



# Modelos lineales generalizados geoestadísticos basados en distancias

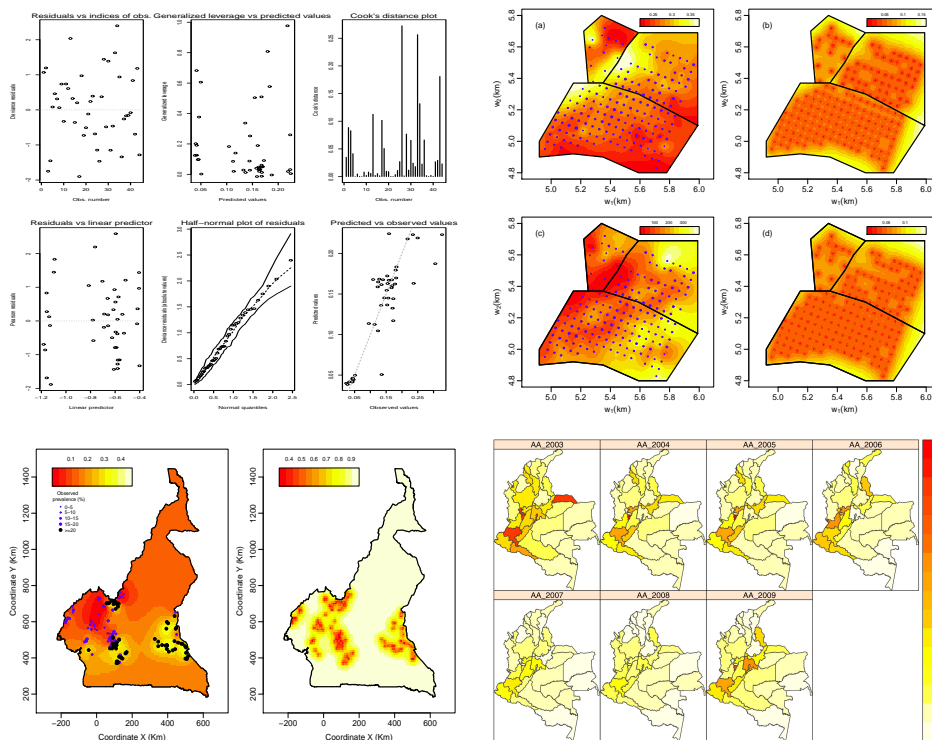
Oscar Orlando Melo Martínez

**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)) i a través del Dipòsit Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) 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 ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (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)) y a través del Repositorio Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) 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 o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (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 and by the UB Digital Repository ([diposit.ub.edu](http://diposit.ub.edu)) 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 nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those 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.

# Modelos lineales generalizados geoestadísticos basados en distancias



Oscar Orlando Melo Martínez





## Modelos lineales generalizados geoestadísticos basados en distancias

MEMORIA PRESENTADA POR:

**Oscar Orlando Melo Martínez**

PARA OPTAR AL TÍTULO DE DOCTOR POR LA UNIVERSIDAD DE BARCELONA

DOCTORANDO:

---

**Oscar Orlando Melo Martínez**

DIRECTOR:

TUTOR:

---

**Dr. Jorge Mateu Mahiques**

Departamento de Matemáticas  
Universidad Jaume I de Castellón

---

**Dr. Josep María Oller Sala**

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 2013



## Agradecimientos

A mi director de tesis y a mi tutor, los profesores Jorge Mateu y Josep María Oller, respectivamente, por ser los motivadores permanentes, por su constante interés, apoyo y porque los dos me han dedicado parte de su valioso tiempo guiándome en la realización de este trabajo. Mi admiración y sincera gratitud a los dos.

A mi hermano Carlos con quien a lo largo de la vida hemos compartido y trabajado en infinidad de temas científicos y no científicos, siendo así el doctorado una excusa más para trabajar en equipo y compartir. Una enorme gratitud por su colaboración y apoyo en mis estudios.

A mi madre su apoyo incondicional durante toda mi vida. Todos sus sacrificios hicieron posible llegar hasta este punto. Gracias por ser la mejor mamá del mundo. A mi padre, que a pesar de la distancia siempre estuvo atento para saber cómo iba mi proceso de estudios en España. Gracias a los dos por ser siempre el motor, el soporte y la razón de ser en todo lo que he emprendido.

A la Universidad Nacional de Colombia por darme la oportunidad de ir a otro país a realizar mis estudios de doctorado y por apoyarme durante estos años de estudio. En esta universidad llevo varios años de formación académica y profesional, pues allí realice mis estudios de Estadística y de Master en Estadística. Además, en esta prestigiosa universidad trabajé como docente y hago una de las labores que más placer me da en la vida que es hacer investigación. En especial muchas gracias, a todos los profesores del Departamento de Estadística por suplir mi ausencia en estos años de la universidad.

A la Universidad de Barcelona y la Universidad Politécnica de Cataluña por acogerme durante estos cuatro años de estudio y por las enseñanzas adquiridas durante este tiempo. En especial a mis profesores. A la Fundación Carolina por haberme otorgado una beca para realizar mis estudios de doctorado en Barcelona y por sus actividades para darme a conocer la ciudad.

A Diana Moreno por sus valiosos comentarios y correcciones sobre el documento escrito. También, a los revisores y editores anónimos por sus valiosos comentarios sobre lo escrito en los diferentes papers sometidos, ya que esta tesis es también producto de las correcciones y sugerencias realizadas por ellos.



A mi familia: hermanos, sobrinas, y en especial, a mis  
padres María y Gustavo, quienes con su sacrificio y  
esfuerzo me iniciaron a muy temprana edad en la pasión  
por el conocimiento.





# Contenido

<b>Lista de figuras</b>	<b>vi</b>
<b>Lista de tablas</b>	<b>x</b>
<b>Abreviaturas</b>	<b>xiii</b>
<b>Introducción</b>	<b>1</b>
<b>Objetivos</b>	<b>9</b>
<b>1 Conceptos básicos</b>	<b>11</b>
1.1 Introducción . . . . .	11
1.2 Análisis geoestadístico clásico . . . . .	12
1.2.1 Definiciones básicas . . . . .	13
1.2.2 El covariograma . . . . .	16
1.2.3 El variograma . . . . .	17
1.2.4 El correlograma . . . . .	17
1.2.5 Forma general de estas funciones . . . . .	18
1.3 Estimación del variograma y del covariograma . . . . .	19
1.4 Principales modelos de variogramas y covariogramas isotrópicos	20
1.5 Estimación de los parámetros del variograma . . . . .	22
1.5.1 Estimación por mínimos cuadrados . . . . .	23
1.6 Distancia, similaridad y descomposición espectral . . . . .	24
<b>2 DBBR con dispersión variable para la predicción de proporciones y tasas</b>	<b>29</b>

2.1	Introducción . . . . .	29
2.2	Modelo de regresión beta basado en distancia con dispersión variable . . . . .	32
2.2.1	Construcción del modelo beta utilizando distancias . . . . .	34
2.2.2	Estimación de parámetros . . . . .	36
2.2.3	Casos especiales . . . . .	40
2.2.4	Inferencia para muestras grandes . . . . .	41
2.3	Ajuste, selección, diagnóstico y predicción del modelo DBBR . . . . .	44
2.3.1	Medidas de bondad de ajuste . . . . .	44
2.3.2	Selección de variables para el modelo beta basado en distancias . . . . .	45
2.3.3	Medidas de diagnóstico . . . . .	48
2.3.4	Predicción de un nuevo individuo . . . . .	50
2.3.5	Tratamiento de datos faltantes bajo la aproximación basada en distancias . . . . .	52
2.4	Relación con el modelo de regresión beta clásico . . . . .	53
2.4.1	Variables continuas . . . . .	54
2.4.2	Variables Cualitativas . . . . .	55
2.4.3	Variables mixtas . . . . .	55
2.4.4	Modelo de regresión beta no lineal basado en distancias . . . . .	56
2.5	Aplicaciones . . . . .	57
2.5.1	Proporción de petróleo crudo convertido a gasolina . . . . .	57
2.5.2	Rendimiento en fondos de inversión . . . . .	60

**3 Modelos lineales generalizados espaciales mixtos basados en distancias 69**

3.1	Introducción . . . . .	69
3.2	Modelos espaciales mixtos basados en distancias . . . . .	72
3.2.1	Modelo mixto basado en distancias . . . . .	73
3.2.2	Algoritmo de máxima verosimilitud de Monte Carlo para DBSGLMM . . . . .	78
3.2.3	Versión Monte Carlo del algoritmo gradiente EM para la MLE de los parámetros del DBSGLMM . . . . .	82

3.2.4	Medidas de bondad de ajuste . . . . .	84
3.2.5	Selección de las coordenadas principales para el DBSGLMM reducido . . . . .	86
3.3	Predicción espacial de un nuevo individuo . . . . .	87
3.3.1	Predicción espacial . . . . .	87
3.4	Relación con el GLMM espacial clásico . . . . .	92
3.4.1	Variables continuas . . . . .	93
3.4.2	Variables cualitativas . . . . .	94
3.4.3	Variables mixtas . . . . .	94
3.4.4	DBSGLMM no lineal . . . . .	95
3.5	Aplicación . . . . .	95
3.5.1	Inferencia sobre los parámetros, diagnóstico y predicción utilizando DBSGLMM . . . . .	99
<b>4</b>	<b>Modelo lineal beta espacial mixto con dispersión variable</b>	<b>103</b>
4.1	Introducción . . . . .	103
4.2	Modelo lineal espacial beta mixto . . . . .	105
4.2.1	Algoritmo de máxima verosimilitud con Monte Carlo pa- ra el BSLMM . . . . .	109
4.2.2	Medidas de bondad de ajuste . . . . .	113
4.2.3	Predicción espacial de nuevos individuos . . . . .	114
4.3	Experimento simulado . . . . .	114
4.4	Aplicaciones . . . . .	118
4.4.1	Contenido de arcilla . . . . .	118
4.4.2	Contenido de Magnesio . . . . .	124
<b>5</b>	<b>Modelo DBGLSTARAR</b>	<b>135</b>
5.1	Introducción . . . . .	135
5.2	GLM dinámico espacio-tiempo utilizando el método basado en distancias . . . . .	138
5.2.1	Modelo lineal generalizado espacio-tiempo basado en dis- tancias . . . . .	144
5.3	Métodos de estimación de los parámetros . . . . .	147

5.3.1	Estimación por máxima verosimilitud . . . . .	147
5.3.2	Algoritmo de máxima verosimilitud vía Monte Carlo para DBGLSTARAR . . . . .	151
5.3.3	Ecuaciones de estimación generalizadas espacio-tiempo . . . . .	153
5.3.4	Elecciones específicas de $\mathbf{R}(\boldsymbol{\vartheta})$ . . . . .	155
5.4	Selección, validación y predicción del modelo ajustado utilizando GEE para espacio-tiempo . . . . .	157
5.4.1	Medida de bondad de ajuste . . . . .	158
5.4.2	Análisis residual . . . . .	158
5.4.3	Selección de las coordenadas principales para el modelo DBGLSTARAR reducido . . . . .	159
5.4.4	Predicción espacio-tiempo de un nuevo individuo . . . . .	160
5.5	Aplicación . . . . .	162
5.5.1	Análisis descriptivo . . . . .	163
5.5.2	Modelo GLSTARAR . . . . .	169
5.5.3	Modelo DBGLSTARAR . . . . .	174
5.5.4	Validación de los supuestos sobre los modelos GLSTARAR y DBGLSTARAR . . . . .	176
<b>6</b>	<b>DBSGLMVAR incorporando la dinámica tanto espacial como temporal</b>	<b>183</b>
6.1	Introducción . . . . .	183
6.2	GLM dinámico espacio-tiempo utilizando el método basado en distancia . . . . .	186
6.2.1	Vectores autorregresivos generalizados espaciales basados en distancias . . . . .	189
6.3	Algoritmo de máxima verosimilitud vía Monte Carlo para DBSGLMVAR . . . . .	196
6.4	Selección, validación y predicción del modelo ajustado . . . . .	198
6.4.1	Medidas de bondad de ajuste . . . . .	198
6.4.2	Selección de las coordenadas principales para el DBSGLMVAR reducido . . . . .	200
6.4.3	Predicción espacial de un nuevo individuo . . . . .	200

**7 Conclusiones y recomendaciones** **203**

**Referencias** **208**



# Lista de figuras

1.1	Forma general del variograma y covariograma de un proceso espacial homogéneo. . . . .	19
2.1	Diagnóstico del modelo con dispersión variable para la proporción de petróleo crudo convertido a gasolina. . . . .	61
2.2	Diagnósticos del modelo de dispersión variable para el rendimiento de fondos de inversión. . . . .	65
3.1	Nube de puntos del variograma (panel izquierdo) y variograma empírico (panel derecho) para la prevalencia de Loa loa utilizando DB. . . . .	97
3.2	Distribución de los parámetros involucrados en el DBSGLMM. . . . .	100
3.3	(a) Residuales de Pearson contra la aldea, (b) residuales de deviance contra la aldea, (c) residuales de Pearson contra valores pronosticados para el DBSGLMM, modelo (3.37), y (d) relación entre prevalencia observada de Loa loa microfilaria y prevalencia pronosticada usando DBSGLMM. . . . .	101
3.4	Prevalencias estimadas para el Loa loa microfilaria utilizando la aproximación DBSGLMM, sobrepuesta con la prevalencia observada en el campo de estudio (panel izquierdo). Raíz cuadrada de las varianzas de la predicción (panel derecho). . . . .	102
4.1	Comportamiento de la cadena muestreada para los parámetros del SMM, modelo (4.17). . . . .	119
4.2	Comportamiento de la cadena muestreada para los parámetros del SVDM, modelo (4.18). . . . .	120



4.3	(a) Ubicaciones del contenido de arcilla, (b) Diagrama de dispersión del contenido de arcilla contra la altura, donde la recta representa la curva lowess, (c) Diagrama de cajas del contenido de arcilla en cada una de las cuatro zonas y (d) Diagrama de cajas del contenido de arcilla en cada uno de los dos grupos de referencia de suelo. . . . .	121
4.4	Deviance residual contra: el ordenamiento de la profundidad de la capa de suelo ( <i>panel izquierdo</i> ) y los valores pronosticados ( <i>panel derecho</i> ) utilizando el BSLMM para el contenido de arcilla.	125
4.5	Deviance residual contra: ordenamiento de la capa de suelo ( <i>panel izquierdo</i> ) y los valores pronosticados ( <i>panel derecho</i> ) utilizando el BSLMM para el contenido de magnesio. . . . .	129
4.6	(a): Estimaciones puntuales utilizando el SMM, superpuesto con los contenidos de magnesio observados en los campos estudiados, donde cada punto es proporcional a la correspondiente medida del contenido de magnesio, y las líneas delimitan las sub-regiones con diferentes prácticas en el manejo del suelo. (b): Errores de predicción estándar para el SMM. (c): Estimaciones puntuales usando el SVDM (modelo de precisión), superpuesto con el contenido de magnesio observado en los campos estudiados. (d): Errores de predicción estándar para el SVDM. . . . .	130
5.1	División político administrativa departamental de Colombia. . .	164
5.2	Mapas de los quintiles del número de acciones armadas por departamento para el período 2003 a 2009 . . . . .	166
5.3	Mapas de estructuras de vecindarios para la construcción de pesos espaciales . . . . .	167
5.4	Gráficos de dispersión de Moran del número de acciones armadas por departamento, período 2003 a 2009 . . . . .	169
5.5	Mapas de influyentes por cuadrante de dispersión de Moran del número de acciones armadas por departamentos, período 2003 a 2009 . . . . .	172
5.6	Mapas del número de acciones armadas por departamento, período 2003 a 2009 . . . . .	172
5.7	Mapas de la predicción del número de acciones armadas por departamento, bajo el modelo GLSTARAR para el período 2003 a 2009 . . . . .	179
5.8	Mapas de residuales de Pearson por departamento bajo el modelo GLSTARAR para el período 2003 a 2009 . . . . .	179

5.9	Mapas de deviances por departamento bajo el modelo GLSTARAR para el período 2003 a 2009 . . . . .	180
5.10	Mapas de la predicción del número de acciones armadas por departamento bajo el modelo DBGLSTARAR para el período 2003 a 2009 . . . . .	180
5.11	Mapas de residuales de Pearson por departamento bajo el modelo DBGLSTARAR para el período 2003 a 2009 . . . . .	181
5.12	Mapas de deviances por departamento, bajo el modelo DBGLSTARAR para el período 2003 a 2009 . . . . .	181



# Lista de tablas

1.1	Formas funcionales de algunos variogramas. . . . .	21
1.2	Formas funcionales de algunos covariogramas. . . . .	22
2.1	Funciones de enlace utilizando los modelos de regresión beta clásico y DBBR con dispersión variable para la proporción de petróleo crudo convertido a gasolina. . . . .	59
2.2	Estimación de parámetros para el modelo DBBR con dispersión variable que relaciona la proporción de petróleo crudo convertido en gasolina con las coordenadas principales. . . . .	60
2.3	Funciones de enlace utilizando los modelos de regresión beta clásico y DBBR con dispersión variable para el retorno promedio de 5 años (%). . . . .	63
2.4	DBBR con dispersión variable sobre el conjunto de datos Morningstar con: 0% (sin datos faltantes), 5%, 10% y 20% de las observaciones con datos faltantes, y elección de $\kappa_v =  \lambda_{x_{k_1}} $ y $\kappa_u =  \lambda_{z_{k_2}} $ . . . . .	66
3.1	Estimaciones de máxima verosimilitud utilizando MCMC para los modelos $H_1, H_2, H_3, H_4$ y $H_5$ . . . . .	98
3.2	Estimaciones e intervalos de confianza para los parámetros involucrados en el DBSGLMM (modelo (3.37)) para ajustar la prevalencia de Loa loa. . . . .	99
4.1	Estimaciones beta espaciales por máxima verosimilitud para la simulación utilizando los modelos $H_1, H_2, H_3, H_4, H_5, H_6, H_7$ y $H_8$ . . . . .	117
4.2	Log-verosimilitudes para la simulación utilizando diferentes funciones de enlace en el SMM y el SVDM, con funciones de correlación esférica en los dos modelos. . . . .	118

4.3	Estimaciones beta espaciales por máxima verosimilitud para el contenido de arcilla utilizando los modelos $H_1, H_2, H_3, H_4, H_5, H_6, H_7$ y $H_8$ . . . . .	122
4.4	Log-verosimilitudes para el contenido de arcilla utilizando diferentes funciones de enlace en el SMM y el SVDM con función de correlación esférica en ambos modelos. . . . .	123
4.5	Estimaciones e intervalos de confianza de 95 % para los parámetros involucrados en el BSLMM para ajustar el contenido de arcilla. . . . .	124
4.6	Estimaciones beta espaciales por máxima verosimilitud para el contenido de magnesio utilizando los modelos $H_1, H_2, H_3, H_4, H_5, H_6, H_7$ and $H_8$ . . . . .	127
4.7	Log-verosimilitudes para el contenido de magnesio utilizando diferentes funciones de enlace en el SMM y el SVDM con funciones de correlación gaussiana y esférica, respectivamente. . . . .	127
4.8	Estimaciones e intervalos de confianza de 95 % para los parámetros involucrados en el ajuste de contenido de magnesio en el BSLMM. . . . .	128
5.1	Características de algunas distribuciones univariadas comunes en la familia exponencial como las dadas en (5.2) o (6.5) . . . .	142
5.2	Estimación de los parámetros del modelo GLSTARAR para el error espacial mediante GEE. . . . .	171
5.3	Estimación de los parámetros del modelo GLSTARAR para el rezago espacial mediante GEE. . . . .	173
5.4	Estimación de los parámetros del modelo GLSTARAR para el rezago espacial mediante GEE (Continuación Tabla 5.3). . . . .	174
5.5	Estimación de los parámetros del modelo DBGLSTARAR para el error espacial mediante GEE. . . . .	176
5.6	Estimación de los parámetros del modelo DBGLSTARAR para el rezago espacial mediante GEE. . . . .	177

# Abreviaturas

<b>DB:</b>	Distance-Based (basado en distancias)
<b>GLM:</b>	Generalised linear model (modelo lineal generalizado)
<b>OLS:</b>	Ordinary least squares (mínimos cuadrados ordinarios)
<b>GLS:</b>	Generalized least squares (mínimos cuadrados generalizados)
<b>WLS:</b>	Weighted least squares (mínimos cuadrados ponderados)
<b>DBBR:</b>	Distance-based beta regression (regresión beta basado en distancias)
<b>FT:</b>	Fund type (tipo de fondo)
<b>DE:</b>	Domestic equity (renta variable nacional)
<b>IE:</b>	International equity (renta variable internacional)
<b>NAV:</b>	Net asset value (patrimonio neto)
<b>FYAR:</b>	5-year average return (rendimiento promedio durante 5 años)
<b>ER:</b>	Expense ratio (índice de gasto)
<b>MR:</b>	Morningstar rank (clasificación de Morningstar)
<b>MCAR:</b>	Missing completely at random (faltantes de forma completamente aleatoria)
<b>MSE:</b>	Mean squared error (error cuadrático medio)
<b>GLMM:</b>	Generalised linear mixed model (modelo lineal generalizado mixto)
<b>MCML:</b>	Monte Carlo maximum likelihood (máxima verosimilitud vía Monte Carlo)
<b>MCMC:</b>	Markov chain Monte Carlo (Monte Carlo vía cadenas de Markov)
<b>DBSGLMM:</b>	Distance-based spatial generalised linear mixed models (modelos lineales generalizados espaciales mixtos basados en distancias)
<b>MCEMG:</b>	Monte Carlo expectation-maximization gradient (gradiente Monte Carlo esperanza-maximización)
<b>ELE:</b>	Elevation (elevación o altura)
<b>NDVI:</b>	Normalised difference vegetation index (índice de vegetación de diferencia normalizada)
<b>EM:</b>	Expectation-maximization (esperanza-maximización)

