.shp,col="gray25",first = FALSE)

```

```

gridded(GrillaDDT) = ~x+y
pdem <- spplot(GrillaDDT["dem"],col.regions=pacol,sp.layout=addCr,cuts=40,
              scales = list(draw =T), xlab="East (m)", ylab = "North (m)",
              key.space=list(space="right", cex=0.8))
pdsea <- spplot(GrillaDDT["dsea"],col.regions=pacol,sp.layout=addCr,cuts=40,
              scales = list(draw =T), xlab="East (m)", ylab = "North (m)",
              key.space=list(space="right", cex=0.8))
ptwi <- spplot(GrillaDDT["twi"],col.regions=pacol,sp.layout=addCr,cuts=40,
              scales = list(draw =T), xlab="East (m)", ylab = "North (m)",
              key.space=list(space="right", cex=0.8))

#-----
#STEP 4: Calculo de coordenadas principales y regresión basada en distancias
#-----

library(cluster)

Data.f <- data.frame(Muestra$X,Muestra$Y,Muestra$MONTH,Muestra$HRdem,
                    Muestra$HRdsea, Muestra$HRtwi,Muestra$EST) #,Muestra$SDTEMP)
names(Data.f) <- c("x","y","t","dem","dsea","twi","est") #,"sdt")
Data.f[,7] <- as.factor(Data.f[,7])

Delta <- daisy(Data.f, metric = "gower")
class(Delta)
library(ade4)
is.euclid(Delta^(1/2))

mds <- cmdscale(Delta^(1/2), k = nrow(Data.f)-1, eig = TRUE)
names(mds)
round(mds$points[,1],4)

m <- sum(mds$eig > 0.007)
mds <- cmdscale(Delta^0.5, k = m, eig = TRUE)
X <- mds$points

ValoresPropios <- mds$eig

CorrCuadrado <- as.vector(cor(Muestra$MTEMP,X)^2)
Porc.Inercia <- ValoresPropios/length(CorrCuadrado)
o<-data.frame(1:length(CorrCuadrado),round(ValoresPropios[1:

```

```

        length(CorrCuadrado)],10),round(CorrCuadrado,10),round(Porc.Inercia
        [1:length(CorrCuadrado)],10))
names(o)<-c("ID", "ValoresProp", "CorrCuad", "Porc.Inercia")
names(o)
o1<-o[o$CorrCuad>0.003,]
#fix(o1)
Xr <- X[,o1$ID]
rdb <- lm(Muestra$MTEMP ~ Xr)
model.db <- summary(rdb)

aux <- CorrCuadrado[1:10]*ValoresPropios[1:10]/sum(CorrCuadrado[1:10]*
        ValoresPropios[1:10])
c.pred <- c(0,cumsum(aux))

plot(0:14,1-c.pred[1:15],
      xlab="Principal Coordinates",
      ylab="1 - Predictability",
      main=c("No Predictability Principal Coordinates"),
      cex.main=1.2, col.main=4, ylim=c(0,1), xlim=c(0,12), type="l")
abline(v=3,lty=2,col="blue")

# -----
# STEP 5: Calculo de las coordenadas principales de nuevos individuos;
# -----

library(gstat)
x.news <- function(ValoresPropios,Data,tendencia,newdata){
Data[length(ValoresPropios)+1, ] <- newdata
d <- as.matrix(gowdis(Data))[-(length(ValoresPropios)+1),
        (length(ValoresPropios)+1)]
b <- diag(tcrossprod(tendencia))
x.new <- (1/2)*diag(ValoresPropios[1:ncol(tendencia)]^(-1))%*%t(tendencia)
        %*%(b-d)
x.new
}

tabla <- function(newdata){
x.new0<-x.news(ValoresPropios=ValoresPropios,Data=Data.f,tendencia=
        Xr,newdata)

```

```
x.new0
}

tabla1 <- apply(Grilla25.m1,1,tarla)
tabla2 <- apply(Grilla25.m2,1,tarla)
tabla3 <- apply(Grilla25.m3,1,tarla)
tabla4 <- apply(Grilla25.m4,1,tarla)
tabla5 <- apply(Grilla25.m5,1,tarla)
tabla6 <- apply(Grilla25.m6,1,tarla)
tabla7 <- apply(Grilla25.m7,1,tarla)
tabla8 <- apply(Grilla25.m8,1,tarla)
tabla9 <- apply(Grilla25.m9,1,tarla)
tabla10 <- apply(Grilla25.m10,1,tarla)
tabla11 <- apply(Grilla25.m11,1,tarla)
tabla12 <- apply(Grilla25.m12,1,tarla)