---

<b>MLE:</b>	Maximum likelihood estimate (estimación de máxima verosimilitud)
<b>EMG:</b>	Expectation-maximization gradient (gradiente esperanza-maximización)
<b>MMSE:</b>	Minimum mean square error (mínimo error cuadrático medio)
<b>BPLUP:</b>	Best pseudo linear unbiased predictor (mejor pseudo predictor lineal insesgado)
<b>IRD:</b>	Institut de Recherche pour le Développement
<b>TCP:</b>	Tropenbos Cameroon Programme
<b>ZONE:</b>	Agro-ecological zone (zona agro-ecológica)
<b>WRB:</b>	World Reference Base
<b>UTM:</b>	Universal Transverse Mercator
<b>ALT:</b>	Altitude (altitud)
<b>SR:</b>	Sub-region (sub-región)
<b>SGLMM:</b>	Spatial generalised linear mixed models (modelos lineales generalizados espaciales mixtos)
<b>BSLMM:</b>	Beta spatial linear mixed model (modelo lineal beta espacial mixto)
<b>SMM:</b>	Spatial mean model (modelo espacial de media)
<b>SVDM:</b>	Spatial variable dispersion model (modelo espacial de dispersión variable)
<b>GEEs:</b>	Generalised estimating equations (ecuaciones de estimación generalizada)
<b>ML:</b>	Maximum likelihood (máxima verosimilitud)
<b>QSTIC:</b>	Quasi-likelihood space-time under the independence model criterion (cuasi-verosimilitud espacio-tiempo bajo el criterio de modelo de independencia)
<b>GLSTARAR:</b>	Generalised linear space-time-autoregressive models with space-time-autoregressive disturbances (modelo lineal generalizado autorregresivo espacio-tiempo con perturbación autorregresiva espacio-tiempo)
<b>SAC:</b>	Spatial autocorrelation (autocorrelación espacial)
<b>SGVARs:</b>	Spatial generalised vector autoregressions (vectores autorregresivos espaciales generalizados)
<b>DBGLMVAR:</b>	Distance-based generalised linear mixed vector autoregressive disturbances (vectores autorregresivos generalizados con perturbaciones basados en distancias)
<b>DBGCS:</b>	Distance-based spatial generalised cross sectional (transversal espacial generalizado basado en distancias)

---

<b>DBGLSTARAR:</b>	Distance-based generalised linear space-time-autoregressive models with space-time-autoregressive disturbances (modelos lineales generalizados autorregresivos espacio-tiempo basados en distancias con perturbaciones autorregresivas espacio-tiempo)
<b>DBSGLMVAR:</b>	Distance-based spatial generalised linear mixed vector autoregression (vector autorregresivo lineal generalizado espacial mixto basado en distancias)
<b>DBGLMVAR:</b>	Distance-based generalised linear mixed vector autoregression (vector autorregresivo lineal generalizado basado en distancia)
<b>AA:</b>	Acciones armadas
<b>AVV:</b>	Atención a víctimas de la violencia
<b>SMMLV:</b>	Salarios mínimos mensuales legales vigentes
<b>DFHR:</b>	desplazamiento forzado - hogares recibidos
<b>CAA:</b>	Confrontaciones armadas por año
<b>FM:</b>	fuerzas militares
<b>URB:</b>	Urbana
<b>SL:</b>	Spatial lag (rezago espacial)





# Introducción

En función del tipo de variable estocástica a analizar y del tipo de localización de los datos (puntos o áreas), deberá utilizarse una metodología distinta de análisis. Los datos espaciales pueden clasificarse en tres grandes grupos:

1. *Datos en rejilla.* Observaciones procedentes de un proceso aleatorio, observadas sobre una colección contable de regiones espaciales, regular o irregularmente distribuidas, complementados con lo que se denomina “estructura de vecindad”. Como ejemplos de éstos se tienen: recuento de algún tipo de casos de enfermedad en los municipios de una región determinada, número de frutos por árbol en una región y recuento del número de accidentes por tramo de carretera.
2. *Procesos puntuales.* Cuando las localizaciones (y no las mediciones) son las variables de interés. Consisten en un número finito de localizaciones observadas en una región determinada. Como ejemplos de éstos se tienen: domicilio de las personas que desarrollan un determinado tumor, determinada especie de árboles en un bosque, coordenadas de los epicentros de terremotos y posición de estrellas en el cielo.
3. *Datos geoestadísticos.* Son mediciones tomadas en puntos fijos pero definidas en cualquier lugar del espacio por lo que sus localizaciones definen una superficie espacialmente continua. Como ejemplos de éstos se tienen: concentraciones de mineral en lugares concretos dentro de una mina, lluvia medida en estaciones meteorológicas, concentraciones de contaminantes medidas en estaciones de control y valores de temperatura medidas en estaciones meteorológicas.

El término estadística espacial se utiliza para describir una amplia gama de modelos estadísticos y métodos destinados al análisis de datos espacialmente referenciados (Diggle et al. 2007, Cressie 1993). La metodología geoestadística convencional resuelve el problema de predecir el valor realizado de un funcional lineal en un proceso estocástico espacial gaussiano, basado en observaciones  $y(\mathbf{s}_i) = v(\mathbf{s}_i) + z(\mathbf{s}_i)$  de un conjunto discreto en localizaciones de muestreo

$\mathbf{s}_i$  dentro de una región espacial  $A$ , donde las  $z(\mathbf{s}_i)$  son variables aleatorias mutuamente independientes, de media cero.

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 del análisis geoestadístico (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 et al. 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.

Es así como 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, el método denominado kriging universal no asume una media constante y es con frecuencia usado en los procesos no estacionarios (Wackernagel 2003).

Diggle et al. (1998) introducen el término modelo basado en geoestadística en el sentido de la aplicación de los principios generales del modelamiento estadístico y la inferencia a los problemas de geoestadística. En particular, agregaron supuestos de distribución gaussiana al modelo clásico  $y(\mathbf{s}_i) = v(\mathbf{s}_i) + z(\mathbf{s}_i)$  y re-expresaron éste como un modelo lineal jerárquico de dos niveles, en el cual  $v(\mathbf{s}_i) + z(\mathbf{s}_i)$  es el valor en la posición  $\mathbf{s}_i$  de un proceso estocástico gaussiano latente y, al condicionar sobre  $v(\mathbf{s}_i) + z(\mathbf{s}_i) : i = 1, \dots, n$ , los valores  $y(\mathbf{s}_i) : i = 1, \dots, n$  son mutuamente independientes, normalmente distribuidos con media  $\mu + v(\mathbf{s}_i)$  y varianza común  $\sigma^2$ . Además, Diggle et al. (1998) extienden este modelo, manteniendo la hipótesis de gaussianidad de  $v(\mathbf{s}_i) + z(\mathbf{s}_i)$ , permitiendo un modelo lineal generalizado (Nelder & Wedderburn 1972, McCullagh & Nelder 1989, McCullagh 2008) para distribuciones condicionales mutuamente independientes de los  $y(\mathbf{s}_i)$  dado  $v(\mathbf{s}_i) + z(\mathbf{s}_i)$ .

De acuerdo a lo anterior, algunos autores han utilizado estos conceptos, por ejemplo en el área de epidemiología espacial, en la cual se busca describir estudios sobre las causas y la prevención de las enfermedades utilizando dife-

rentes perspectivas de análisis en las que la localización de los eventos es un componente fundamental. En esta tesis en particular se hace lo mismo teniendo en cuenta adicionalmente el tiempo, utilizando el método basado en distancias (Cuadras 1989) en el modelamiento del problema. El objetivo de esta clase de datos es mostrar qué parte de la variación espacial o espacio-temporal de la distribución de la ocurrencia de una enfermedad no está explicada ni por la distribución espacial o espacio-temporal de factores explicativos conocidos ni por una variación aleatoria.

En la epidemiología espacial o espacio-temporal, por lo general se hacen estudios en áreas pequeñas ya que se estudian fenómenos poco frecuentes. Ésta puede ser abordada desde tres grandes perspectivas: a) mapas de enfermedades, b) estudios de asociación geográfica y c) aglomeraciones de casos o clustering.

La idea de utilizar este tipo de análisis es proporcionar un rápido resumen visual de información geográfica compleja e identificar patrones en los datos que de otro modo podrían pasar inadvertidos en las presentaciones tabulares. Además en esta tesis se utiliza este tipo de análisis con propósitos descriptivos con el objetivo de generar hipótesis etiológicas, para la vigilancia epidemiológica y/o sociológica, según sea el caso de estudio, con el fin de detectar áreas con un aparente mayor riesgo, como ayuda en la definición de políticas sociales y de asignación de recursos y, para localizar clusters específicos.

En este análisis se utiliza habitualmente los datos en rejilla. Cuando se presentan datos espaciales o espacio-temporales en rejilla, se utilizan modelos estadísticos apropiados para suavizar los estadísticos de resumen (habitualmente razones estandarizadas). La suavización (reducción de la extra-variabilidad) se consigue introduciendo efectos aleatorios, uno recoge la heterogeneidad y el otro la dependencia espacial o espacio-temporal. El modelo puede permitir variables explicativas (con efectos fijos y/o aleatorios), otras estructuras de dependencia, otras distribuciones de probabilidad y/o la dimensión temporal.

El tipo de datos espaciales o espacio-temporales dependen de las unidades de medida adoptadas y varía considerablemente en función del contexto de aplicación. Por ejemplo, en aplicaciones de mapeo de una enfermedad es posible obtener la dirección residencial y la fecha de diagnóstico de un caso. En este caso, una realización completa de un proceso de punto espacio-temporal es observada. Sin embargo, en otros casos, solamente se cuentan los eventos disponibles dentro de períodos espacio-temporales fijos, o las mediciones en los sitios de monitoreo espacial se realizan en intervalos de tiempo fijos como es el caso del estudio presentado en el Capítulo 5.

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 (distance-based, DB) para la predicción de variables mixtas (continuas y categóricas) y exploran el problema de información faltante dando una solución utilizando DB. Algunos de los trabajos más recientes son los de Esteve et al. (2009) y Boj et al. (2010) quienes proponen algunas alternativas invariantes de los métodos de distancias para el modelamiento de variable respuestas continuas y no continuas.

En términos generales muchos métodos en la estadística se basan en el cálculo de distancias geométricas, 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, son determinantes en la generación de pronósticos y permite 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 son 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, y covariables regionalizadas continuas, categóricas y binarias. La selección de variables explicativas se hace a partir de técnicas muy populares del análisis de regresión tradicional. Sin embargo, aquí se presentan algunas alternativas a partir de la propuesta de Cuadras et al. (1996) para seleccionar las coordenadas 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 lineal mixto generalizado geoestadístico, ya que también el método basado en distancias es robusto ante los errores de especificación en la correlación de los parámetros.

En esta tesis se hace una mezcla del método de distancias (Cuadras & Arenas 1990) con los modelos lineales generalizados mixtos tanto en lo espacial como en lo espacio-temporal. Con el empleo de las distancias se logran buenas predicciones y menores variabilidades en el espacio o espacio-tiempo de la región de estudio, provocando todo esto que se tomen mejores decisiones en los diferentes problemas de interés.

Por otro lado, la variabilidad de un proceso juega un papel determinante al momento de ajustar un modelo, por ser un elemento primordial para la toma de decisiones. Ya sea un modelo homocedástico o heterocedástico, si hay diferencias considerables entre la dispersión real y la estimación de la misma arrojada por el modelo, éste será inútil al aplicarse en la práctica. Se generan así problemas de sobredispersión (variabilidad del modelo mayor que la real) o subdispersión (variabilidad del modelo menor que la real). Los factores asociados comúnmente a los problemas de sobredispersión o subdispersión son, entre otros, la variabilidad del material experimental, la correlación entre respuestas, los datos atípicos e influyentes, la mezcla de distribuciones, el muestreo por conglomerados y la omisión de variables. Aunque, en algunos casos, es casi imposible determinar el origen de la sobredispersión o la subdispersión, ya que puede deberse a la naturaleza misma de los datos (Hinde & Demétrio 1998, Jowaheer & Sutradhar 2002, McCullagh 2008).

La sobredispersión o subdispersión del modelo afecta la significancia de los parámetros, debido a que perturba la estimación del error estándar, originando interpretaciones erradas de los efectos de las variables explicativas sobre la respuesta, generando predicciones e inferencias equivocadas. En esta tesis, también se modela una variable respuesta espacio o espacio-temporal tipo binomial o poisson con problemas de sobredispersión o subdispersión a través del modelo lineal generalizado (generalised linear model, GLM) utilizando el método de distancias con diferentes funciones de enlace. Básicamente la idea es involucrar los problemas de variación de manera directa en el modelo ajustado, utilizando la función de cuasi-verosimilitud o de ecuaciones de estimación generalizada, con el fin de ajustar apropiadamente las predicciones que se hacen con el modelo ajustado.

Adicionalmente, cuando se tienen tasas de mortalidad medidas en las diferentes localizaciones de un país e inclusive cuando se considera la evolución de las mismas a través del tiempo, los modelos espaciales o espacio-temporales clásicos no son apropiados ya que la respuesta está restringida al intervalo  $(a, b)$  o  $(0, 1)$ . Los métodos de estimación como mínimos cuadrados ponderados o máxima verosimilitud generalizada pueden generar valores ajustados que excedan dichas cotas inferior y superior. En este caso, como se muestra en Kieschnick & McCullough (2003), Ferrari & Cribari-Neto (2004) y Vasconcellos & Cribari-Neto (2005), una posible solución es transformar la variable dependiente para asumir que ésta toma valores sobre la recta real y luego modelar la media de la respuesta transformada como un predictor lineal basado en un conjunto de variables exógenas.

De acuerdo a Ferrari & Cribari-Neto (2004), el modelo de regresión propuesto es generado para situaciones donde la variable respuesta  $y$  es continua y definida sobre el intervalo unitario estandarizado  $0 < y(\mathbf{s}_i) < 1$  y, la estructura de regresión involucra regresores y parámetros desconocidos. Ospina

et al. (2006) obtiene el sesgo de segundo orden de los estimadores de máxima verosimilitud y los utiliza para definir los estimadores de sesgo ajustado, los cuales son muy útiles para solucionar el problema en muestras pequeñas. Una variante del modelo de regresión beta que permite modelar la no linealidad y la variable de dispersión fue propuesta por Simas et al. (2010). En particular, en este modelo más general, el parámetro que representa la precisión de los datos no se supone que es constante a través de observaciones, sino que puede variar, lo que conduce al *modelo de regresión beta de dispersión variable*.

En esta tesis se mezclan los modelos de regresión beta espacial o espacio-temporal con el método de distancias. El modelamiento y los procedimientos inferenciales propuestos son similares a los de los modelos lineales generalizados obtenidos para el caso de una variable respuesta tipo binomial. Aunque la respuesta no es miembro de la familia exponencial, se hace una adaptación a esta familia siguiendo las propuestas realizadas en modelos generalizados por McCulloch & Searle (2001), Lee & Nelder (2002), Dobson (2002) y Smith & Ridout (2003).

Recientemente los métodos espaciales y espacio-tiempo cada vez se han aplicado más en una amplia gama de investigaciones empíricas en los campos más tradicionales de la economía y las ciencias sociales, incluyendo entre otros, estudios en el análisis de la demanda, el crecimiento económico, economía internacional, mercado laboral, índices de empleo, el desplazamiento por violencia armada, economía pública, finanzas públicas locales, producción agrícola y contaminación ambiental. Muchos de éstos estudios establecen la importancia de integrar rezagos espaciales y temporales en el análisis de datos panel cuando la variable respuesta no es normal. Sin embargo, la literatura en modelos con dinámica espacial y temporal solo han presentado algunos progresos al tratar con esta clase de variable respuesta, pero en muchos casos por separado. En esta tesis, se presenta una solución a problemas donde la variable respuesta es un conteo, una razón o una respuesta binaria (dicotómica) utilizando modelos lineales generalizados autorregresivos espacio-tiempo basados en distancias con perturbaciones autorregresivas espacio-tiempo.

Por lo tanto, en esta tesis se proponen métodos alternos de interpolación espacial o espacio-temporal con variables explicativas mixtas utilizando distancias entre individuos, tales como la distancia de Gower (Gower 1968); aunque, algunas otras distancias euclidianas se pueden utilizar. El método DB se utiliza en los modelos lineales mixtos generalizados geoestadísticos no sólo en la etapa de estimación de la tendencia, sino también en la etapa de estimación de la correlación espacial o espacio-temporal, cuando las variables explicativas son mixtas. En este caso, el método DB espacial o espacio-temporal propuesto se basa en 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.

En todos los métodos de predicción presentados en esta tesis, las coordenadas principales obtenidas mediante el método de distancias se obtienen a partir 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 sobre los parámetros 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, es de resaltar que todos los procesos de análisis de las diferentes aplicaciones presentadas en cada uno de los capítulos fueron desarrolladas en programa estadístico R Development Core Team (2013); los diferentes códigos para cada una de las aplicaciones y sus respectivas funciones se anexan en el CD adjunto.

Este trabajo lo he dividido en siete 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. El capítulo termina con una corta exposición de algunos conceptos básicos sobre distancias euclidianas muy útiles cuando se tienen variables explicativas continuas, categóricas, binarias, e inclusive una mezcla de todas las anteriores.

**Capítulo 2.** Se propone un método alternativo para ajustar una variable respuesta tipo beta con dispersión variable usando distancias euclidianas entre los individuos. Se emplea el método de máxima verosimilitud para estimar los parámetros desconocidos del modelo propuesto y se presentan las principales propiedades de estos estimadores. Además, se realiza la inferencia estadística sobre los parámetros utilizando las aproximaciones obtenidas a partir de la normalidad asintótica del estimador de máxima verosimilitud; se desarrolla el diagnóstico y predicción de una nueva observación, y se estudia el problema de datos faltantes utilizando la metodología propuesta.

**Capítulo 3.** Se propone una solución alterna para resolver problemas como el de prevalencia de Loa loa utilizando distancias euclidianas entre individuos; se describe un modelo lineal generalizado espacial mixto incorporando medidas generales de distancia o disimilaridad que se pueden aplicar a variables explicativas: numéricas, categóricas o una mezcla de ellas. Los parámetros involucrados en el modelo propuesto se estiman utilizando máxima verosimilitud mediante el método de Monte Carlo vía cadenas de Markov, la cual es una técnica muy útil para el análisis de este tipo de información.



**Capítulo 4.** Se propone un modelo lineal beta espacial mixto con dispersión variable utilizando máxima verosimilitud mediante el método de Monte Carlo vía cadenas de Markov. El método propuesto se utiliza en situaciones donde la variable respuesta es una razón o proporción que esta relacionada con determinadas variables explicativas. Para este fin, se desarrolla una aproximación utilizando modelos lineales generalizados espaciales mixtos empleando la transformación Box-Cox en el modelo de precisión. Por lo tanto, se realiza el proceso de optimización de los parámetros tanto para modelo espacial de media como para el modelo espacial de dispersión variable. Además, se realiza la inferencia estadística sobre los parámetros utilizando las aproximaciones obtenidas a partir de la normalidad asintótica del estimador de máxima verosimilitud. También se desarrolla el diagnóstico del modelo y la predicción de nuevas observaciones. Por último, el método se ilustra a través de los contenidos de arcilla y magnesio.

**Capítulo 5.** Se describe el modelo propuesto basado en distancias para la predicción espacio-temporal usando modelos lineales generalizados. Utilizando el modelo propuesto se realiza: el proceso de estimación de los parámetros involucrados en el modelo propuesto mediante el método de ecuaciones de estimación generalizada y la inferencia estadística sobre los parámetros empleando las aproximaciones obtenidas a partir de la normalidad asintótica del estimador de máxima verosimilitud. Además, se desarrolla el diagnóstico del modelo y la predicción de nuevas observaciones. Ese capítulo finaliza con una aplicación de la metodología propuesta para el número de acciones armadas estandarizada por cada  $1000 \text{ km}^2$  de los grupos irregulares Fuerzas Armadas Revolucionarias de Colombia - Ejército del Pueblo (FARC-EP) y Ejército de Liberación Nacional (ELN) en los diferentes departamentos de Colombia entre los años 2003 a 2009.

**Capítulo 6.** Se presenta un modelo autorregresivo espacial lineal generalizado mixto utilizando el método basado en distancias propuesto por Cuadras (1989). Este modelo incluye retrasos tanto espaciales como temporales entre vectores de variables de estado estacionarias. Se utilizó la dinámica espacial de los datos econométricos tipo panel para estimar el modelo propuesto; los parámetros involucrados en el modelo se estiman utilizando el método MCMC mediante máxima verosimilitud. Además, se discute en este capítulo la interacción entre estacionariedad temporal y espacial, y se derivan las respuestas al impulso para el modelo propuesto, lo cual naturalmente depende de la dinámica temporal y espacial del modelo.

**Capítulo 7** En este último capítulo se resumen las principales aportaciones de la presente tesis y se realizan algunas recomendaciones.

# Objetivos

## Objetivos Generales

Proponer una metodología a través del método basado en distancias utilizando modelos lineales generalizados para generar innovaciones en el espacio y espacio-tiempo.

## Objetivos Específicos

- Proponer una metodología para ajustar modelos con variable respuesta tipo beta por medio del método basado en distancias, utilizando procedimientos inferenciales similares a los propuestos en los modelos lineales generalizados.
- Generar un nuevo método basado en distancias para modelar variables aleatorias no continuas a través del modelo lineal generalizado espacial o espacio-temporal.
- Proponer una metodología para ajustar modelos espaciales con variable respuesta tipo beta utilizando procedimientos inferenciales similares a los propuestos en los modelos lineales generalizados mixtos.
- Analizar las bondades y limitantes de las metodologías propuestas al compararlas con otras existentes.
- Aplicar los métodos propuestos a casos reales de ciencias de la tierra, casos epidemiológicos, casos de economía clásica con respuesta de distribución beta y casos de economía espacio-temporal.



# Capítulo 1

## Conceptos básicos

### 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 et al. 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.

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 sí los métodos geoestadísticos se basan en el cálculo de distancias geométricas, 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 como se verán en los capítulos posteriores.

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. En estos últimos se definen los principales conceptos involucrados en su estudio y sus propiedades, así como un par de procedimientos para el ajuste del variograma o semivariograma.

La Sección 1.2 introduce los conceptos más relevantes que se utilizarán 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 la Sección 1.5 se profundiza en la estimación de los parámetros del variograma utilizando mínimos cuadrados. Finalmente, en la Sección 1.6 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.

## 1.2 Análisis geoestadístico clásico

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  $Y(\mathbf{s})$  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  $\{Y(\mathbf{s}), \mathbf{s} \in D\}$ , 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  $Y(\mathbf{s})$  podría representar, por ejemplo, el número de personas enfermas 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

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

**Definición 1.1.** Sea  $Y(\mathbf{s})$  un proceso estocástico de segundo orden. Se define su función de covarianza como

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

Generalmente en la práctica sólo se dispone de un conjunto  $\{y(\mathbf{s}_1), \dots, y(\mathbf{s}_n)\}$  de observaciones del proceso aleatorio  $\{Y(\mathbf{s}), \mathbf{s} \in D\}$  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 1.2.** Se dice que el proceso  $Y(\mathbf{s})$  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  $\{Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n)\}$  permanece invariable ante una traslación. Sea  $F_{\mathbf{s}_1, \dots, \mathbf{s}_n}(y(\mathbf{s}_1), \dots, y(\mathbf{s}_n)) = P(Y(\mathbf{s}_1) \leq y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n) \leq y(\mathbf{s}_n))$  la función de distribución conjunta, entonces se cumple que

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

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 1.3.** Se dice que un proceso espacial  $Y(\mathbf{s})$  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

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

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

**Definición 1.4.** Se dice que el proceso  $Y(\mathbf{s})$  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,  $\text{Var}(Y(\mathbf{s}_i) - Y(\mathbf{s}_j)) = 2\gamma(\mathbf{h}), \forall \mathbf{s}_i, \mathbf{s}_j \in D$ , 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  $Y(\mathbf{s})$  de segundo orden con covariograma  $C(\cdot)$ , entonces

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

por lo que el proceso  $Y(\mathbf{s})$  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 1.5.** *Se dice que el proceso  $Y(\mathbf{s})$  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 1.6.** *Se dice que el proceso  $Y(\mathbf{s})$  es homogéneo si es intrínsecamente estacionario e isotrópico.*

Si  $Y(\mathbf{s})$  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  $Y(\cdot)$  es anisotrópico, la dependencia entre  $Y(\mathbf{s})$  y  $Y(\mathbf{s} + \mathbf{h})$  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 1.7.** *Se dice que el proceso  $Y(\mathbf{s})$  tiene anisotropía geométrica si su variograma es de la forma*

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

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

De otro lado, se tiene que dados  $Y_1(\cdot), \dots, Y_n(\cdot)$ ,  $n$  procesos intrínsecamente estacionarios independientes, entonces  $Y_1(\cdot) + \dots + Y_n(\cdot)$  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  $Y_i(\cdot)$ . Esta propiedad permite definir la siguiente generalización de la anisotropía geométrica.

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

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

siendo  $\gamma_0$  un semivariograma isotrópico y  $\mathbf{A}_1, \dots, \mathbf{A}_n$  matrices  $d \times d$ .



Otro tipo de tratamiento de la anisotropía es la de suponer que, dado el proceso original  $Y(\mathbf{s})$ , existe una función no lineal  $f^*(\mathbf{s})$ , de forma que el proceso  $Y(f^*(\mathbf{s}))$  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  $Y(\cdot)$ , se ha definido su covariograma como

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

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

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

### 1.2.3 El variograma

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

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

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

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

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  $Y(\cdot)$  un proceso estacionario de segundo orden con función de covarianza  $C(\cdot)$ . Se tiene que  $C(\mathbf{0}) = \text{Cov}(Y(\mathbf{s}), Y(\mathbf{s})) = \text{Var}(Y(\mathbf{s}))$ , por lo que  $C(\mathbf{0}) > 0$

a no ser que  $Y(\cdot)$  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  $\lim_{h \rightarrow 0} \gamma(\mathbf{h}) = c_0 > 0$ , 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  $\{y(\mathbf{s}_i), i = 1, \dots, n\}$ , 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  $Y(\cdot)$  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.

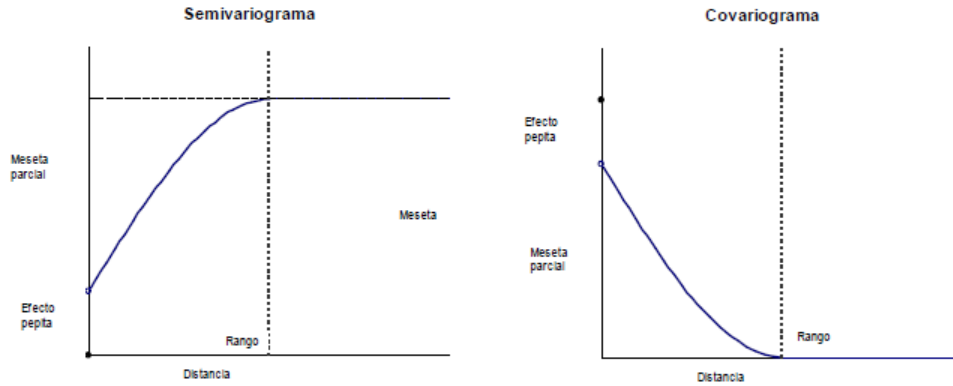


FIGURA 1.1: Forma general del variograma y covariograma de un proceso espacial homogéneo.

### 1.3 Estimación del variograma y del covariograma

Dado un proceso espacial  $Y(\cdot)$  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).

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

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

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

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

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

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

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

## 1.4 Principales modelos de variogramas y covariogramas isotrópicos

En la sección anterior se vio un estimador del variograma o covariograma de los procesos espaciales. El problema es que esta estimación no se puede utilizar directamente en la práctica geoestadística, ya que no satisface en general la condición de ser condicionalmente definida negativa (o definida positiva para el covariograma); condición que debe verificar el variograma. 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 semi-variograma. 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  $h$  entre localizaciones únicamente por su módulo  $h = \|\mathbf{h}\|$ ), ya que son el punto de arranque sobre los que se construyen modelos más complejos.

TABLA 1.1: Formas funcionales de algunos variogramas.

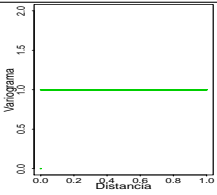
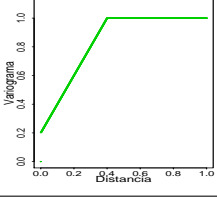
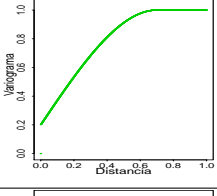
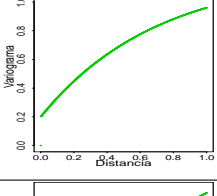
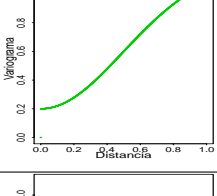
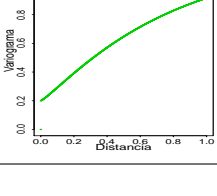
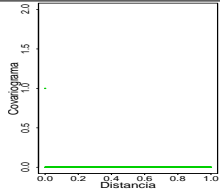
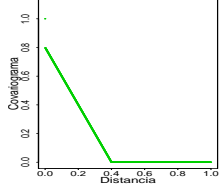
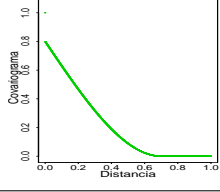
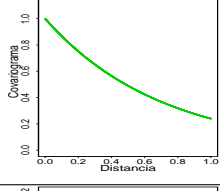
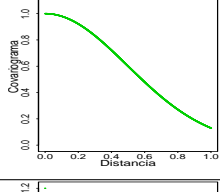
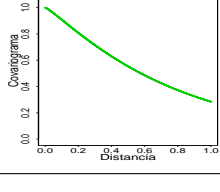
Modelo	Función de Semivariograma	Variograma
Efecto pepita	$\gamma(h) = \begin{cases} c_0 & \text{si } h > 0 \\ 0 & \text{si } h = 0 \end{cases}$	
Lineal con meseta	$\gamma(h) = \begin{cases} 0 & \text{si } h = 0 \\ c_0 + c_s \left(\frac{h}{a_l}\right) & \text{si } 0 < h \leq a_l \\ c_0 + c_s & \text{si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

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	$C(h) = \begin{cases} c_0 + c_s & \text{si } h = 0 \\ c_s \left(1 - \frac{h}{a_l}\right) & \text{si } 0 < h \leq a_l \\ 0 & \text{si } h > a_l \end{cases}$	
Esférico	$C(h) = \begin{cases} c_0 + c_s & \text{si } h = 0 \\ c_s \left(1 - \frac{3}{2}\left(\frac{h}{a_s}\right) + \frac{1}{2}\left(\frac{h}{a_s}\right)^3\right) & \text{si } 0 < h \leq a_s \\ 0 & \text{si } h > a_s \end{cases}$	
Exponencial	$C(h) = \begin{cases} c_s \left(\exp(-3h/a_e)\right) & \text{si } h > 0 \\ c_0 + c_s & \text{si } h = 0 \end{cases}$	
Gaussiano	$C(h) = \begin{cases} c_s \left(\exp(-3h^2/a_g^2)\right) & \text{si } h > 0 \\ c_0 + c_s & \text{si } h = 0 \end{cases}$	
Circular	$C(h) = \begin{cases} c_0 + c_s & \text{si } h = 0 \\ c_1 \left(\frac{2}{\pi} \cos^{-1}\left(\frac{h}{a_c}\right) + \frac{2h}{\pi a_c} \sqrt{\left(1 - \frac{h}{a_c}\right)^2}\right) & \text{si } 0 < h \leq a_c \\ 0 & \text{si } h > a_c \end{cases}$	

## 1.5 Estimación de los parámetros del variograma

Sea  $Y(\cdot)$  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  $\gamma(\mathbf{h}, \boldsymbol{\theta})$  de los que se han presentado anteriormente que se aproxime a la dependencia espacial encontrada 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  $\{\gamma(\mathbf{h}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$  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 (ordinary least squares, OLS) consiste en obtener el valor  $\hat{\boldsymbol{\theta}}$  que minimiza

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

siendo  $\hat{\boldsymbol{\gamma}} = [\hat{\gamma}(\mathbf{h}_1), \dots, \hat{\gamma}(\mathbf{h}_n)]^t$  y  $\boldsymbol{\gamma}(\boldsymbol{\theta}) = [\gamma(\mathbf{h}_1, \boldsymbol{\theta}), \dots, \gamma(\mathbf{h}_n, \boldsymbol{\theta})]^t$ . 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 (generalized least squares, GLS), que consiste en minimizar

$$[\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})]^t V^{-1}(\boldsymbol{\theta}) [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})]$$

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

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

$$\sum_{j=1}^n w_j [\hat{\gamma}(\mathbf{h}_j) - \gamma(\mathbf{h}_j, \boldsymbol{\theta})]^2 = [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})]^t \mathbf{W}^{-1}(\boldsymbol{\theta}) [\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})] \quad (1.6)$$

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



localizaciones a distancia  $\mathbf{h}_j$ . Por tanto, los pesos de (1.6) vendrán dados por  $w_j = N(\mathbf{h}_j)/(2\gamma(\mathbf{h}_j, \boldsymbol{\theta})^2)$ .

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.

## 1.6 Distancia, similaridad y descomposición espectral

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.

El concepto de distancia entre objetos o individuos permite interpretar geoméricamente muchas técnicas clásicas del análisis multivariante, equivalentes a representar estos objetos como puntos de un espacio métrico adecuado. Esta interpretación es posible no solamente cuando se dispone de variables cuantitativas, sino también cuando las variables observadas son cualitativas o mixtas (cualitativas y cuantitativas), siempre que tenga sentido obtener una medida de proximidad entre los objetos o individuos.

**Definición 1.9.** *Una distancia  $d$  sobre un conjunto (finito o no)  $\Omega$  es una aplicación que a cada par de individuos  $(\omega_i, \omega_{i'}) \in \Omega \times \Omega$ , le hace corresponder un número real  $d(\omega_i, \omega_{i'}) = d_{ii'}$ , que cumple con las siguientes propiedades básicas:*

- i.  $d_{ii'} \geq 0$ .
- ii.  $d_{ii} = 0$ .
- iii.  $d_{ii'} = d_{j'i}$ .
- iv.  $d_{ii'} \leq d_{ij} + d_{j'i'}$ .

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

**Definición 1.10.** Si  $\Omega$  es un conjunto finito, que se indica por  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ , las distancias  $d_{ii'}$  se expresan mediante la matriz simétrica  $\mathbf{D}$ , llamada matriz de distancias sobre  $\Omega$ .

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

con  $d_{ii} = 0$ ,  $d_{ii'} = d_{i'i}$ . 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:

$$d_{i_1 i'_1} \leq d_{i_2 i'_2} \leq \cdots \leq d_{i_q i'_q}$$

es decir, la ordenación de los pares  $(\omega_i, \omega_{i'})$  de  $\Omega$  de acuerdo con su proximidad.

Una matriz de distancias  $\mathbf{D}$  puede ser transformada de diversos modos. Por ejemplo:

$$\tilde{d}_{ii'} = \begin{cases} 0 & \text{si } i = i' \\ d_{ii'} + c & \text{si } i \neq i' \end{cases} \quad (1.7)$$

La transformación (1.7), consiste en sumar una constante fuera de la diagonal de  $\mathbf{D}$ , 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 1.11.** Una similaridad  $m$  en un conjunto  $\Omega$ , es una aplicación que asigna a cada par  $(\omega_i, \omega_{i'}) \in \Omega \times \Omega$  un número real  $m_{ii'} = m(i, i')$ , que cumple:

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

ii.  $m_{ii'} = m_{i'i}$ .

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

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

se denomina matriz de similaridades sobre  $\Omega$ .

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

$$d_{ii'} = 1 - m_{ii'} \quad (1.8)$$

y

$$d_{ii'} = \sqrt{1 - m_{ii'}} \quad (1.9)$$

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

$$d_{ii'} = \sqrt{m_{ii} + m_{i'i} - 2m_{ii'}} \quad (1.10)$$

Por diversas razones (1.9) es preferible a (1.8). Pero en general, (1.10) 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  $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_p$ , donde cada  $\mathbf{V}_j$  ( $j = 1, \dots, p$ ) 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  $\omega_i, \omega_{i'}$ .

$$\begin{aligned} m_{ii'} &= \frac{c_{1ii'} + c_{4ii'}}{c_{1ii'} + c_{2ii'} + c_{3ii'} + c_{4ii'}} && (\text{Sokal} - \text{Michener}) \\ m_{ii'} &= \frac{c_{1ii'}}{c_{1ii'} + c_{2ii'} + c_{3ii'}} && (\text{Jaccard}) \end{aligned} \quad (1.11)$$

donde  $c_{1ii'}$ ,  $c_{2ii'}$ ,  $c_{3ii'}$ ,  $c_{4ii'}$  son las frecuencias de (1, 1), (1, 0), (0, 1) y (0, 0), respectivamente. Note que  $p = c_{1ii'} + c_{2ii'} + c_{3ii'} + c_{4ii'}$ . Estas similaridades pueden ser transformadas en distancias utilizando (1.8) o (1.9).

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

$$m_{ii'} = \frac{\sum_{j=1}^{p_c} \left(1 - \frac{|v_{ij} - v_{i'j}|}{G_j}\right) + c_{1ii'} + v_{ii'}}{p_c + (p_b - c_{4ii'}) + p_q}, \quad i, i' = 1, \dots, n \quad (1.12)$$

donde  $G_j$  es el rango de la  $j$ -ésima variable cuantitativa,  $p_c$  es el número de variables cuantitativas,  $c_{1ii'}$  y  $c_{4ii'}$  corresponden al número de coincidencias y no coincidencias para las  $p_b$  variables binarias, respectivamente, y  $v_{ii'}$  es el número de coincidencias para las  $p_q$  variables cualitativas. Este coeficiente admite la posibilidad de tratar datos faltantes y se reduce al coeficiente de Jaccard cuando  $p_c = p_b = 0$ . Además, en este caso, se utiliza la distancia (1.9) elevada al cuadrado, es decir  $d_{ii'}^2 = 1 - m_{ii'}$ .

Una vez se utiliza alguna de las anteriores distancias o cualquier otra distancia (ver Mardia et al. (2002)) según los intereses del investigador, se realiza la descomposición espectral con la finalidad de realizar el modelo de regresión basado en distancias como se presentará en los siguientes capítulos. En este sentido, sea  $\mathbf{A}_{n \times n} = (a_{ii'})$  la matriz con elementos  $a_{ii'} = -d_{ii'}^2/2$ , y sea  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$  donde  $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$  es la matriz centrada, con  $\mathbf{I}_n$  una matriz identidad  $n \times n$  y  $\mathbf{1}_n$  un vector de unos  $n \times 1$ . Además,  $\mathbf{B}$  es una matriz semidefinida positiva (Mardia et al. 2002) de rango  $n - 1$ , entonces la matriz  $\mathbf{X}$  de coordenadas principales se puede obtener a partir de la siguiente descomposición espectral

$$\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}^t = \mathbf{X}\mathbf{X}^t \quad (1.13)$$

donde  $\mathbf{\Lambda}$  es una matriz diagonal conformada por los valores propios de  $\mathbf{B}$ ,  $\mathbf{X} = \mathbf{L}\mathbf{\Lambda}^{1/2}$  es una matriz  $n \times n$  de rango  $n - 1$  porque  $\mathbf{X}$  tiene un valor propio igual a  $\mathbf{1}_n$ , y  $\mathbf{L}$  contiene las coordenadas estandarizadas. La matriz  $\mathbf{B}$  proporciona las coordenadas euclídeas del conjunto  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ . Cada fila  $\mathbf{x}_i$  de  $\mathbf{X}$  contiene las coordenadas, llamadas coordenadas principales del individuo  $i$  ( $i = 1, \dots, n$ ).



## Capítulo 2

# Regresión beta basada en distancias con dispersión variable para la predicción de proporciones y tasas

### 2.1 Introducción

En una variedad de áreas científicas, se quiere evaluar la relación entre variables explicativas (continuas, categóricas y binarias) con una variable respuesta que representa una tasa, una proporción o partes por millón. Este es el caso de, por ejemplo, el porcentaje del gasto en alimentos en una familia, la prevalencia de Loa-loa en muestras de sangre, la proporción de petróleo crudo convertido a gasolina después de la destilación, la rentabilidad media anual en porcentaje en un fondo, la composición de una cartera de inversiones, el tiempo de uso diario en diferentes actividades y la distribución de las ventas en las diferentes regiones.