# -----
# STEP 6: optimización y generación de mapas de interpolación;
# -----

# optimización CRS FBR
rbf.st.cvop(c(0.0001,0.001),z=Muestra$MTEMP,coordinates=
           scale(Muestra[,c(14,15,17)]), trend=Xr, n.neigh=30, func="CRS")
rbf.st.cvop(c(0.001,0),z=Muestra$MTEMP, coordinates=
           scale(Muestra[,c(14,15,17)]), trend=Xr, n.neigh=30, func="CRS")
rbf.st.cvop(c(0.01,0),z=Muestra$MTEMP, coordinates=
           scale(Muestra[,c(14,15,17)]), trend=Xr, n.neigh=30, func="CRS")

# optimización TPS RBF
rbf.st.cvop(c(0.001,0),z=Muestra$MTEMP, coordinates=
           scale(Muestra[,c(14,15,17)]),trend=Xr, n.neigh=30, func="TPS")
rbf.st.cvop(c(0.01,0),z=Muestra$MTEMP, coordinates=
           scale(Muestra[,c(14,15,17)]),trend=Xr, n.neigh=30, func="TPS")
rbf.st.cvop(c(0.1,0),z=Muestra$MTEMP, coordinates=
           scale(Muestra[,c(14,15,17)]),trend=Xr, n.neigh=30, func="TPS")

# Las demás funciones se trabajaron de manera similar
```



```

pal2a <- colorRampPalette(c("blue3","wheat1","red3"))

#####
# CRS Spline Interpolation MTEMP #
#####

# Month 1
fbr.pred.CRS<-rbf.trst(eta=0.001,z=Muestra$MTEMP,coordinates=scale(Muestra
  [,c(14,15,17)]),trend=Xr,nd.trend=nd.trend1,rho=0,n.neigh=30,func="CRS")
fbr.CRS <- data.frame(Grilla25.m1[,1:2],fbr.pred.CRS[,4:5])

gridded(fbr.CRS) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(fbr.CRS["var1.pred"], col.regions=pal2, cuts=30,scales=list(draw=T),
  xlab="East (m)",
  ylab = "North (m)", key.space=list(space="right", cex=0.8))

#Month 4
fbr.pred.CRS4 <- rbf.trst(eta=0.001, z=Muestra$MTEMP,
  coordinates=scale(Muestra[,c(14,15,17)]),
  trend=Xr, nd.trend=nd.trend4, rho=0, n.neigh=30, func="CRS")
fbr.CRS4 <- data.frame(Grilla25.m4[,1:2],fbr.pred.CRS4 [,4:5])

gridded(fbr.CRS4) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(fbr.CRS4["var1.pred"], col.regions=pal2,cuts=30,scales=list(draw=T),
  xlab="East (m)",ylab="North (m)",key.space=list(space="right",cex=0.8))

#Month 7
fbr.pred.CRS7 <-rbf.trst(eta=0.001,z=Muestra$MTEMP,coordinates=scale(Muestra
  [,c(14,15,17)]),trend=Xr,nd.trend=nd.trend7,rho=0,n.neigh=30,func="CRS")
fbr.CRS7 <- data.frame(Grilla25.m7[,1:2],fbr.pred.CRS7 [,4:5])
#fbr.CRS7@data$var1.pred[fbr.CRS7@data$var1.pred>40] <-
# mean(fbr.CRS7@data$var1.pred[fbr.CRS7@data$var1.pred<40]+9)

gridded(fbr.CRS7) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(fbr.CRS7["var1.pred"], col.regions=by.colors(60), cuts=30, scales=
  list(draw =T), xlab="East (m)",
  ylab = "North (m)", key.space=list(space="right", cex=0.8))

```

```

# Month 10
fbr.pred.CRS10<-rbf.trst(eta=0.001,z=Muestra$MTEMP,coordinates=scale(Muestra
  [,c(14,15,17)]),trend=Xr,nd.trend=nd.trend10,rho=0,n.neigh=30,func="CRS")
fbr.CRS10 <- data.frame(Grilla25.m10[,1:2],fbr.pred.CRS10[,4:5])

gridded(fbr.CRS10) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(fbr.CRS10["var1.pred"], col.regions=bpy.colors(60), cuts=30,scales=
  list(draw=T), xlab="East (m)", ylab = "North (m)",
  key.space=list(space="right", cex=0.8))

## AGREGADO EN UN MAPA
fbr.CRS.A<-data.frame(Grilla25.m10[,1:2],fbr.pred.CRS[,4],fbr.pred.CRS4[,4],
  fbr.pred.CRS7[,4],fbr.pred.CRS10[,4])
names(fbr.CRS.A) <- c("x","y","MONTH1","MONTH4","MONTH7","MONTH10")