En estos casos, el modelo de regresión lineal clásico no es apropiado ya que la respuesta está restringida al intervalo  $(0,1)$ , y la estimación por OLS puede generar valores que exceden los límites inferior y superior. Por lo tanto, como se muestra en Rocke (1993), Cox (1996), Papke & Wooldridge (1996), Paolina (2001), Kieschnick & McCullough (2003), Ferrari & Cribari-Neto (2004) y Vasconcellos & Cribari-Neto (2005), una solución es transformar la variable dependiente de manera que asuma valores en la recta real, y así, la respuesta media transformada es modelada como un predictor lineal basado en un conjunto de variables explicativas.

La distribución beta es una familia muy flexible para modelar dicha clase

de datos debido a que su función de densidad tiene diferentes formas en función de los valores de los parámetros. Bury (1999) enumera las aplicaciones de la distribución beta en ingeniería. Johnson et al. (1995) presentaron y discutieron una serie de aplicaciones de la distribución beta. De acuerdo con los anteriores autores, esta distribución es una de las más frecuentemente empleadas en el modelamiento de distribuciones teóricas. Así mismo, Krysicki (1999) presenta algunas nuevas propiedades de esta distribución.

El modelo de regresión beta propuesto por Ferrari & Cribari-Neto (2004) es útil para situaciones en donde la variable respuesta  $y$  es continua y definida sobre el intervalo unidad estándar  $0 < y < 1$ . En su modelo, los parámetros de regresión, son interpretados en términos de la media de  $y$ ; el modelo es naturalmente heterocedástico y se adapta fácilmente a las asimetrías. Ospina et al. (2006) obtiene el sesgo de segundo orden de los estimadores de máxima verosimilitud y los utiliza para definir el sesgo de los estimadores ajustados, que son muy útiles para resolver el problema en muestras pequeñas. Sin embargo, el parámetro de precisión en estos casos tiene una varianza grande. El problema de modelar la varianza ha sido ampliamente discutida en la literatura estadística por autores como Cook & Weisberg (1983), Atkinson (1985), Botter & Cordeiro (1997) y Cysneiros et al. (2007).

Una variante del modelo de regresión beta que permite modelar la no linealidad y dispersión variable fue propuesta por Simas et al. (2010). En particular, en este modelo general, el parámetro que representa la precisión de los datos no se supone constante a través de observaciones, sino que puede variar, lo que conduce al modelo de regresión beta con dispersión variable. En varios casos, el problema de heterocedasticidad o dispersión variable puede ser resuelto por el modelo beta inflacionado de ceros, propuesto por Cook et al. (2008).

Con el fin de comprobar la bondad de ajuste del modelo beta estimado, es importante realizar el análisis de diagnóstico del modelo ajustado. En este caso, Espinheira et al. (2008*b*) propone dos nuevos residuales para esta clase de modelos y evalúa numéricamente su comportamiento en relación con el residual propuesto por Ferrari & Cribari-Neto (2004). Sin embargo, Espinheira et al. (2008*a*) propone un tipo de distancia de Cook para evaluar la influencia de las observaciones, así como las medidas de influencia local bajo diferentes esquemas de perturbación. Ferrari et al. (2011) deriva las matrices apropiadas para la evaluación de la influencia local y residual sobre los parámetros estimados bajo diferentes esquemas de perturbación.

Por otra parte, muchos métodos estadísticos y análisis de datos utilizan el concepto geométrico de distancia entre individuos o poblaciones; estos métodos se aplican en campos como la economía, la antropología, la biología, la genética, la psicología, entre otros (Arenas & Cuadras 2002). El concepto de distancia es una herramienta útil en muchos aspectos de la estadística: contras-

te de hipótesis, estimación de parámetros, regresión, análisis discriminante, etc. Cuadras & Arenas (1990) proponen el método de regresión múltiple basada en el análisis de distancias utilizando diferentes métricas para trabajar con variables explicativas continuas y categóricas. Cuadras et al. (1996) presentaron algunos resultados adicionales del modelo DB para la predicción con variables mixtas y exploró el problema de datos faltantes dando una solución con DB. Algunos de los trabajos más recientes son presentados por Esteve et al. (2009) y Boj et al. (2010), quienes proponen métodos adiciones que incluyen términos polinomiales y de interacción en la regresión DB, y además, hacen regresiones lineales locales utilizando predictores funcionales para crear los pesos en la regresión DB.

Aunque muchos autores han considerado variables explicativas mixtas en la regresión beta, no existe un enfoque totalmente general para resolver el problema bajo datos mixtos, e incluso la presencia de multicolinealidad que puede ser causado por esta mezcla de variables puede traer graves problemas en la estimación de parámetros. Un método tradicional es el de asignar puntajes a variables cualitativas con el fin de hacerlos continuo; sin embargo, este procedimiento puede aumentar la multicolinealidad. Por lo tanto, en este trabajo se propone un método alternativo para la solución de tales problemas usando distancias euclidianas entre los individuos. Además, el problema de heterocedasticidad es solucionado modelando la dispersión variable empleando la metodología propuesta por Simas et al. (2010) y Ferrari et al. (2011). En este contexto, se propone una metodología DB para ajustar una variable respuesta tipo beta con dispersión variable. El modelo propuesto se basa en las metodologías presentadas por Cuadras & Arenas (1990), Cuadras et al. (1996), Ferrari & Cribari-Neto (2004) y Simas et al. (2010).

En las siguientes secciones se emplea el método de máxima verosimilitud para estimar los parámetros desconocidos del modelo propuesto y se presentan las principales propiedades de estos estimadores. Además, se realiza la inferencia estadística sobre los parámetros utilizando las aproximaciones obtenidas a partir de la normalidad asintótica del estimador de máxima verosimilitud; se desarrolla el diagnóstico y predicción de una nueva observación, y se estudia el problema de datos faltantes utilizando la metodología propuesta.

Los procedimientos de modelado y de inferencia propuestos son similares a los modelos lineales generalizados discutidos por McCullagh & Nelder (1989) y Myers et al. (2002), excepto que la distribución de la respuesta no es un miembro de la familia exponencial. Aunque, la respuesta no es un miembro de la familia exponencial, una adaptación a esta familia se lleva a cabo siguiendo las propuestas en modelos generalizados desarrollado por McCulloch & Searle (2001), Lee & Nelder (2002), Dobson (2002), y Smith & Ridout (2003). Una ventaja del modelo propuesto es que es bastante general ya que sólo se debe elegir en función del tipo de variables explicativas una distancia adecuada para



el modelo de la media, y otra o la misma, para el modelo de dispersión variable.

A partir del enfoque planteado se hacen dos aplicaciones en donde se ajustan los modelos de regresión beta basado en distancias con dispersión (precisión) variable. La primera aplicación es la proporción de petróleo crudo convertido a gasolina después de la destilación; en ésta, se utiliza la distancia Gower para los parámetros del modelo de media, y la distancia clásica, para los parámetros del modelo de dispersión variable. Mientras la segunda aplicación es el porcentaje promedio de los retornos durante 5 años; en este estudio de fondos de inversión se utiliza la distancia Gower, tanto en el modelo de media como el modelo con dispersión variable. En este último estudio, se supone que hay un 5 %, 10 % y 20 % de valores perdidos en las variables explicativas; con estos porcentajes de pérdida se muestra que el método de regresión beta basado en distancias (distance-based beta regression, DBBR) claramente funciona bien a pesar de la pérdida de información.

Este capítulo está dividido en las siguientes secciones: en la Sección 2.2 se presenta el modelo propuesto, se desarrolla la estimación de los parámetros, las pruebas de hipótesis. En la Sección 2.3 se presentan algunas medidas de bondad de ajuste, se hace la selección de las coordenadas principales, el diagnóstico y predicción del modelo DBBR, y se estudia el problema de datos faltantes. En la Sección 2.4 se demuestra que las estimaciones obtenidas en el modelo DBBR son las mismas que las obtenidas en el modelo de regresión beta tradicional propuesto por Ferrari & Cribari-Neto (2004), inclusive estas predicciones pueden mejorar si se incluyen más coordenadas principales en el DBBR. Por último, en la Sección 2.5 se presentan dos aplicaciones de la metodología propuesta.

## 2.2 Modelo de regresión beta basado en distancia con dispersión variable

El modelo se basa en el supuesto que la respuesta tiene distribución beta, la función de densidad beta está dada por

$$\pi(y, p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{(p-1)}(1-y)^{(q-1)}, \quad 0 < y < 1$$

donde  $p > 0$ ,  $q > 0$  y  $\Gamma(\cdot)$  es la función Gamma. La media y la varianza de  $y$  son, respectivamente,

$$E(y) = \frac{p}{p+q} \tag{2.1}$$

y

$$Var(y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{2.2}$$

Las estimaciones de  $\mathbf{p}$  y  $\mathbf{q}$  por máxima verosimilitud y la aplicación de los ajustes en los sesgos en pequeñas muestras de los estimadores de máxima verosimilitud para los parámetros son analizados en Cribari-Neto & Vasconcellos (2002), Ospina et al. (2006) y Espinheira et al. (2008a).

Con el fin de obtener una estructura de regresión para la media de la respuesta junto con el parámetro de precisión, se trabaja con una parametrización diferente de la densidad beta. De acuerdo a Ferrari & Cribari-Neto (2004), sea  $\mu = \mathbf{p}/(\mathbf{p} + \mathbf{q})$  y  $\phi = \mathbf{p} + \mathbf{q}$ , es decir,  $\mathbf{p} = \mu\phi$  y  $\mathbf{q} = (1 - \mu)\phi$ . De lo anterior para las ecuaciones (2.1) y (2.2) se sigue que  $E(y) = \mu$  y  $\text{Var}(y) = \mu(1 - \mu)/(1 + \phi)$ .  $\mu$  es la media de la variable respuesta y  $\phi$  puede ser interpretado como un parámetro de precisión en el sentido de que para un  $\mu$  fijo, un valor grande de  $\phi$  conlleva a un menor valor de la varianza de  $y$ . En Cepeda-Cuervo & Achcar (2010) se puede encontrar otra reparametrización de la distribución beta que presenta resultados similares a la presentada en este trabajo.

En términos de la nueva parametrización, la densidad de  $y$  se puede reescribir como (Ferrari & Cribari-Neto 2004)

$$f(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1 - \mu)\phi)} y^{(\mu\phi-1)}(1 - y)^{((1-\mu)\phi-1)}, \quad 0 < y < 1 \quad (2.3)$$

donde  $0 < \mu < 1$ ,  $\phi > 0$  y  $\Gamma(\cdot)$  es la función gamma. La función de densidad en (2.3) presenta diferentes formas de dependencia en los valores de estos dos parámetros. En particular, ésta puede ser simétrica cuando  $\mu = 1/2$  o asimétrica cuando  $\mu \neq 1/2$ . Adicionalmente, la dispersión de la distribución para un  $\mu$  fijo decrece a medida que  $\phi$  crece. En particular cuando  $\mu = 1/2$  y  $\phi = 2$  la densidad se reduce a la distribución uniforme.

Aunque en este trabajo la respuesta está restringida al intervalo unitario  $(0, 1)$ , el modelo propuesto es útil para situaciones donde la respuesta está restringida al intervalo  $(a, b)$  donde  $a$  y  $b$  son escalares conocidos, con  $a < b$ . En este caso se debe modelar  $(y - a)/(b - a)$  en cambio de modelar  $y$  directamente. Además, si  $y$  también asume los extremos 0 y 1, una transformación útil en la práctica es  $(y(n - 1) + 0.5)/n$ , donde  $n$  es el tamaño de la muestra Smithson & Verkuilen (2006).

Para formar un modelo de regresión utilizando la aproximación del GLM extendido, se utilizan dos funciones de enlace: una para el parámetro de localización ( $\mu$ ) y la otra para el parámetro de precisión ( $\phi$ ). La función de enlace es una función no lineal, suave y monótona que asigna el espacio ilimitado del predictor lineal al espacio muestral apropiado de las observaciones; por lo tanto, enlaza el predictor lineal con las observaciones.

Sea  $y_1, y_2, \dots, y_n$  variables aleatorias independientes, cada  $y_i$  sigue la función de densidad en (2.3) con media  $\mu_i$  y dispersión desconocida  $\phi_i$ ,  $i = 1, \dots, n$ . Asuma que la media y la dispersión variable de  $y_i$  se pueden escribir como

(Smithson & Verkuilen 2006, Simas et al. 2010)

$$\eta_{1i} = g_1(\mu_i) = \zeta_0 + \mathbf{v}_i^t \boldsymbol{\zeta} \quad \eta_{2i} = g_2(\phi_i) = \varphi_0 + \mathbf{u}_i^t \boldsymbol{\varphi} \quad (2.4)$$

donde  $\zeta_0$  y  $\varphi_0$  son parámetros asociados a los interceptos desconocidos,  $\boldsymbol{\zeta}^t = (\zeta_1, \dots, \zeta_{p_1})$ ,  $\boldsymbol{\varphi}^t = (\varphi_1, \dots, \varphi_{p_2})$  son vectores de parámetros de regresión desconocidos e independientes, con  $\boldsymbol{\zeta} \in \mathbb{R}^{p_1}$  y  $\boldsymbol{\varphi} \in \mathbb{R}^{p_2}$ ,  $p_1 + p_2 < n$ . Adicionalmente,  $\mathbf{v}_i^t = (v_{i1}, \dots, v_{ip_1})$  y  $\mathbf{u}_i^t = (u_{i1}, \dots, u_{ip_2})$  son vectores de las  $p_1$  y  $p_2$  variables explicativas observadas centradas, respectivamente. Note que el parámetro de precisión no es constante para todas las observaciones, sino que en cambio es modelado de forma similar al parámetro de la media.

Las funciones de enlace  $g_1 : (0, 1) \rightarrow \mathbb{R}$  y  $g_2 : (0, \infty) \rightarrow \mathbb{R}$  son estrictamente monótonas y dos veces diferenciables. Algunas posibles elecciones para la función de enlace en el modelo de localización,  $g_1(\mu_i)$ , son: logit,  $g_1(\mu_i) = \log\{\mu_i/(1 - \mu_i)\}$ ; probit,  $g_1(\mu_i) = \Phi^{-1}(\mu_i)$  donde  $\Phi(\cdot)$  es la función de distribución acumulativa de una variable normal estándar; complemento log-log (cloglog),  $g_1(\mu_i) = \log\{-\log(1 - \mu_i)\}$ ; y loglog,  $g_1(\mu_i) = -\log\{-\log(\mu_i)\}$ . Una rica discusión de las funciones de enlace para el modelo de localización es presentada en Atkinson (1985) y McCullagh & Nelder (1989). En el modelo precisión algunas funciones de enlace para  $g_2$  son: logaritmo,  $g_2(\phi_i) = \log \phi_i$ ; raíz cuadrada,  $g_2(\phi_i) = \sqrt{\phi_i}$ ; y la identidad,  $g_2(\phi_i) = \phi_i$ , entre otras.

En forma matricial, los modelos en (2.4) se pueden reescribir como:

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) = \zeta_0 \mathbf{1} + \mathbf{V} \boldsymbol{\zeta} \quad \boldsymbol{\eta}_2 = g_2(\boldsymbol{\phi}) = \varphi_0 \mathbf{1} + \mathbf{U} \boldsymbol{\varphi} \quad (2.5)$$

donde  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^t$ ,  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)^t$ ,  $\mathbf{1}$  es un vector de unos de tamaño  $n \times 1$ ,  $\mathbf{V} = \mathbf{H}\mathbf{V}^*$  y  $\mathbf{U} = \mathbf{H}\mathbf{U}^*$  con  $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$  una matriz centrada e  $\mathbf{I}_n$  la matriz identidad de tamaño  $n \times n$ .  $\mathbf{V}^*$  y  $\mathbf{U}^*$  son matrices conformadas por las variables explicativas originales asociadas con el modelo de media y dispersión variable, respectivamente. Además, tanto  $\mathbf{V}^*$  como  $\mathbf{U}^*$  pueden incluir variables continuas, categóricas y binarias, o incluso una mezcla de ellas.

### 2.2.1 Construcción del modelo beta utilizando distancias

Considere un conjunto de  $n$  individuos donde se observan las variables aleatorias independientes  $y_1, y_2, \dots, y_n$ , y las variables explicativas,  $\mathbf{v}_i^t = (v_{i1}, \dots, v_{ip_1})$  y  $\mathbf{u}_i^t = (u_{i1}, \dots, u_{ip_2})$ . A través de las funciones de distancia  $d(\cdot, \cdot)$  y  $\delta(\cdot, \cdot)$  que dependen de  $\mathbf{v}_i$ 's y  $\mathbf{u}_i$ 's, respectivamente (como se muestra en la Sección 2.3), se encuentran las matrices de distancia  $\mathbf{D}_v = (d_{ii'})$  y  $\mathbf{D}_u = (\delta_{ii'})$  donde  $d_{ii'}$  y  $\delta_{ii'}$  son distancias entre los individuos  $i$  e  $i'$  para los modelos de media y de dispersión variable, respectivamente. Obsérvese que

si las mismas variables explicativas se toman en los modelos de media y de dispersión variable, entonces  $\mathbf{D}_v = (d_{ii'})$  y  $\mathbf{D}_u = (\delta_{ii'})$  coinciden.

Estas distancias satisfacen que la distancia es cercana a 0 si  $\mathbf{v}$  o  $\mathbf{u}$  medidas en  $i$  e  $i'$  son muy similares, es decir,  $d_{ii'} \cong 0$  o  $\delta_{ij} \cong 0$  si  $\mathbf{v}_i \cong \mathbf{v}_{i'}$  o  $\mathbf{u}_i \cong \mathbf{u}_{i'}$ , respectivamente. Además, si  $d_{ii'} \cong 0$  se observa que los individuos  $i$  e  $i'$  están cerca en el modelo de media, y/o si  $\delta_{ii'} \cong 0$  se dice que los individuos  $i$  e  $i'$  están cerca en el modelo de dispersión variable.

Siguiendo el procedimiento presentado en la Sección 1.6 y la propuesta realizada por Cuadras (1989): como  $\mathbf{D}_v = (d_{ii'})$  y  $\mathbf{D}_u = (\delta_{ii'})$  son matrices de distancias euclidianas, sean  $\mathbf{A}_v = (a_{ii'}^v)$  y  $\mathbf{A}_u = (a_{ii'}^u)$  donde  $a_{ii'}^v = -d_{ii'}^2/2$  y  $a_{ii'}^u = -\delta_{ii'}^2/2$ , y además, sean  $\mathbf{B}_v = \mathbf{H}\mathbf{A}_v\mathbf{H}$  y  $\mathbf{B}_u = \mathbf{H}\mathbf{A}_u\mathbf{H}$ . Entonces, de acuerdo a Mardia et al. (2002),  $\mathbf{B}_v$  y  $\mathbf{B}_u$  son matrices semi-definidas positivas de rango  $p_1$  y  $p_2$ , respectivamente. Por lo tanto, estas matrices se pueden expresar como

$$\begin{aligned} \mathbf{B}_v &= \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right) \mathbf{A}_v \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right) & \mathbf{B}_u &= \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right) \mathbf{A}_u \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right) \\ &= \mathbf{L}_x \mathbf{\Lambda}_x \mathbf{L}_x^t = \mathbf{X}\mathbf{X}^t & &= \mathbf{L}_z \mathbf{\Lambda}_z \mathbf{L}_z^t = \mathbf{Z}\mathbf{Z}^t \end{aligned}$$

donde  $\mathbf{\Lambda}_x$  y  $\mathbf{\Lambda}_z$  son matrices diagonales conformadas por los valores propios de  $\mathbf{B}_v$  y  $\mathbf{B}_u$ , respectivamente.  $\mathbf{X} = \mathbf{L}_x \mathbf{\Lambda}_x^{1/2}$  es una matriz  $n \times (n-1)$  de rango  $(n-1)$ ,  $\mathbf{Z} = \mathbf{L}_z \mathbf{\Lambda}_z^{1/2}$  es una matriz  $n \times (n-1)$  de rango  $(n-1)$ , y  $\mathbf{L}_x$  y  $\mathbf{L}_z$  contienen las coordenadas estandarizadas para el modelo de medias y el modelo de dispersión variable.

Además, las filas  $\mathbf{x}_1^t, \dots, \mathbf{x}_n^t$  de  $\mathbf{X}$  son las coordenadas principales de  $\mathbf{B}_v$  con respecto a la matriz de distancia  $\mathbf{D}_v$ . Cuando un individuo  $i$  es similar a un individuo  $i'$  en (2.4) entonces  $\mathbf{v}_i \cong \mathbf{v}_{i'}$  y por consiguiente, es evidente que  $\mathbf{x}_i \cong \mathbf{x}_j$ . Lo mismo sucede con las filas  $\mathbf{z}_1^t, \dots, \mathbf{z}_n^t$  de  $\mathbf{Z}$ . Por lo tanto, los modelos en (2.4) pueden ser reexpresados como los siguientes modelos de media y de dispersión variable basados en distancias, respectivamente,

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta}_* \quad \boldsymbol{\eta}_2 = g_2(\boldsymbol{\phi}) = \alpha_0 \mathbf{1} + \mathbf{Z}\boldsymbol{\alpha}_* \quad (2.6)$$

donde  $\beta_0$  y  $\alpha_0$  son los parámetros relacionados a los interceptos desconocidos,  $\boldsymbol{\beta}_*^t = (\beta_1, \dots, \beta_{n-1})$ ,  $\boldsymbol{\alpha}_*^t = (\alpha_1, \dots, \alpha_{n-1})$  son los vectores de parámetros de regresión desconocidos y formalmente independientes, con  $\boldsymbol{\beta}_* \in \mathbb{R}^{n-1}$  y  $\boldsymbol{\alpha}_* \in \mathbb{R}^{n-1}$ .

Particionando  $\mathbf{X}$  como  $\mathbf{X} = (\mathbf{X}_{(k_1)} \quad \mathbf{X}_{n-k_1})$  donde  $\mathbf{X}_{(k_1)}$  contiene un subconjunto de  $k_1$  columnas de  $\mathbf{X}$ , y  $\mathbf{X}_{n-k_1}$  contiene las restantes columnas de  $\mathbf{X}$ . Realizando el mismo procedimiento con  $\mathbf{Z}$ ,  $\mathbf{Z} = (\mathbf{Z}_{(k_2)} \quad \mathbf{Z}_{n-k_2})$  donde  $\mathbf{Z}_{(k_2)}$  contiene un subconjunto de  $k_2$  columnas de  $\mathbf{Z}$ , y  $\mathbf{Z}_{n-k_2}$  contiene las restantes columnas de  $\mathbf{Z}$ . Entonces, los modelos reducidos DB se pueden expresar como

$$\boldsymbol{\eta}_1 = \beta_0 \mathbf{1} + \mathbf{X}_{(k_1)} \boldsymbol{\beta}_{(k_1)} \quad \boldsymbol{\eta}_2 = \alpha_0 \mathbf{1} + \mathbf{Z}_{(k_2)} \boldsymbol{\alpha}_{(k_2)} \quad (2.7)$$

donde  $\mathbf{X}_{(k_1)} = (\mathbf{X}_1, \dots, \mathbf{X}_{k_1})$  con  $\mathbf{X}_j$  ( $j = 1, \dots, k_1$ ) la  $j$ -ésima columna de  $\mathbf{X}$ ,  $\mathbf{Z}_{(k_2)} = (\mathbf{Z}_1, \dots, \mathbf{Z}_{k_2})$  con  $\mathbf{Z}_h$  ( $h = 1, \dots, k_2$ ) la  $h$ -ésima columna de  $\mathbf{Z}$ ; siendo además cada  $\mathbf{X}_j$  y  $\mathbf{Z}_h$  coordenadas principales de  $\mathbf{X}$  y  $\mathbf{Z}$ , respectivamente.

El modelo propuesto en (2.7) se puede reescribir como

$$\boldsymbol{\eta}_1 = \beta_0 \mathbf{1} + \sum_{j=1}^{k_1} \beta_j \mathbf{X}_j \quad \boldsymbol{\eta}_2 = \alpha_0 \mathbf{1} + \sum_{h=1}^{k_2} \alpha_h \mathbf{Z}_h \quad (2.8)$$

o alternativamente como

$$\eta_{1i} = g_1(\mu_i) = \sum_{j=0}^{k_1} x_{ij} \beta_j = \mathbf{x}_i^t \boldsymbol{\beta} \quad \eta_{2i} = g_2(\phi_i) = \sum_{h=0}^{k_2} z_{ih} \alpha_h = \mathbf{z}_i^t \boldsymbol{\alpha} \quad (2.9)$$

$i = 1, \dots, n$ , donde  $x_{i0} = 1$ ,  $z_{i0} = 1$ ,  $\mathbf{x}_i^t = (x_{i0}, \dots, x_{ik_1})$ ,  $\mathbf{z}_i^t = (z_{i0}, \dots, z_{ik_2})$ ,  $\boldsymbol{\beta}^t = (\beta_0, \dots, \beta_{k_1})$  y  $\boldsymbol{\alpha}^t = (\alpha_0, \dots, \alpha_{k_2})$ .

Obsérvese que para los modelos presentados en (2.7) se cumple:  $\mathbf{X}_j^t \mathbf{1} = 0$  y  $\mathbf{Z}_h^t \mathbf{1} = 0$ ,  $\mathbf{X}_j^t \mathbf{X}_j = \lambda_{x_j}$  con  $\lambda_{x_j}$  un valor propio de  $\mathbf{B}_v$ ,  $\mathbf{Z}_h^t \mathbf{Z}_h = \lambda_{z_h}$  con  $\lambda_{z_h}$  un valor propio de  $\mathbf{B}_u$ ,  $\mathbf{X}_j^t \mathbf{X}_{j'} = 0$  para  $j \neq j'$  y  $\mathbf{Z}_h^t \mathbf{Z}_{h'} = 0$  para  $h \neq h'$ , con  $j, j' = 1, \dots, k_1$  y  $h, h' = 1, \dots, k_2$ . Por lo tanto, esos modelos están en una forma ortogonal centrada.

## 2.2.2 Estimación de parámetros

En esta subsección se utilizarán algunos de los resultados de la regresión beta presentados en Ferrari & Cribari-Neto (2004), junto con la extensión del modelo de regresión beta con dispersión variable realizada por Smithson & Verkuilen (2006) y formalmente introducida (junto con otras extensiones) por Simas et al. (2010).

Sea la densidad para cada  $y_i$  como la presentada en la ecuación (2.3), el logaritmo de la función de verosimilitud es:

$$l_i(\mu_i, \phi_i) = \log \Gamma(\phi_i) - \log \Gamma(\mu_i \phi_i) - \log \Gamma((1 - \mu_i) \phi_i) \\ + (\mu_i \phi_i - 1) \log(y_i) + ((1 - \mu_i) \phi_i - 1) \log(1 - y_i) \quad (2.10)$$

con  $\mu_i = g_1^{-1}(\eta_{1i})$  y  $\phi_i = g_2^{-1}(\eta_{2i})$  definidas en (2.9).

Entonces la función de log-verosimilitud para esta clase de modelo de regresión beta está dada por

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n l_i(\mu_i, \phi_i) \quad (2.11)$$

Derivando parcialmente la ecuación (2.11) con respecto a cada  $\beta_j$ , para  $j = 0, 1, \dots, k_1$ , se obtiene la siguiente expresión

$$U_j(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i(\mu_i, \phi_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_{1i}} \frac{\partial \eta_{1i}}{\partial \beta_j} \quad (2.12)$$

donde  $\partial \eta_{1i}/\partial \beta_j = x_{ij}$ , y de (2.9),  $\partial \mu_i/\partial \eta_{1i} = 1/g'_1(\mu_i) = dg_1^{-1}(\eta_{1i})/d\eta_{1i}$ . Además, derivando parcialmente (2.10) con respecto a cada  $\mu_i$ , para  $i = 1, \dots, n$ , se tiene que

$$\frac{\partial l_i(\mu_i, \phi_i)}{\partial \mu_i} = \phi_i \left\{ \log \left( \frac{y_i}{1 - y_i} \right) - [\psi(\mu_i \phi_i) - \psi((1 - \mu_i) \phi_i)] \right\} \quad (2.13)$$

donde  $\psi(\cdot)$  es una función digamma, es decir,  $\psi(z) = \partial \log \Gamma(z)/\partial z$  para  $z > 0$ .

Sea  $y_i^* = \log(y_i/(1 - y_i))$  y  $\mu_i^* = \psi(\mu_i \phi_i) - \psi((1 - \mu_i) \phi_i)$ , entonces

$$U_j(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_j} = \sum_{i=1}^n \phi_i (y_i^* - \mu_i^*) \frac{1}{g'_1(\mu_i)} x_{ij} \quad (2.14)$$

Ahora, derivando parcialmente (2.11) con respecto a cada  $\alpha_h$ , se obtiene

$$U_h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_h} = \sum_{i=1}^n \frac{\partial l_i(\mu_i, \phi_i)}{\partial \phi_i} \frac{\partial \phi_i}{\partial \eta_{2i}} \frac{\partial \eta_{2i}}{\partial \alpha_h}, \quad h = 0, 1, \dots, k_2 \quad (2.15)$$

donde  $\partial \eta_{2i}/\partial \alpha_h = z_{ih}$ , y de (2.9),  $\partial \phi_i/\partial \eta_{2i} = 1/g'_2(\phi_i) = dg_2^{-1}(\eta_{2i})/d\eta_{2i}$ . Además, derivando parcialmente (2.10) con respecto a cada  $\phi_i$ , para  $i = 1, \dots, n$ , se obtiene

$$\frac{\partial l_i(\mu_i, \phi_i)}{\partial \phi_i} = \psi(\phi_i) + \mu_i (y_i^* - \mu_i^*) - \psi((1 - \mu_i) \phi_i) + \log(1 - y_i) \quad (2.16)$$

Por lo tanto,

$$U_h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n [\psi(\phi_i) + \mu_i (y_i^* - \mu_i^*) - \psi((1 - \mu_i) \phi_i) + \log(1 - y_i)] \frac{1}{g'_2(\phi_i)} z_{ih},$$

para  $h = 0, 1, \dots, k_2$ .

Bajo condiciones de regularidad, se sabe que

$$\begin{aligned} \mathbb{E} \left( \log \left( \frac{y_i}{1 - y_i} \right) \right) &= \psi(\mu_i \phi_i) - \psi((1 - \mu_i) \phi_i) \\ \mathbb{E}[\log(1 - y_i)] &= \psi((1 - \mu_i) \phi_i) - \psi(\phi_i) \end{aligned}$$

Considerando el vector de parámetros completo  $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{\alpha}^t)^t$ , la expresión matricial de la función score obtenida por diferenciación de la función de log-verosimilitud con respecto a los parámetros desconocidos está dada por  $\mathbf{U}(\boldsymbol{\theta}) = (\mathbf{U}_\beta^t(\boldsymbol{\theta}), \mathbf{U}_\alpha^t(\boldsymbol{\theta}))^t$ , donde

$$\mathbf{U}_\beta(\boldsymbol{\theta}) = \mathbf{X}^t \boldsymbol{\Upsilon} \mathbf{T}_1 (\mathbf{y}^* - \boldsymbol{\mu}^*) \quad \text{y} \quad \mathbf{U}_\alpha(\boldsymbol{\theta}) = \mathbf{Z}^t \mathbf{T}_2 \boldsymbol{\vartheta} \quad (2.17)$$

y donde  $\mathbf{y}^* = (y_1^*, \dots, y_n^*)^t$ ,  $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_n^*)^t$ ,  $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_n)^t$ ,  $\mathbf{T}_1 = \text{diag}(d\mu_i/d\eta_{1i})$ ,  $\mathbf{T}_2 = \text{diag}(d\phi_i/d\eta_{2i})$  y  $\boldsymbol{\Upsilon} = \text{diag}(\phi_i)$ , con  $\text{diag}(\mu_i)$  que denota la matriz diagonal  $n \times n$  ( $i = 1, \dots, n$ ) y  $\vartheta_i = \mu_i(y_i^* - \mu_i^*) + \psi(\phi_i) - \psi((1 - \mu_i)\phi_i) + \log(1 - y_i)$ .

El siguiente paso es encontrar la segunda derivada de  $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$  con respecto a  $\boldsymbol{\beta}$ , es decir, una expresión para la matriz de información de Fisher. Derivando parcialmente (2.12) con respecto a  $\beta_j$ , se tiene

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_j \partial \beta_{j'}} = \sum_{i=1}^n \left( \frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \mu_i^2} \frac{d\mu_i}{d\eta_{1i}} + \frac{\partial l_i(\mu_i, \phi_i)}{\partial \mu_i} \frac{\partial}{\partial \mu_i} \frac{d\mu_i}{d\eta_{1i}} \right) \frac{d\mu_i}{d\eta_{1i}} x_{ij} x_{ij'}$$

Como  $E(\partial l_i(\mu_i, \phi_i)/\partial \mu_i) = 0$ , entonces

$$E \left( \frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_j \partial \beta_{j'}} \right) = \sum_{i=1}^n E \left( \frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \mu_i^2} \right) \left( \frac{d\mu_i}{d\eta_{1i}} \right)^2 x_{ij} x_{ij'}$$

Derivando (2.13) ahora parcialmente con respecto a  $\mu_i$ , se obtiene

$$\frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \mu_i^2} = -\phi_i^2 [\psi'(\mu_i \phi_i) + \psi'((1 - \mu_i)\phi_i)]$$

y entonces

$$E \left( \frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_j \partial \beta_{j'}} \right) = - \sum_{i=1}^n \phi_i^2 a_i x_{ij} x_{ij'}$$

con

$$a_i = \{\psi'(\mu_i \phi_i) + \psi'((1 - \mu_i)\phi_i)\} \frac{1}{g_1'(\mu_i)^2}$$

donde  $\psi'(\cdot)$  es la función trigamma.

De (2.12), la segunda derivada de  $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$  con respecto a  $\alpha_h$  se puede expresar como

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_h \partial \beta_j} = \sum_{i=1}^n \left( \frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \phi_i \partial \mu_i} \frac{d\mu_i}{d\eta_{1i}} + \frac{\partial l_i(\mu_i, \phi_i)}{\partial \mu_i} \frac{\partial}{\partial \phi_i} \frac{d\mu_i}{d\eta_{1i}} \right) \frac{d\phi_i}{d\eta_{2i}} z_{ih} x_{ij'}$$

Tomando los valores esperados a ambos lados de la expresión anterior, se tiene que

$$E \left( \frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_h \partial \beta_j} \right) = \sum_{i=1}^n E \left( \frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \phi_i \partial \mu_i} \right) \left( \frac{d\mu_i}{d\eta_{1i}} \right) \left( \frac{d\phi_i}{d\eta_{2i}} \right) z_{ih} x_{ij'}$$

Por otra parte, derivando parcialmente (2.13) con respecto a  $\phi_i$ , se obtiene

$$\frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \phi_i \partial \mu_i} = y_i^* - \mu_i^* - \phi_i [\mu_i a_i - \psi'((1 - \mu_i)\phi_i)]$$

y entonces

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_h \partial \beta_j}\right) = - \sum_{i=1}^n \phi_i [\mu_i a_i - \psi'((1 - \mu_i)\phi_i)] \left(\frac{d\mu_i}{d\eta_{1i}}\right) \left(\frac{d\phi_i}{d\eta_{2i}}\right) z_{ih} x_{ij'}$$

Por último, derivando parcialmente (2.15) con respecto a  $\phi_i$ , se llega a

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_h \partial \alpha_{h'}} = \sum_{i=1}^n \left( \frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \phi_i^2} \frac{d\phi_i}{d\eta_{2i}} + \frac{\partial l_i(\mu_i, \phi_i)}{\partial \phi_i} \frac{\partial}{\partial \phi_i} \frac{d\phi_i}{d\eta_{2i}} \right) \frac{d\phi_i}{d\eta_{2i}} z_{ih} z_{ih'}$$

Como  $E(\partial l_i(\mu_i, \phi_i)/\partial \phi_i) = 0$  y tomando los valores esperados a ambos lados de la expresión anterior, se encuentra que

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_h \partial \alpha_{h'}}\right) = \sum_{i=1}^n E\left(\frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \phi_i^2}\right) \left(\frac{d\phi_i}{d\eta_{2i}}\right)^2 z_{ih} z_{ih'}$$

Derivando parcialmente (2.16) con respecto a  $\phi_i$ , se tiene que

$$\frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial \phi_i^2} = \psi'(\phi_i) - \mu_i^2 \psi'(\mu_i \phi_i) - (1 - \mu_i)^2 \psi'((1 - \mu_i)\phi_i)$$

y entonces

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_h \partial \alpha_{h'}}\right) = - \sum_{i=1}^n b_i \left(\frac{d\phi_i}{d\eta_{2i}}\right)^2 z_{ih} z_{ih'}$$

donde  $b_i = (1 - \mu_i)^2 \psi'((1 - \mu_i)\phi_i) + \mu_i^2 \psi'(\mu_i \phi_i) - \psi'(\phi_i)$ .

Los estimadores de máxima verosimilitud (MLE) de  $\boldsymbol{\beta}$  y  $\boldsymbol{\alpha}$  se obtienen del sistema no lineal  $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$ , con  $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{\alpha}^t)^t$ . En la práctica, los MLE's se obtienen por maximización numérica de (2.11), utilizando un algoritmo de optimización no lineal, tal como Newton-Raphson, Fisher's scoring o quasi-Newton; para más detalles ver Press et al. (1992) y Nocedal & Wright (1999). Como valores iniciales para  $\boldsymbol{\beta}$  y  $\boldsymbol{\alpha}$  se sugiere utilizar los siguientes modelos de regresión lineal normal

$$g_1(\mu_i) = \sum_{j=0}^{k_1} x_{ij} \beta_j \quad \text{y} \quad g_2(\sigma_i^{-2}) = \sum_{h=0}^{k_2} z_{ih} \alpha_h$$

Este procedimiento producirá  $\hat{\boldsymbol{\beta}}^{(0)}$  y  $\hat{\boldsymbol{\alpha}}^{(0)}$ , los cuales se utilizan como valores iniciales. Note que se está asumiendo que  $Y_i \sim N(\mu_i, \sigma_i^2)$ .



### 2.2.3 Casos especiales

En esta subsección se presentan algunos modelos especiales que se encuentran comúnmente en la práctica. Sea  $\mathbf{P}$  y  $\mathbf{W}$  matrices de tamaños  $2n \times (k_1 + k_2)$  y  $2n \times 2n$ , respectivamente, las cuales están dadas por

$$\mathbf{P} = \begin{pmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \quad \text{y} \quad \mathbf{W} = \begin{pmatrix} \mathbf{W}_{\beta\beta} & \mathbf{W}_{\beta\alpha} \\ \mathbf{W}_{\alpha\beta} & \mathbf{W}_{\alpha\alpha} \end{pmatrix} \quad (2.18)$$

donde  $\mathbf{W}_{\beta\alpha} = \text{diag} \left( \phi_i [\mu_i a_i - \psi'((1 - \mu_i)\phi_i)] \left( \frac{d\mu_i}{d\eta_{1i}} \right) \left( \frac{d\phi_i}{d\eta_{2i}} \right) \right)$ ,  $\mathbf{W}_{\alpha\beta} = \mathbf{W}_{\beta\alpha}^t$ ,  $\mathbf{W}_{\beta\beta} = \text{diag} \left( \phi_i^2 a_i \left( \frac{d\mu_i}{d\eta_{1i}} \right)^2 \right)$  y  $\mathbf{W}_{\alpha\alpha} = \text{diag} \left( b_i \left( \frac{d\phi_i}{d\eta_{2i}} \right) \right)$ .

Por lo tanto, la matriz de información de Fisher para el vector de parámetros,  $\boldsymbol{\theta}$ , está dado por

$$\mathbf{K}(\boldsymbol{\theta}) = \mathbf{K}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{P}^t \mathbf{W} \mathbf{P} = \begin{pmatrix} \mathbf{K}_{\beta\beta} & \mathbf{K}_{\beta\alpha} \\ \mathbf{K}_{\alpha\beta} & \mathbf{K}_{\alpha\alpha} \end{pmatrix} \quad (2.19)$$

En este caso, como  $\mathbf{W}_{\beta\alpha} \neq \mathbf{0}$  los parámetros  $\boldsymbol{\beta}$  y  $\boldsymbol{\alpha}$  no son ortogonales, en contraste con lo que sucede con los modelos lineales generalizados donde esta ortogonalidad se satisface (McCullagh & Nelder 1989). Sin embargo, los MLE's de  $\hat{\boldsymbol{\theta}}$  y  $\mathbf{K}(\hat{\boldsymbol{\theta}})$  son estimadores consistentes de  $\boldsymbol{\theta}$  y  $\mathbf{K}(\boldsymbol{\theta})$ , respectivamente, donde  $\mathbf{K}(\hat{\boldsymbol{\theta}})$  es la matriz de información de Fisher evaluada en  $\hat{\boldsymbol{\theta}}$ .

Observe que si las variables originales,  $\mathbf{V}$  y  $\mathbf{U}$ , se juntan en una sola matriz de variables explicativas y luego, el proceso de distancias y descomposición espectral se desarrollan, la matriz  $\mathbf{W}_{\beta\alpha} = \mathbf{0}$  porque las coordenadas principales de  $\mathbf{X}$  son independientes de las coordenadas de  $\mathbf{Z}$  debido al proceso de ortogonalización, que produce el mismo resultado de McCullagh & Nelder (1989). Sin embargo, este procedimiento no se utilizó debido a la dificultad de distinguir entre cuál coordenada debe ir al modelo de media y cuál al modelo de dispersión variable.