gridded(fbr.CRS.A) = ~x+y
pr.list <- c("MONTH1","MONTH4","MONTH7","MONTH10")
pr.list1 <- c("a","b","c","d")
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 4))
fbr.CRS.ag = fbr.CRS
fbr.CRS.ag[["a"]] = fbr.CRS[["var1.pred"]]
fbr.CRS.ag[["b"]] = fbr.CRS4[["var1.pred"]]
fbr.CRS.ag[["c"]] = fbr.CRS7[["var1.pred"]]
fbr.CRS.ag[["d"]] = fbr.CRS10[["var1.pred"]]
g2 <- spplot(fbr.CRS.ag,c("a","b","c","d"),names.attr=c("MONTH 1","MONTH 4",
  "MONTH 7","MONTH 10"),as.table=TRUE,col.regions=pal2a,cuts=30,scales=
  list(draw=T),xlab="East (m)",ylab="North (m)",key.space=list(space=
  "right", cex=0.8))
#spplot(fbr.CRS.A[pr.list], names.attr=c("MONTH 1", "MONTH 4", "MONTH 7",
  "MONTH 10"),col.regions=pal2,cuts=30,scales=list(draw=T),xlab="East (m)",
  ylab = "North (m)", key.space=list(space="right", cex=0.8))
#xlim=c(450000,600000), ylim=c(4950000,5100000),col.regions=bpy.colors(10)

#####
# EXP Spline Interpolation MTEMP #
#####

```

```

# Month 1
fbr.pred.EXP <- rbf.trst(eta=0.001, z=Muestra$MTEMP, coordinates=
  scale(Muestra[,c(14,15,17)]),trend=Xr,nd.trend=nd.trend1,rho=0,n.neigh=30,
  func="EXP")
fbr.EXP <- data.frame(Grilla25.m1[,1:2],fbr.pred.EXP[,4:5])

gridded(fbr.EXP) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(fbr.EXP["var1.pred"],col.regions=pal2,cuts=30,scales=list(draw =T),
  xlab="East (m)",
  ylab = "North (m)", key.space=list(space="right", cex=0.8))

# Month 4
fbr.pred.EXP4 <- rbf.trst(eta=0.001, z=Muestra$MTEMP, coordinates=
  scale(Muestra[,c(14,15,17)]),trend=Xr,nd.trend=nd.trend4,rho=0,n.neigh=30,
  func="EXP")
fbr.EXP4 <- data.frame(Grilla25.m1[,1:2],fbr.pred.EXP4[,4:5])

gridded(fbr.EXP4) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(fbr.EXP4["var1.pred"],col.regions=pal2, cuts=30,scales=list(draw =T),
  xlab="East (m)",ylab="North (m)",key.space=list(space="right",cex=0.8))

# Month 7
fbr.pred.EXP7 <- rbf.trst(eta=0.001, z=Muestra$MTEMP, coordinates=
  scale(Muestra[,c(14,15,17)]),trend=Xr,nd.trend=nd.trend7,rho=0,n.neigh=30,
  func="EXP")
fbr.EXP7 <- data.frame(Grilla25.m1[,1:2],fbr.pred.EXP7[,4:5])

gridded(fbr.EXP7) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(fbr.EXP7["var1.pred"],col.regions=terrain.colors(100),cuts=30,scales=
  list(draw =T),xlab="East (m)",ylab="North (m)",key.space=list(space="right",
  cex=0.8))

# Month 10
fbr.pred.EXP10 <- rbf.trst(eta=0.001, z=Muestra$MTEMP, coordinates=
  scale(Muestra[,c(14,15,17)]),trend=Xr,nd.trend=nd.trend10,rho=0,n.neigh=30,
  func="EXP")
fbr.EXP10 <- data.frame(Grilla25.m1[,1:2],fbr.pred.EXP10[,4:5])

```

```
gridded(fbr.EXP10) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 1))
spplot(fbr.EXP10["var1.pred"], col.regions=terrain.colors(100), cuts=30, scales=
  list(draw=T), xlab="East (m)", ylab="North (m)", key.space=list(space="right",
    cex=0.8))

## AGREGADO EN UN MAPA
fbr.EXP.A<-data.frame(Grilla25.m10[,1:2], fbr.pred.EXP[,4], fbr.pred.EXP4[,4],
  fbr.pred.EXP7[,4], fbr.pred.EXP10[,4])
names(fbr.EXP.A) <- c("x", "y", "MONTH1", "MONTH4", "MONTH7", "MONTH10")

gridded(fbr.EXP.A) = ~x+y
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 4))
fbr.EXP.ag = fbr.EXP
fbr.EXP.ag[["a"]] = fbr.EXP[["var1.pred"]]
fbr.EXP.ag[["b"]] = fbr.EXP4[["var1.pred"]]
fbr.EXP.ag[["c"]] = fbr.EXP7[["var1.pred"]]
fbr.EXP.ag[["d"]] = fbr.EXP10[["var1.pred"]]
g5<-spplot(fbr.EXP.ag, c("a", "b", "c", "d"), names.attr=c("MONTH 1", "MONTH 4",
  "MONTH 7", "MONTH 10"), as.table=TRUE, col.regions=pal2a, cuts=30, scales=
  list(draw=T), xlab="East (m)", ylab="North (m)", key.space=list(space=
    "right", cex=0.8))

# Los demás mapas se generaron de manera similar
```