#### Modelo de regresión beta basado en distancia con dispersión constante

Escribiendo  $g_1(\mu_i) = g(\mu_i)$  y  $g_2(\phi_i) = \phi_i$  en (2.9), donde  $g(\cdot)$  es una función de enlace, los modelos presentados en (2.9) se pueden escribir como

$$g(\mu_i) = \eta_{1i} = \mathbf{x}_i^t \boldsymbol{\beta} \quad \text{y} \quad \phi_i = \phi$$

donde  $\phi > 0$  es constante, es decir,  $\mathbf{Z} = \mathbf{1}$  y  $\mathbf{T}_2 = \mathbf{I}$ . Por lo tanto, el vector score definido en (2.17) toma la siguiente forma

$$\mathbf{U}_{\beta}(\boldsymbol{\beta}, \phi) = \phi \mathbf{X}^t \mathbf{T}(\mathbf{y}^* - \boldsymbol{\mu}^*) \quad \text{y} \quad \mathbf{U}_{\phi}(\boldsymbol{\beta}, \phi) = \sum_{i=1}^n \vartheta_i \quad (2.20)$$

donde  $\mathbf{T} = \text{diag}(d\mu_i/d\eta_i)$ , y  $\mathbf{y}^*$ ,  $\boldsymbol{\mu}^*$  y  $\vartheta_i$  se definieron anteriorente. Por otra parte, las matrices  $\mathbf{P}$  y  $\mathbf{W}$  definidas en (2.18) se pueden expresar como

$$\mathbf{P} = \begin{pmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \quad \text{y} \quad \mathbf{W} = \begin{pmatrix} \mathbf{W}_{\beta\beta} & \mathbf{W}_{\beta\phi} \\ \mathbf{W}_{\phi\beta} & \mathbf{W}_{\phi\phi} \end{pmatrix}$$

con  $\mathbf{W}_{\beta\beta} = \text{diag}\left(\phi^2 a_i \left(\frac{d\mu_i}{d\eta_i}\right)\right)$ ,  $\mathbf{W}_{\beta\phi} = \text{diag}\left(\phi\{\mu_i a_i - \psi'((1 - \mu_i)\phi)\} \left(\frac{d\mu_i}{d\eta_i}\right)\right)$  y  $\mathbf{W}_{\phi\phi} = \text{diag}(b_i)$ , donde  $a_i$  y  $b_i$  se definieron en la subsección 2.2.2. Por lo tanto, la matriz de información de Fisher para  $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \phi)^t$  es  $\mathbf{K}(\boldsymbol{\theta}) = \mathbf{P}^t \mathbf{W} \mathbf{P}$ .

### Modelo de regresión beta basado en distancia con dispersión variable

En esta clase de modelos lineales generalizados, los modelos de DBBR con dispersión variable a diferencia de con dispersión constante, presentados anteriormente, permiten que el parámetro de precisión  $\phi$  varíe a través de una estructura de regresión lineal. Más precisamente, la ecuación (2.9) para este modelo se convierte en

$$g_1(\mu_i) = \eta_{1i} = \mathbf{x}_i^t \boldsymbol{\beta} \quad \text{y} \quad g_2(\phi_i) = \eta_{2i} = \mathbf{z}_i^t \boldsymbol{\alpha}$$

donde  $\boldsymbol{\beta} \in \mathbb{R}^{k_1+1}$  y  $\boldsymbol{\alpha} \in \mathbb{R}^{k_2+1}$ . El vector score para este modelo es idéntico al que se presentó en la expresión (2.17). Además, con  $\mathbf{W}$  definida como en (2.18), la matriz de información de Fisher para el vector de parámetros  $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{\alpha}^t)^t$  es  $\mathbf{K}(\boldsymbol{\theta}) = \mathbf{P}^t \mathbf{W} \mathbf{P}$ .

#### 2.2.4 Inferencia para muestras grandes

En esta sección, se desarrolla la razón de verosimilitud, las pruebas de Score y de Wald para los parámetros del modelo DBBR con dispersión variable. También, se obtienen los intervalos de confianza para los parámetros de regresión asociados a los modelos de media y de dispersión variable.

Suponiendo que  $\mathbf{J}(\boldsymbol{\theta}) = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{K}(\boldsymbol{\theta})$  existe y es no singular, es decir,

$$\mathbf{K}^{-1} = \mathbf{K}^{-1}(\boldsymbol{\theta}) = \mathbf{K}^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{pmatrix} \mathbf{K}^{\beta\beta} & \mathbf{K}^{\beta\alpha} \\ \mathbf{K}^{\alpha\beta} & \mathbf{K}^{\alpha\alpha} \end{pmatrix}$$

Además, no es difícil mostrar que  $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_{k_1+k_2}(\mathbf{0}, \mathbf{J}^{-1}(\boldsymbol{\theta}))$ , donde  $\xrightarrow{d}$  denota convergencia en distribución (McCullagh & Nelder 1989, Dobson 2002).

En el modelo DBBR, se considera la hipótesis

$$H_0 : \boldsymbol{\theta}_* = (\boldsymbol{\beta}_{r_1}^t, \boldsymbol{\alpha}_{r_2}^t)^t = \boldsymbol{\theta}_*^{(0)} \quad \text{contra} \quad H_1 : \boldsymbol{\theta}_* \neq \boldsymbol{\theta}_*^{(0)} \quad (2.21)$$

donde  $\boldsymbol{\theta}_* = (\boldsymbol{\beta}_{r_1}^t, \boldsymbol{\alpha}_{r_2}^t)^t = (\beta_1, \dots, \beta_{r_1}, \alpha_1, \dots, \alpha_{r_2})^t$  y  $\boldsymbol{\theta}_*^{(0)} = (\beta_1^{(0)}, \dots, \beta_{r_1}^{(0)}, \alpha_1^{(0)}, \dots, \alpha_{r_2}^{(0)})^t$  para  $r_1 \leq k_1$  y  $r_2 \leq k_2$ .

El estadístico de razón de log-verosimilitud para juzgar (2.21) está dado por

$$w_1 = 2 \left( l(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}) - l(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) \right) \quad (2.22)$$

donde  $l(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$  es la función de log-verosimilitud evaluada en los valores estimados y  $(\widetilde{\boldsymbol{\beta}}^t, \widetilde{\boldsymbol{\alpha}}^t)^t$  es el estimador máximo verosímil restringido de  $(\boldsymbol{\beta}^t, \boldsymbol{\alpha}^t)^t$  obtenido bajo la hipótesis nula. Bajo las condiciones usuales de regularidad y bajo  $H_0$ ,  $w_1 \xrightarrow{d} \chi_{(r_1+r_2)}^2$ . Así, el juzgamiento de (2.21) se puede realizar utilizando los valores críticos de las aproximaciones a la distribución asintótica  $\chi_{(r_1+r_2)}^2$ .

En particular para juzgar  $H_0 : \boldsymbol{\beta}_{r_1} = \boldsymbol{\beta}_{r_1}^{(0)}$  contra  $H_1 : \boldsymbol{\beta}_{r_1} \neq \boldsymbol{\beta}_{r_1}^{(0)}$ , se describe el uso del estadístico score: sea  $\mathbf{U}_{1\beta}$  un vector de dimensión  $r_1 \times 1$  que contiene los primeros  $r_1$  elementos de la función score para  $\boldsymbol{\beta}$  y sea  $\mathbf{K}_{11}^{\beta\beta}$  una matriz de tamaño  $(r_1 \times r_1)$  formado a partir de las primeras  $r_1$  filas y las primeras  $r_1$  columnas de  $\mathbf{K}^{\beta\beta}$ . De la ecuación (2.17), se puede demostrar que  $\mathbf{U}_{1\beta} = \mathbf{X}_1^t \boldsymbol{\Upsilon} \mathbf{T}_1 (\mathbf{y}^* - \boldsymbol{\mu}^*)$ , donde  $\mathbf{X}$  es particionada como  $[\mathbf{X}_1 : \mathbf{X}_2]$  siguiendo la partición de  $\boldsymbol{\beta}$ . Por lo tanto, la estadística Score de Rao para juzgar  $H_0 : \boldsymbol{\beta}_{r_1} = \boldsymbol{\beta}_{r_1}^{(0)}$  está dada por

$$w_2 = \widetilde{\mathbf{U}}_{1\beta}^t \widetilde{\mathbf{K}}_{11}^{\beta\beta} \widetilde{\mathbf{U}}_{1\beta} \quad (2.23)$$

donde las tildes indican que las cantidades son evaluadas en el estimador de máxima verosimilitud restringida. Bajo las usuales condiciones de regularidad y bajo  $H_0$ ,  $w_2 \xrightarrow{d} \chi_{r_1}^2$ .

La inferencia asintótica también se puede realizar usando el estadístico de Wald, el cual está dado por

$$w_3 = (\widehat{\boldsymbol{\beta}}_{r_1} - \boldsymbol{\beta}_{r_1}^{(0)})^t \left( \widehat{\mathbf{K}}_{11}^{\beta\beta} \right)^{-1} (\widehat{\boldsymbol{\beta}}_{r_1} - \boldsymbol{\beta}_{r_1}^{(0)}) \quad (2.24)$$

donde  $\widehat{\mathbf{K}}_{11}^{\beta\beta}$  es igual a  $\mathbf{K}_{11}^{\beta\beta}$  evaluada en el estimador máximo verosímil sin restricción y  $\widehat{\boldsymbol{\beta}}_{r_1}$  es el estimador máximo verosímil de  $\boldsymbol{\beta}_{r_1}$ . Bajo las condiciones usuales de regularidad y bajo  $H_0$ ,  $w_3 \xrightarrow{d} \chi_{r_1}^2$ .

En particular, para juzgar la significancia del  $j$ -ésimo parámetro de regresión  $\beta_j$ ,  $j = 1, \dots, k_1$ , se puede utilizar la raíz cuadrada positiva del estadístico de Wald, es decir, si se desea contrastar la hipótesis

$$H_0 : \beta_j = 0 \quad \text{contra} \quad H_1 : \beta_j \neq 0$$

El estadístico de prueba está dado por

$$w_{\beta_j} = \frac{\hat{\beta}_j}{ee(\hat{\beta}_j)}, \quad j = 0, 1, \dots, k_1 \quad (2.25)$$

donde  $ee(\hat{\beta}_j)$  es el error estándar asintótico del estimador máximo verosímil de  $\hat{\beta}_j$  obtenido de la inversa de la matriz de información de Fisher evaluada en las estimaciones máximo verosímiles. El límite de la distribución del estadístico de prueba dado en (2.25) bajo  $H_0$  cierta es una distribución normal estándar.

Además, un intervalo de confianza aproximado  $(1 - q)100\%$  para  $\beta_j$ ,  $j = 1, \dots, k_1$  y  $0 < q < 1/2$  está dado por

$$IC_{(1-q)100\%}(\beta_j) = \left( \hat{\beta}_j - \Phi_{(1-q/2)}^{-1} ee(\hat{\beta}_j); \hat{\beta}_j + \Phi_{(1-q/2)}^{-1} ee(\hat{\beta}_j) \right)$$

donde  $\Phi^{-1}$  es la inversa de función de distribución normal estándar. Observe que este intervalo no se construye en términos del odds ratio porque éste se obtiene de las coordenadas principales usando distancias; el odds ratio no tiene el mismo sentido que el odds ratio en el modelo clásico.

Si se desean calcular regiones de confianza aproximadas para grupos de parámetros, éstas se pueden obtener utilizando cualquiera de las tres pruebas para muestras grandes dadas en (2.23), (2.24) y (2.25).

De forma similar, para juzgar  $H_0 : \alpha_{r_2} = \alpha_{r_2}^{(0)}$  contra  $H_1 : \alpha_{r_2} \neq \alpha_{r_2}^{(0)}$  se describe el estadístico de score: sea  $\mathbf{U}_{1\alpha}$  el vector de dimensión  $r_2$  que contiene los primeros  $r_2$  elementos de la función score para  $\alpha$ . Además, sea  $\mathbf{K}_{11}^{\alpha\alpha}$  una matriz de tamaño  $(r_2 \times r_2)$  formado por las primeras  $r_2$  filas y las primeras  $r_2$  columnas de  $\mathbf{K}^{\alpha\alpha}$ . Por lo tanto, de la ecuación (2.17) se puede mostrar que  $\mathbf{U}_{1\alpha} = \mathbf{Z}_2^t \mathbf{T}_2 \boldsymbol{\theta}$ , donde  $\mathbf{Z}$  es particionada como  $[\mathbf{Z}_1 : \mathbf{Z}_2]$  siguiendo la misma partición de  $\alpha$ . Por lo tanto, la estadística Score de Rao para juzgar  $H_0 : \alpha_{r_2} = \alpha_{r_2}^{(0)}$  está dada por

$$w_4 = \widetilde{\mathbf{U}}_{1\alpha}^t \widetilde{\mathbf{K}}_{11}^{\alpha\alpha} \widetilde{\mathbf{U}}_{1\alpha} \quad (2.26)$$

donde las tildes indican que las cantidades son evaluadas en el estimador restringido de máxima verosimilitud. Bajo las condiciones usuales de regularidad y bajo  $H_0$ ,  $w_4 \xrightarrow{d} \chi_{r_2}^2$ .

La inferencia asintótica también se puede realizar usando el estadístico de Wald, el cual está dado por

$$w_5 = (\widehat{\alpha}_{r_2} - \alpha_{r_2}^{(0)})^t \left( \widehat{\mathbf{K}}_{11}^{\alpha\alpha} \right)^{-1} (\widehat{\alpha}_{r_2} - \alpha_{r_2}^{(0)}) \quad (2.27)$$

donde  $\widehat{\mathbf{K}}_{11}^{\alpha\alpha}$  es igual a  $\mathbf{K}_{11}^{\alpha\alpha}$  evaluada en el estimador máximo verosímil sin restricción y  $\widehat{\alpha}_{r_2}$  es el estimador máximo verosímil de  $\alpha_{r_2}$ . Bajo las usuales condiciones de regularidad y bajo  $H_0$ ,  $w_5 \xrightarrow{d} \chi_{r_2}^2$ .

En este caso, para juzgar la significancia del  $h$ -ésimo parámetro de regresión  $\alpha_h$ ,  $h = 1, \dots, k_2$ , se puede utilizar la raíz cuadrada positiva del estadístico de Wald, es decir, si se desea contrastar la hipótesis

$$H_0 : \alpha_h = 0 \quad \text{contra} \quad H_1 : \alpha_h \neq 0$$

El estadístico de prueba está dado por

$$w_{\alpha_h} = \frac{\hat{\alpha}_h}{ee(\hat{\alpha}_h)}, \quad h = 0, 1, \dots, k_2 \quad (2.28)$$

donde  $ee(\hat{\alpha}_s)$  es el error estándar asintótico del estimador máximo verosímil de  $\hat{\alpha}_h$  obtenido de la inversa de la matriz de Fisher evaluada en las estimaciones máximo verosímiles. El límite de la distribución del estadístico dado en (2.28) bajo  $H_0$  cierta es una normal estándar.

Además, un intervalo de confianza aproximado del  $(1 - q)100\%$  para  $\alpha_h$ ,  $h = 1, \dots, k_2$  y  $0 < q < 1/2$ , está dado por

$$IC_{(1-q)100\%}(\alpha_h) = \left( \hat{\alpha}_h - \Phi_{(1-q/2)}^{-1} ee(\hat{\alpha}_h); \hat{\alpha}_h + \Phi_{(1-q/2)}^{-1} ee(\hat{\alpha}_h) \right)$$

## 2.3 Ajuste, selección, diagnóstico y predicción del modelo DBBR

En esta subsección se exponen algunas medidas de bondad de ajuste, se hace la selección de las coordenadas principales, el diagnóstico y predicción del modelo DBBR, y se estudia el problema de datos faltantes.

### 2.3.1 Medidas de bondad de ajuste

Después de ajustar el modelo DBBR es importante llevar a cabo un análisis de diagnóstico, con la finalidad de verificar la bondad de ajuste del modelo estimado. Una medida global de la variación explicada se obtiene calculando el pseudo  $R^2$  definido como

$$R_{k_1}^2 = r^2(\hat{\boldsymbol{\eta}}_1, g_1(\mathbf{y})) \quad 0 \leq R_{k_1}^2 \leq 1 \quad (2.29)$$

donde  $r(\hat{\boldsymbol{\eta}}_1, g_1(\mathbf{y}))$  es el coeficiente de correlación muestral entre  $\hat{\boldsymbol{\eta}}_1$  y  $g_1(\mathbf{y})$ . Cuando  $R_{k_1}^2 = 1$  existe un acuerdo perfecto entre  $\hat{\boldsymbol{\eta}}_1$  y  $g_1(\mathbf{y})$ , por lo tanto entre  $\hat{\boldsymbol{\mu}}$  e  $\mathbf{y}$ . Se observa además que las estimaciones de  $\boldsymbol{\eta}_1$  dependen de la estimación del vector de parámetro  $\boldsymbol{\alpha}$ , es decir, del modelo de dispersión variable,  $\boldsymbol{\eta}_2$ . Por lo tanto,  $R_{k_1}^2$  depende de sí se modela la dispersión variable o no.

La discrepancia del modelo ajustado se puede determinar a través de qué tanto el modelo ajustado es significativamente diferente del modelo saturado, el cual contiene tantos parámetros como observaciones hay en el modelo. Para ello, sea

$$D(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\phi}) = \sum_{i=1}^n 2 [l_i(\tilde{\mu}_i, \phi_i) - l_i(\mu_i, \phi_i)]$$

donde  $\tilde{\mu}_i$  es el valor de  $\mu_i$  que resuelve  $\partial l_i / \partial \mu_i = 0$ , es decir,  $\phi_i(y_i^* - \mu_i^*) = 0$ . Cuando  $\min\{\phi_1, \dots, \phi_n\}$  es grande,  $\mu_i^* \approx \log\{\mu_i / (1 - \mu_i)\}$ , y de esto se sigue que  $\tilde{\mu}_i \approx y_i$ .

Para  $\boldsymbol{\phi}$  conocido, la medida de discrepancia entre los dos modelos es  $D(\mathbf{y}; \hat{\boldsymbol{\mu}}, \boldsymbol{\phi})$ , donde  $\hat{\boldsymbol{\mu}}$  es el estimador de máxima verosimilitud de  $\boldsymbol{\mu}$  bajo el modelo estudiado. Cuando  $\boldsymbol{\phi}$  es desconocido, una aproximación de esta cantidad es

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\phi}}) = \sum_{i=1}^n r_{D_i}^2 \quad (2.30)$$

que es conocida como la *deviance*, y donde

$$r_{D_i} = \text{sign}(y_i - \hat{\mu}_i) \{2 [l_i(\tilde{\mu}_i, \tilde{\phi}_i) - l_i(\hat{\mu}_i, \hat{\phi}_i)]\}^{1/2} \quad (2.31)$$

con  $l_i(\tilde{\mu}_i, \tilde{\phi}_i)$  la máxima verosimilitud del modelo saturado y  $l_i(\hat{\mu}_i, \hat{\phi}_i)$  es la máxima verosimilitud del modelo restringido bajo  $H_0$ . Ferrari & Cribari-Neto (2004) llaman a  $r_{D_i}$  la  $i$ -ésima residual de deviance porque una observación con un valor absoluto grande de  $r_{D_i}$  puede ser vista como una discrepancia. Tal como se esperaba, la log-verosimilitud asociada al modelo saturado debe ser mayor que la asociada a un modelo con  $k_1 + k_2 < n$  parámetros.

El estadístico (3.36) tiene una distribución asintótica  $\chi_{(n-k_1-k_2)}^2$ . Claramente es deseable una deviance pequeña, y si esta no es significativa, la conclusión es que el desempeño del modelo bajo estudio no es significativamente diferente del modelo saturado; por lo tanto, la estimación de los otros parámetros involucrados en el modelo saturado es innecesaria.

### 2.3.2 Selección de variables para el modelo beta basado en distancias

A continuación se presenta cómo elegir la dimensión de predicción adecuada. Se quiere obtener dos permutaciones (una de  $k_1$  coordenadas principales para el modelo de media y otra de  $k_2$  coordenadas principales para el modelo de dispersión variable) de tal manera que  $\mathbf{X}$  y  $\mathbf{Z}$  son dos conjuntos óptimos de  $k_1$  y  $k_2$  variables predictoras en (1.4), de acuerdo con algún criterio adecuado. Para este objetivo, se adaptan las propuestas del modelo de regresión clásico formuladas por Cuadras et al. (1996) con la finalidad de decidir, si una columna

de  $\mathbf{X}$  en el modelo de media o una columna de  $\mathbf{Z}$  en el modelo de dispersión variable, debe ser incluida o eliminada.

Lo anterior es de interés para evitar el problema de obtener un pseudo  $R^2 \simeq 1$  en el modelo de media cuando el rango de  $\mathbf{X}$  es  $k_1 = n - 1$ . Por lo tanto, es necesario considerar solo las coordenadas principales  $(\mathbf{X}_1, \dots, \mathbf{X}_{k_1}, \dots, \mathbf{X}_{n-1})$  más correlacionados de  $\mathbf{B}_v$  con la variable respuesta  $\mathbf{y}$ . Algunas alternativas para incluir o eliminar una variable predictora (una columna de  $\mathbf{X}$ ) son:

M1. Las columnas de  $\mathbf{X}$  se organizan en orden decreciente de acuerdo a la correlación absoluta con  $g_1(\mathbf{y})$ , es decir,

$$r^2(\hat{\boldsymbol{\eta}}_1(\mathbf{X}_1), g_1(\mathbf{y})) > \dots > r^2(\hat{\boldsymbol{\eta}}_1(\mathbf{X}_{k_1}), g_1(\mathbf{y})) > \dots > r^2(\hat{\boldsymbol{\eta}}_1(\mathbf{X}_{n-1}), g_1(\mathbf{y}))$$

donde  $r(\hat{\boldsymbol{\eta}}_1(\mathbf{X}_j), g_1(\mathbf{y}))$  es el coeficiente de correlación entre  $\hat{\boldsymbol{\eta}}_1(\mathbf{X}_j)$  y  $g_1(\mathbf{y})$ , incluyendo sólo la  $j$ -ésima ( $j = 1, \dots, k_1, \dots, n - 1$ ) coordenada principal. Esto se hace dejando las mismas funciones de enlace en  $g_1$  y  $g_2$  (por ejemplo, en todos los modelos, el enlace logit para  $g_1$  y el enlace log para  $g_2$ ). Con este procedimiento, las coordenadas principales menos correlacionadas con  $\mathbf{y}$  se eliminan, es decir,  $n - k_1$  ( $k_1 < n$ ) coordenadas en el modelo final no se consideran.

M2. Lo mismo que en M1, pero utilizando el estadístico (2.25) en valor absoluto para seleccionar las coordenadas principales que deben ser incluidas en el modelo reducido DB, es decir,

$$w_{\beta_1} > \dots > w_{\beta_{k_1}} > \dots > w_{\beta_{n-1}}$$

M3. Se lleva a cabo de manera similar para la selección del número de variables explicativas en la regresión multivariada, siguiendo el conocido  $C_p$ -Mallows. Esto es, se construye una gráfica representando los puntos  $(j, 1 - c(j))$ ;  $j = 1, \dots, k_1, \dots, n - 1$ ; de este modo, el punto en el que hay una caída significativa debido a la falta de predecibilidad es determinado. La predecibilidad,  $c(j)$ , está dada por la expresión

$$c(0) = 0, \quad c(j) = \frac{\sum_{i=1}^j r^2(\hat{\boldsymbol{\eta}}_1(\mathbf{X}_i), g_1(\mathbf{y})) \lambda_{x_i}}{\sum_{i=1}^{n-1} r^2(\hat{\boldsymbol{\eta}}_1(\mathbf{X}_i), g_1(\mathbf{y})) \lambda_{x_i}}$$

donde  $\lambda_{x_i}$  es el  $i$ -ésimo valor propio asociado a  $\mathbf{X}_i$ ,  $i = 1, 2, \dots, k_1, \dots, n - 1$  (Cuadras et al. 1996). La selección del número de coordenadas principales corresponde a la caída significativa de  $1 - c(j)$ .

Finalmente, siguiendo cualquiera de los tres métodos presentados anteriormente, las coordenadas principales  $\mathbf{X}_{k_1+1}, \dots, \mathbf{X}_{n-1}$  se deben eliminar o no considerar en el modelo de medias.

Ahora supóngase que  $(\boldsymbol{\eta}_1 - \widehat{\beta}_0 \mathbf{1}) \in E_{n-1}^{(1)}$ , donde  $E_{n-1}^{(1)}$  es un subespacio generado por las columnas de  $\mathbf{X}$ , entonces el modelo de media se puede expresar como

$$\begin{aligned}\boldsymbol{\eta}_1 &= \widehat{\beta}_0 \mathbf{1} + \mathbf{X}_{(k_1)} \widehat{\boldsymbol{\beta}}_{(k_1)} + \mathbf{X}_{(n-1-k_1)} \widehat{\boldsymbol{\beta}}_{(n-1-k_1)} \\ &= \widehat{\boldsymbol{\eta}}_1 + \widehat{\boldsymbol{e}}_x\end{aligned}$$

donde  $\mathbf{X}_{(n-1-k_1)}$  contiene las columnas que son pobremente correlacionadas con  $g_1(\boldsymbol{\mu})$ , así  $\mathbf{X}_{(n-1-k_1)} \widehat{\boldsymbol{\beta}}_{(n-1-k_1)}$  puede ser interpretado como un término de error  $\widehat{\boldsymbol{e}}_x$ , y en este sentido, el modelo es consistente con el modelo de regresión beta lineal clásico.

Por otro lado, para la selección de las  $k_2$  coordenadas principales asociadas a  $\mathbf{Z}$  en el modelo (2.8), se puede utilizar cualquiera de los procedimientos presentados para la selección de coordenadas en el modelo de medias. En el caso del modelo de dispersión variable, una coordenada principal debe incluirse en (2.8) siguiendo cualquiera de las siguientes alternativas:

S1. Utilizando (3.36), pero con

$$r_{D_i} = r_{D_i}(\mathbf{Z}_h) = \text{sign}(y_i - \widehat{\mu}_i) \left\{ 2 \left( l_i(\widehat{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\phi}}_i(\mathbf{Z}_h)) - l_i(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\phi}}_i(\mathbf{Z}_h)) \right) \right\}^{1/2}$$

donde  $r_{D_i}(\mathbf{Z}_h)$  es el residual obtenido a partir de la inclusión de  $\mathbf{Z}_h$ ,  $h = 1, \dots, k_2, \dots, n-1$  en el modelo. De esta manera, las deviances están ordenadas de menor a mayor, las primeras  $k_2$  deviances son

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\phi}}(\mathbf{Z}_1)) < D(\mathbf{y}; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\phi}}(\mathbf{Z}_2)) < \dots < D(\mathbf{y}; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\phi}}(\mathbf{Z}_{k_2}))$$

S2. Se utiliza el estadístico presentado en (2.28) en valor absoluto considerando cada variable por separado, luego se ordena de mayor a menor los valores del estadístico, y finalmente, se seleccionan las primeras  $k_2$  coordenadas principales, es decir,

$$w_{\alpha_1} > \dots > w_{\alpha_{k_2}} \dots > w_{\alpha_{n-1}}$$

S3. Ordenar los valores propios asociados con cada columna de  $\mathbf{Z}$  de mayor a menor, es decir,  $\lambda_{z_1} > \dots > \lambda_{z_{k_2}} > \dots > \lambda_{z_{n-1}}$ , y luego, se seleccionan las primeras  $k_2$  coordenadas principales para formar  $\mathbf{Z}$ . Sin embargo, como se indica en Cuadras & Fortiana (1993), si hay valores propios pequeños que estén correlacionados con la variable respuesta, las coordenadas principales asociadas se correlacionarán con el “ ruido ” en vez de con la variabilidad principal de los datos.



Al igual que en el modelo de medias, siguiendo cualquiera de los tres métodos presentados anteriormente, las coordenadas principales  $\mathbf{Z}_{k_1+1}, \dots, \mathbf{Z}_{n-1}$  se deben eliminar o no considerar en el modelo de dispersión variable. Observe que si las mismas coordenadas principales se incluyen en los modelos de media y dispersión variable,  $\boldsymbol{\eta}_1$  y  $\boldsymbol{\eta}_2$ , las coordenadas principales  $\mathbf{X}$  y  $\mathbf{Z}$  coinciden.

Adicionalmente, supóngase que  $(\boldsymbol{\eta}_2 - \hat{\alpha}_0 \mathbf{1}) \in E_{n-1}^{(2)}$ , donde  $E_{n-1}^{(2)}$  es un subespacio generado por las columnas de  $\mathbf{Z}$ , entonces el modelo de dispersión variable se puede expresar como

$$\begin{aligned}\boldsymbol{\eta}_2 &= \hat{\alpha}_0 \mathbf{1} + \mathbf{Z}_{(k_2)} \hat{\boldsymbol{\alpha}}_{(k_2)} + \mathbf{Z}_{(n-1-k_2)} \hat{\boldsymbol{\alpha}}_{(n-1-k_2)} \\ &= \hat{\boldsymbol{\eta}}_2 + \hat{\boldsymbol{e}}_z\end{aligned}$$

donde  $\mathbf{Z}_{(n-1-k_2)}$  contiene las columnas que son pobremente correlacionadas con  $g_2(\boldsymbol{\phi})$ , así  $\mathbf{Z}_{(n-1-k_2)} \hat{\boldsymbol{\alpha}}_{(n-1-k_2)}$  puede ser interpretado como un término de error  $\hat{\boldsymbol{e}}_z$ .

### 2.3.3 Medidas de diagnóstico

Debido a que la  $i$ -ésima observación en (2.31) contribuye una cantidad  $r_{D_i}^2$  a la deviance, una observación con un alto valor absoluto de  $r_{D_i}$  puede ser atípica. Por esto, a  $r_{D_i}$  se le llama el  $i$ -ésimo residual de deviance.

Hay varias clases de residuos para los modelos DBBR, una alternativa natural son los residuales de Pearson, los cuales Ferrari & Cribari-Neto (2004) y Espinheira et al. (2008b) llaman residuales ordinarios estandarizados y se definen como

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{Var}}(y_i)}}$$

donde  $\widehat{\text{Var}}(y_i) = \hat{\mu}_i (1 - \hat{\mu}_i) / (1 + \hat{\phi}_i)$ ,  $\hat{\mu}_i = g_1^{-1}(\mathbf{x}_i^t \hat{\boldsymbol{\beta}})$  y  $\hat{\phi}_i = g_2^{-1}(\mathbf{z}_i^t \hat{\boldsymbol{\alpha}})$ . Similarmente, se pueden definir los residuales de deviance como se hizo en la ecuación (3.36) vía las contribuciones del signo al exceso de la verosimilitud. Además, Espinheira et al. (2008b) propusieron unos residuales con mejores propiedades que los residuales de Pearson, éstos son llamados los residuales estandarizados ponderados 2:

$$r_{SW2_i} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\hat{\nu}_i (1 - h_{ii})}}$$

donde  $y_i^* = \log(y_i / (1 - y_i))$  y  $\hat{\mu}_i^* = \psi(\hat{\mu}_i \hat{\phi}_i) - \psi((1 - \hat{\mu}_i) \hat{\phi}_i)$ ,  $\psi(\cdot)$  denota la función digamma. Además  $\hat{\nu}_i = [\psi'(\hat{\mu}_i \hat{\phi}_i) + \psi'((1 - \hat{\mu}_i) \hat{\phi}_i)]$  y  $h_{ii}$  es el  $i$ -ésimo elemento de la matriz hat definida en (2.32).

Un gráfico de estos residuales contra las observaciones indexadas ( $i$ ) puede evidenciar patrones no detectables. Adicionalmente, una tendencia detectable

en el gráfico de  $r_{D_i}$  o  $r_{P_i}$  o  $r_{SW2_i}$  contra  $g_1^{-1}(\hat{\eta}_{1i})$  puede sugerir una falta de especificación de la función de enlace.

Dado que la distribución de los residuales no se conoce, los gráficos Half-Normal con valores de franjas simuladas son una buena estrategia de diagnóstico, véase Atkinson (1985), Neter et al. (1996), Ferrari & Cribari-Neto (2004) y Espinheira et al. (2008b) para mayores detalles. La idea principal es aumentar el gráfico Half-Normal usual al adicionar una franja simulada, la cual es útil para decidir cuándo los residuales observados son consistentes con el modelo ajustado. Este tipo de gráficos se obtienen siguiendo los siguientes pasos (Ferrari & Cribari-Neto 2004):

1. Ajuste el modelo y genere una muestra simulada de  $n$  observaciones independientes utilizando el modelo ajustado como si éste fuera el modelo verdadero.
2. Ajuste un modelo con los datos de la muestra generada y calcule los valores absolutos ordenados de los residuales.
3. Repita los pasos 1. y 2.  $l$  veces.
4. Considere los  $n$  conjuntos de  $l$  estadísticas de orden, para cada conjunto calcule el promedio, y los valores mínimo y máximo.
5. Grafique los valores obtenidos y los residuales ordenados de la muestra original contra los puntajes Half-Normal de la forma:  $\Phi^{-1}((i + n - 1/8)/(2n + 1/2))$ .

Los valores mínimo y máximo de las  $l$  estadísticas de orden generan la cubierta. Atkinson (1985) sugiere usar  $l = 19$ , ya que la probabilidad de que un residual absoluto caiga más allá de la banda superior proporcionada por la franja es aproximadamente igual a  $1/20 = 0.05$ . Las observaciones correspondientes a los residuales absolutos que se encuentren fuera de los límites proporcionados por la franja simulada se deben estudiar a mayor profundidad ya que pueden ser observaciones atípicas o influyentes. Adicionalmente, si una proporción considerable de puntos cae fuera de la cubierta, se tiene evidencia en contra del adecuado ajuste del modelo propuesto.

Después de hacer una identificación de las observaciones influyentes y un análisis de residuales, se hace uso de la matriz “sombrero” de costumbre,

$$\mathbf{H} = \mathbf{W}_{\beta\beta}^{1/2} \mathbf{X} (\mathbf{X}^t \mathbf{W}_{\beta\beta} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W}_{\beta\beta}^{1/2} \quad (2.32)$$

cuando los  $\phi_i$ 's son conocidos y  $\min\{\phi_1, \dots, \phi_n\}$  es grande. En este caso, el  $i$ -ésimo elemento de la diagonal de  $\mathbf{W}_{\beta\beta}^{1/2}$  es aproximadamente igual a

$\{g'_1(\mu_i)\text{Var}(y_i)^{1/2}\}^{-1}$  (Ferrari & Cribari-Neto 2004, Espinheira et al. 2008a). Cuando los  $\phi_i$ 's son desconocidos se pueden revisar en Ferrari et al. (2011).

Una medida de influencia de cada observación sobre los parámetros de regresión estimados es la distancia de Cook (Cook 1977) dada por  $k_1^{-1}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^t \mathbf{X}^t \mathbf{W}_{\beta\beta} \mathbf{X} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})$ , donde  $\widehat{\boldsymbol{\beta}}_{(i)}$  es el vector de parámetros estimados sin la  $i$ -ésima observación. Esta medida es la distancia al cuadrado entre  $\widehat{\boldsymbol{\beta}}$  y  $\widehat{\boldsymbol{\beta}}_{(i)}$ . Para evitar el ajuste del modelo  $n + 1$  veces, se utilizará la aproximación usual de la distancia de Cook dada por

$$C_i = \frac{h_{ii} r_{D_i}^2}{k_1(1 - h_{ii})^2}$$

Esta expresión combina leverage y residuales. Además es común hacer un gráfico de  $C_i$  contra  $i$  para verificar posibles observaciones influyentes. También se pueden utilizar otras medidas de diagnóstico, tales como las medidas de influencia local (Cook 1986).

### 2.3.4 Predicción de un nuevo individuo

Si se supone que sobre las variables mixtas explicativas iniciales se ha observado un nuevo individuo  $n + 1$  del que se conoce las observaciones sobre las variables independientes,  $\mathbf{v}_{n+1} = (v_{(n+1)1}, \dots, v_{(n+1)p_1})^t$ . Con estas observaciones se puede calcular las distancias entre el individuo  $n + 1$  y cada uno de los individuos involucrados en el modelo planteado en (2.4), es decir,

$$d_{(n+1)i} = d(\mathbf{v}_{n+1}, \mathbf{v}_i), \quad i = 1, \dots, n$$

A partir de estas distancias, se puede hacer una predicción utilizando un resultado presentado en Gower (1968), Gower (1971) y Cuadras (1989), el cual relaciona el vector  $\mathbf{d} = (d_{(n+1)1}^2, \dots, d_{(n+1)n}^2)^t$  de distancias al cuadrado con el vector  $\mathbf{x}_{n+1} = (x_{(n+1)1}, \dots, x_{(n+1)(n-1)})^t$  de coordenadas principales atribuible al nuevo individuo:

$$d_{(n+1)i}^2 = (\mathbf{x}_{n+1} - \mathbf{x}_i)^t (\mathbf{x}_{n+1} - \mathbf{x}_i) = \mathbf{x}_{n+1}^t \mathbf{x}_{n+1} + \mathbf{x}_i^t \mathbf{x}_i - 2\mathbf{x}_{n+1}^t \mathbf{x}_i \quad (2.33)$$

Sumando para  $i$  de 1 a  $n$ , y teniendo en cuenta que las columnas de la matriz  $\mathbf{X}$  de coordenadas suman 0, se obtiene la siguiente expresión

$$\sum_{i=1}^n d_{(n+1)i}^2 = n\mathbf{x}_{n+1}^t \mathbf{x}_{n+1} + \text{tr}(\mathbf{B}_v)$$

Sustituyendo esta última ecuación en (2.33) y haciendo algunos procedimientos algebraicos, se tiene

$$\mathbf{x}_{n+1} = \frac{1}{2} \boldsymbol{\Lambda}_x^{-1} \mathbf{X}^t (\mathbf{b} - \mathbf{d}) \quad (2.34)$$

donde  $\mathbf{b} = (b_{11}, \dots, b_{nm})^t$ ,  $\Lambda_x^{-1} = (\mathbf{X}^t \mathbf{X})^{-1}$  y  $b_{ii} = \mathbf{x}_i^t \mathbf{x}_i$ ,  $i = 1, \dots, n$ .

La predicción está entonces dada por

$$\hat{\eta}_{(n+1)\beta} = \hat{\beta}_0 + \mathbf{x}_{n+1}^t \hat{\beta}$$

Al considerar el modelo DBBR en dimensión  $k_1$  y hacer la siguiente partición

$$\mathbf{x}_{n+1} = \begin{pmatrix} \mathbf{x}_{(k_1)} \\ \mathbf{x}_{(n-1-k_1)} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_{(k_1)} & \mathbf{X}_{(n-1-k_1)} \end{pmatrix} \text{ y } \Lambda = \begin{pmatrix} \Lambda_{x_{k_1}} & \mathbf{0} \\ \mathbf{0} & \Lambda_{x_{n-1-k_1}} \end{pmatrix}$$

donde  $\mathbf{x}_{(k_1)} = (x_1, \dots, x_{k_1})^t$  es el vector de  $k_1$  coordenadas principales relacionadas a la  $k_1$ -dimensiones predictivas asociadas con al  $n + 1$  individuo, y  $\Lambda_{x_{k_1}}$  es una matriz diagonal que contiene sus valores propios asociados, entonces

$$\hat{\eta}_{(n+1)\beta} = \hat{\beta}_0 + \mathbf{x}_{(k_1)}^t \hat{\beta}_{(k_1)} + \mathbf{x}_{(n-1-k_1)}^t \hat{\beta}_{(n-1-k_1)}$$

donde  $\mathbf{x}_{(n-1-k_1)}$  contiene las coordenadas menos correlacionadas con el nuevo individuo. Entonces, la predicción de un nuevo individuo en dimensión reducida se puede escribir como

$$\hat{\eta}_{(n+1)\beta}(k_1) = \hat{\beta}_0 + \mathbf{x}_{(k_1)}^t \hat{\beta}_{(k_1)} \quad (2.35)$$

Observe que si  $\mathbf{x}_{(n-1-k_1)}$  es muy grande, entonces (2.35) no funciona bien y la observación  $\mathbf{v}_{n+1}$  podría ser un dato atípico en el modelo de media.

Por ejemplo usando la función de enlace logit se encuentra que la predicción en términos de la variable original sería

$$\hat{y}_{(n+1)} = \frac{1}{1 + \exp[-(\hat{\beta}_0 + \mathbf{x}_{(k_1)}^t \hat{\beta}_{(k_1)})]}$$

Por último, un intervalo de confianza aproximado del  $(1 - q)100\%$  para la predicción de un nuevo individuo,  $n + 1$ , utilizando el modelo de media se puede obtener por

$$g^{-1} \left( \hat{\eta}_{(n+1)\beta}(k_1) \mp \Phi_{(1-q/2)}^{-1} ee(\hat{\eta}_{(n+1)\beta}(k_1)) \right)$$

donde  $ee(\hat{\eta}_{(n+1)\beta}(k_1)) = \sqrt{\mathbf{x}_{(k_1)}^t \widehat{\text{Cov}}(\hat{\beta}) \mathbf{x}_{(k_1)}}$ , con  $\widehat{\text{Cov}}(\hat{\beta})$  obtenida de la inversa de la matriz de información de Fisher evaluada en las estimaciones máximo verosímiles. Observe que el anterior intervalo es válido para funciones enlace estrictamente crecientes.

De otro lado, para la dispersión variable asociada con el  $n + 1$  individuo, siguiendo un procedimiento similar y asumiendo que las variables adicionales  $\mathbf{u}_{n+1} = (u_{(n+1)1}, \dots, u_{(n+1)p_2})^t$  fueron observadas, entonces

$$\mathbf{z}_{n+1} = \frac{1}{2} \Lambda_z^{-1} \mathbf{Z}^t (\mathbf{c} - \delta) \quad (2.36)$$

donde  $\mathbf{c} = (c_{11}, \dots, c_{nn})^t$  y  $\boldsymbol{\delta} = (\delta_{(n+1)1}^2, \dots, \delta_{(n+1)n}^2)^t$ , con  $c_{ii} = \mathbf{z}_i \mathbf{z}_i^t$  y  $\delta_{(n+1)i} = \delta(\mathbf{u}_{n+1}, \mathbf{u}_i)$ ,  $i = 1, \dots, n$ .

En este caso, la predicción asociada a la dispersión variable es

$$\hat{\eta}_{(n+1)\alpha} = \hat{\alpha}_0 + \mathbf{z}_{(k_2)}^t \hat{\boldsymbol{\alpha}}_{(k_2)} + \mathbf{x}_{(n-1-k_2)}^t \hat{\boldsymbol{\alpha}}_{(n-1-k_2)}$$

donde  $\mathbf{x}_{(n-1-k_2)}$  contiene la coordenadas menos correlacionadas con el nuevo individuo. Entonces, la predicción en dimensión reducida asociada a la dispersión variable es

$$\hat{\eta}_{(n+1)\alpha}(k_2) = \hat{\alpha}_0 + \mathbf{z}_{(k_2)}^t \hat{\boldsymbol{\alpha}}_{(k_2)} \quad (2.37)$$

Al igual que para la media, (2.35), observe que si  $\mathbf{x}_{(n-1-k_2)}$  es muy grande entonces (2.37) no funciona bien y la observación  $\mathbf{u}_{n+1}$  podría ser un dato atípico para el modelo de dispersión variable.

Por ejemplo, utilizando la función de enlace log se encuentra que la predicción para el termino de precisión (dispersión variable) sería

$$\hat{\phi}_{(n+1)} = \exp[\hat{\alpha}_0 + \mathbf{z}_{(k_2)}^t \hat{\boldsymbol{\alpha}}_{(k_2)}]$$

En este caso, un intervalo de confianza aproximado del  $(1 - q)100\%$  para la predicción de un nuevo individuo usando el modelo de dispersión variable se puede calcular mediante

$$g_2^{-1} \left( \hat{\eta}_{(n+1)\alpha}(k_2) \mp \Phi_{(1-q/2)}^{-1} ee(\hat{\eta}_{(n+1)\alpha}(k_2)) \right)$$

donde  $ee(\hat{\eta}_{(n+1)\alpha}(k_2)) = \sqrt{\mathbf{z}_{(k_2)}^t \widehat{\text{Cov}}(\hat{\boldsymbol{\alpha}}) \mathbf{z}_{(k_2)}}$ , con  $\widehat{\text{Cov}}(\hat{\boldsymbol{\alpha}})$  obtenida de la inversa de la matriz de información de Fisher, evaluada en las estimaciones máximo verosímiles.

### 2.3.5 Tratamiento de datos faltantes bajo la aproximación basada en distancias

Las matrices  $\mathbf{D}_v$  y  $\mathbf{D}_u$  son definidas positivas, y las matrices  $\mathbf{B}_v$  y  $\mathbf{B}_u$  son semi-definidas positivas; sin embargo, la presencia de valores faltantes pueden causar que  $\mathbf{B}_u$  y  $\mathbf{B}_v$  pierdan esta propiedad. En este caso no se puede descomponer  $\mathbf{B}_v = \mathbf{X}\mathbf{X}^t$  y  $\mathbf{B}_u = \mathbf{Z}\mathbf{Z}^t$ , excluyendo la posibilidad de proyectar  $\boldsymbol{\eta}_1$  y  $\boldsymbol{\eta}_2$  en las columnas de  $\mathbf{X}$  y  $\mathbf{Z}$ , respectivamente.

Según Cuadras et al. (1996), para superar el anterior problema de trabajar con matrices de distancia no euclidiana  $\mathbf{D}_v$  y  $\mathbf{D}_u$ ; estas matrices se transforman en  $\tilde{\mathbf{D}}_v = (\tilde{d}_{ii'})$  y  $\tilde{\mathbf{D}}_u = (\tilde{\delta}_{ii'})$ , donde

$$\tilde{d}_{ii'}^2 = \begin{cases} d_{ii'}^2 + 2\kappa_v & \text{si } i \neq i' \\ 0 & \text{si } i = i' \end{cases} \quad \tilde{\delta}_{ii'}^2 = \begin{cases} \delta_{ii'}^2 + 2\kappa_u & \text{si } i \neq i' \\ 0 & \text{si } i = i' \end{cases} \quad (2.38)$$

donde  $\kappa_v$  y  $\kappa_u$  son constantes apropiadas. Esta transformación es llamada solución de Lingoes (Cuadras et al. 1996, Mardia et al. 2002). Ahora, las matrices  $\mathbf{B}_v$  y  $\mathbf{B}_u$  son transformadas a

$$\widetilde{\mathbf{B}}_v = \mathbf{B}_v + \kappa_v \mathbf{H} \qquad \widetilde{\mathbf{B}}_u = \mathbf{B}_u + \kappa_u \mathbf{H}$$

Si  $\mathbf{x}_i$  y  $\mathbf{z}_j$  son vectores propios de  $\mathbf{B}_v$  y  $\mathbf{B}_u$ , respectivamente, con los correspondientes valores propios  $\lambda_{x_i}$  y  $\lambda_{z_j}$ , como  $\mathbf{1}^t \mathbf{x}_i = 0$  y  $\mathbf{1}^t \mathbf{z}_j = 0$ , respectivamente, se encuentra que

$$\widetilde{\mathbf{B}}_v \mathbf{x}_i = (\lambda_{x_i} + \kappa_v) \mathbf{x}_i \qquad \widetilde{\mathbf{B}}_u \mathbf{z}_j = (\lambda_{z_j} + \kappa_u) \mathbf{z}_j$$

y también  $\mathbf{x}_i$  y  $\mathbf{z}_j$  son los vectores propios de  $\widetilde{\mathbf{B}}_v$  y  $\widetilde{\mathbf{B}}_u$ , respectivamente, con valores propios  $\lambda_{x_i} + \kappa_v$  y  $\lambda_{z_j} + \kappa_u$ .

Por consiguiente, si  $\kappa_v \geq |\lambda_{x_{k_1}}|$  y  $\kappa_u \geq |\lambda_{z_{k_2}}|$  se eligen, donde  $\lambda_{x_{k_1}}$  y  $\lambda_{z_{k_2}}$  son los valores propios más pequeños (posiblemente negativos) de  $\mathbf{B}_v$  y  $\mathbf{B}_u$  respectivamente, entonces  $\widetilde{\mathbf{B}}_v$  y  $\widetilde{\mathbf{B}}_u$  son semidefinidas positivas, y así,  $\widetilde{\mathbf{D}}_v$  y  $\widetilde{\mathbf{D}}_u$  se convierten en matrices de distancias euclidianas. Además, se puede utilizar la transformada empleando el modelo DBBR, la cual está dada por

$$\boldsymbol{\eta}_1 = \beta_0 \mathbf{1} + \widetilde{\mathbf{X}} \boldsymbol{\beta} \qquad \boldsymbol{\eta}_2 = \alpha_0 \mathbf{1} + \widetilde{\mathbf{Z}} \boldsymbol{\alpha}$$

donde  $\widetilde{\mathbf{B}}_v = \widetilde{\mathbf{X}} \widetilde{\mathbf{X}}^t$  y  $\widetilde{\mathbf{B}}_u = \widetilde{\mathbf{Z}} \widetilde{\mathbf{Z}}^t$ .

Por tanto, siguiendo a Cuadras et al. (1996), las predicciones del modelo de media y el modelo de dispersión variable con respecto a la distancia (2.38) están dadas respectivamente por

$$\begin{aligned} \widehat{\eta}_{(n+1)\beta} &= \widehat{\beta}_0 + \frac{1}{2} (\mathbf{b} - \mathbf{d})^t \widetilde{\mathbf{X}} (\boldsymbol{\Lambda}_x + \kappa_v \mathbf{I})^{-1} \widehat{\boldsymbol{\beta}} \\ \widehat{\eta}_{(n+1)\alpha} &= \widehat{\alpha}_0 + \frac{1}{2} (\mathbf{c} - \boldsymbol{\delta})^t \widetilde{\mathbf{Z}} (\boldsymbol{\Lambda}_z + \kappa_u \mathbf{I})^{-1} \widehat{\boldsymbol{\alpha}} \end{aligned}$$

## 2.4 Relación con el modelo de regresión beta clásico

Los modelos presentados en (2.6) dependen de las distancias elegidas,  $d_{ij}$  y  $\delta_{ij}$ . Cuando las variables de predicción son continuas y se utiliza la distancia euclidianas, el modelo DBBR con dispersión variable es compatible con el modelo de regresión beta con dispersión variable clásico. Esta equivalencia también se mantiene para variables explicativas cualitativas (o una mezcla de variables explicativas continuas, categóricas y binarias) cuando se utiliza el método DB sobre los coeficientes de coincidencias,  $m_{ii'}$ .

### 2.4.1 Variables continuas

Si todas las variables explicativas en (2.5) ( $\mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_{p_1})$  y  $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_{p_2})$ ) son continuas, la distancia euclidiana para cada caso está dada por

$$d_{ii'}^2 = \mathbf{v}_i^t \mathbf{v}_i + \mathbf{v}_{i'}^t \mathbf{v}_{i'} - 2\mathbf{v}_i^t \mathbf{v}_{i'} \quad \delta_{ii'}^2 = \mathbf{u}_i^t \mathbf{u}_i + \mathbf{u}_{i'}^t \mathbf{u}_{i'} - 2\mathbf{u}_i^t \mathbf{u}_{i'} \quad (2.39)$$

De esta manera, las matrices de distancias  $\mathbf{D}_v = (d_{ii'})$  y  $\mathbf{D}_u = (\delta_{ii'})$  se obtienen, y además,

$$\begin{aligned} \mathbf{Q}_v &= \text{diag}(\mathbf{V}\mathbf{V}^t)\mathbf{1}^t + \mathbf{1}(\text{diag}(\mathbf{V}\mathbf{V}^t))^t - 2\mathbf{V}\mathbf{V}^t \\ \mathbf{Q}_u &= \text{diag}(\mathbf{U}\mathbf{U}^t)\mathbf{1}^t + \mathbf{1}(\text{diag}(\mathbf{U}\mathbf{U}^t))^t - 2\mathbf{U}\mathbf{U}^t \end{aligned}$$

donde  $\mathbf{Q}_v = (d_{ii'}^2)$  and  $\mathbf{Q}_u = (\delta_{ii'}^2)$  son matrices de distancias al cuadrado,  $\text{diag}(\mathbf{V}\mathbf{V}^t)$  y  $\text{diag}(\mathbf{U}\mathbf{U}^t)$  son vectores que contienen los términos diagonales de las matrices  $\mathbf{V}\mathbf{V}^t$  y  $\mathbf{U}\mathbf{U}^t$ , respectivamente.

Observe que  $\mathbf{A}_v = -\frac{1}{2}\mathbf{Q}_v$  y  $\mathbf{A}_u = -\frac{1}{2}\mathbf{Q}_u$ , por lo tanto,

$$\begin{aligned} \mathbf{B}_v &= -\frac{1}{2}\mathbf{H}\mathbf{Q}_v\mathbf{H} = \mathbf{H}\mathbf{V}\mathbf{V}^t\mathbf{H} & \mathbf{B}_u &= -\frac{1}{2}\mathbf{H}\mathbf{Q}_u\mathbf{H} = \mathbf{H}\mathbf{U}\mathbf{U}^t\mathbf{H} \\ &= \mathbf{X}\mathbf{X}^t & &= \mathbf{Z}\mathbf{Z}^t \end{aligned}$$

ya que  $\mathbf{H}[\text{diag}(\mathbf{V}\mathbf{V}^t)\mathbf{1}^t]\mathbf{H} = \mathbf{0}$ ,  $\mathbf{H}[\mathbf{1}(\text{diag}(\mathbf{V}\mathbf{V}^t))^t]\mathbf{H} = \mathbf{0}$ ,  $\mathbf{H}[\text{diag}(\mathbf{U}\mathbf{U}^t)\mathbf{1}^t]\mathbf{H} = \mathbf{0}$  y  $\mathbf{H}[\mathbf{1}(\text{diag}(\mathbf{U}\mathbf{U}^t))^t]\mathbf{H} = \mathbf{0}$ , y donde  $\mathbf{B}_v$  y  $\mathbf{B}_u$  fueron definidas en la subsección 2.2.1. Entonces, el modelo DBBR con dispersión variable introducido en (2.6) en cada uno de los modelos propuestos en  $p_1$  y  $p_2$  dimensiones, respectivamente, produce las mismas predicciones que el modelo dado en (2.5).

Sin embargo, no es necesario considerar una distancia euclidiana  $p_1$ -dimensional para el modelo de media y una distancia euclidiana  $p_2$ -dimensional para el modelo de dispersión variable, como las dadas en las ecuaciones (2.39), porque se puede utilizar alguna otra distancia (por ejemplo, la distancia valor absoluto dada en la ecuación (2.42)). Sean  $E_{l_1}^{(1)}$  ( $k_1 \leq l_1 \leq n-1$ ) y  $E_{l_2}^{(2)}$  ( $k_2 \leq l_2 \leq n-1$ ) los espacios generados por las columnas de  $\mathbf{X}$  y  $\mathbf{Z}$ , respectivamente; donde  $\mathbf{X}$  y  $\mathbf{Z}$  son soluciones métricas escaladas obtenidos a partir de una distancia aplicada a los mismos datos. Entonces, tomando  $k_1 > p_1$  y  $k_2 > p_2$  (o  $k_1 > p_1$  y  $k_2 = p_2$ , o  $k_1 = p_1$  y  $k_2 > p_2$ ), es decir, las columnas más adecuadas de  $\mathbf{X}$  y  $\mathbf{Z}$ , respectivamente, el modelo DBBR supera al modelo de regresión beta clásico cuando  $(\boldsymbol{\eta}_1 - \hat{\beta}_0\mathbf{1}) \in E_{l_1}$  y  $(\boldsymbol{\eta}_2 - \hat{\alpha}_0\mathbf{1}) \in E_{l_2}$ . Observe que esto siempre es cierto para  $k_1 + k_2 = n-1$  con  $k_1 > p_1$  y  $k_2 > p_2$ .

### 2.4.2 Variables Cualitativas

Suponga que todas las variables explicativas  $\mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_{p_1})$  y  $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_{p_2})$  son cualitativas en (2.5), donde ahora cada  $\mathbf{V}_j$  y  $\mathbf{U}_h$  son variables en los  $q_j^v$  y  $q_h^u$  estados, respectivamente,  $j = 1, \dots, p_1$  y  $h = 1, \dots, p_2$ . En cada caso, una medida de similaridad entre los individuos  $i$  e  $i'$  es el número de coincidencias  $m_{ii'}^v$  y  $m_{ii'}^u$  para las variables cualitativas involucradas en los modelos de media y de dispersión variable, respectivamente. Observe que  $0 \leq m_{ii'}^v \leq p_1$  y  $0 \leq m_{ii'}^u \leq p_2$ , y cuando las variables explicativas son binarias,  $m_{ii'}^v/p_1$  y  $m_{ii'}^u/p_2$  son los coeficientes de coincidencia correspondientes.

En cada modelo, las distancias están definidas como:

$$d_{ii'}^2 = m_{ii}^v + m_{i'i'}^v - 2m_{ii'}^v = 2(p_1 - m_{ii'}^v) \quad \delta_{ii'}^2 = m_{ii}^u + m_{i'i'}^u - 2m_{ii'}^u = 2(p_2 - m_{ii'}^u)$$

De una manera similar, como se demostró para las variables continuas, haciendo  $\mathbf{A}_v = -\frac{1}{2}[\mathbf{M}_r^v + (\mathbf{M}_r^v)^t - 2\mathbf{M}^v]$  y  $\mathbf{A}_u = -\frac{1}{2}[\mathbf{M}_r^u + (\mathbf{M}_r^u)^t - 2\mathbf{M}^u]$ . Por lo tanto,

$$\begin{aligned} \mathbf{B}_v &= \mathbf{H}\mathbf{A}_v\mathbf{H} = \mathbf{H}\mathbf{M}^v\mathbf{H} & \mathbf{B}_u &= \mathbf{H}\mathbf{A}_u\mathbf{H} = \mathbf{H}\mathbf{M}^u\mathbf{H} \\ &= \mathbf{H}\mathbf{V}\mathbf{V}^t\mathbf{H} = \mathbf{X}\mathbf{X}^t & &= \mathbf{H}\mathbf{U}\mathbf{U}^t\mathbf{H} = \mathbf{Z}\mathbf{Z}^t \end{aligned}$$

donde todas las filas de las matrices  $\mathbf{M}_r^v$  y  $\mathbf{M}_r^u$  son iguales,  $\mathbf{M}^v = (m_{ii'}^v)$ ,  $\mathbf{M}^u = (m_{ii'}^u)$ , y  $\mathbf{H}\mathbf{M}_r^v = (\mathbf{M}_r^v)^t\mathbf{H} = \mathbf{0}$  y  $\mathbf{H}\mathbf{M}_r^u = (\mathbf{M}_r^u)^t\mathbf{H} = \mathbf{0}$ . Por tanto, no hay ninguna ventaja sobre el modelo de regresión beta clásico, excepto que el problema de multicolinealidad puede ser resuelto automáticamente mediante el uso de distancias.

### 2.4.3 Variables mixtas

Suponga ahora que en (2.5) las matrices  $\mathbf{V} = (\mathbf{V}^{(1)} \ \mathbf{V}^{(2)})$  y  $\mathbf{U} = (\mathbf{U}^{(1)} \ \mathbf{U}^{(2)})$ , donde  $\mathbf{V}^{(1)}$  y  $\mathbf{U}^{(1)}$  son matrices de variables continuas, y  $\mathbf{V}^{(2)}$  y  $\mathbf{U}^{(2)}$  son matrices de variables cualitativas de  $\mathbf{V}$  y  $\mathbf{U}$ , respectivamente. En este caso, de acuerdo a Cuadras & Arenas (1990) y Cuadras et al. (1996), es apropiado usar las similitudes  $m_{ii'}^v$  y  $m_{ii'}^u$  dadas por Gower (1971) entre el individuo  $i$  e  $i'$  para cada modelo, respectivamente, las cuales están dadas por

$$\begin{aligned} m_{ii'}^v &= \frac{\sum_{j=1}^{p_v^c} \left( \frac{1 - |v_{ij} - v_{i'j}|}{G_j^v} \right) + c_{1ii'}^v + v_{ii'}^v}{p_v^c + (p_v^b - c_{4ii'}^v) + p_v^q} \\ m_{ii'}^u &= \frac{\sum_{j=1}^{p_u^c} \left( \frac{1 - |u_{ij} - u_{i'j}|}{G_j^u} \right) + c_{1ii'}^u + v_{ii'}^u}{p_u^c + (p_u^b - c_{4ii'}^u) + p_u^q} \end{aligned} \quad (2.40)$$



donde, para el modelo de media,  $p_v^c$  es el número de variables continuas,  $c_{1ii'}^v$  y  $c_{4ii'}^v$  son el número de coincidencias positivas y negativas, respectivamente para las  $p_v^b$  variables binarias, y  $v_{ii'}^v$  es el número de concordancias para las  $p_v^q$  variables multiestado cualitativas.  $G_j^v$  es el rango (o distancia) de la  $j$ -ésima variable continua. En el caso del modelo de dispersión variable, los parámetros se definen del mismo modo.

Por lo tanto, las distancias al cuadrado entre los individuos  $i$  e  $i'$  son  $d_{ii'}^2 = 1 - m_{ii'}^v$  y  $\delta_{ii'}^2 = 1 - m_{ii'}^u$  para los modelos de media y dispersión variable, respectivamente. Por consiguiente,  $\mathbf{D}_v = (d_{ii'})$  y  $\mathbf{D}_u = (\delta_{ii'})$  son matrices de distancias euclidianas en el conjunto de  $n$  individuos. Entonces, como se demostró para las variables continuas y cualitativas, el modelo DBBR con dispersión variable produce las mismas predicciones que el modelo de regresión clásico con dispersión variable clásico. Sin embargo, al igual que en la propuesta con variables cuantitativas y cualitativas, el modelo DBBR para datos mixtos podría superar al modelo de regresión beta clásico después de escalar los datos, siempre que el número de coordenadas principales satisfaga  $k_1 > p_1$  y  $k_2 > p_2$  en los modelos de media y de dispersión variable, respectivamente.

#### 2.4.4 Modelo de regresión beta no lineal basado en distancias

El enfoque DB se usa también para llevar a cabo regresión no lineal. Sean  $\mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_{p_1})$  y  $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_{p_2})$  variables explicativas continuas, la regresión beta no lineal DB está dada por

$$\boldsymbol{\eta}_1 = f_1(\mathbf{V}_1, \dots, \mathbf{V}_{p_1}, \boldsymbol{\mu}) \quad \boldsymbol{\eta}_2 = f_2(\mathbf{U}_1, \dots, \mathbf{U}_{p_2}, \boldsymbol{\phi}) \quad (2.41)$$

donde  $f_1$  y  $f_2$  son dos funciones no lineales (posiblemente desconocidas) dependientes de vectores paramétricos  $\boldsymbol{\mu}$  y  $\boldsymbol{\phi}$ , respectivamente.

El modelo DBBR usando distancias euclidianas es equivalente a interpretar (2.41) en términos de una regresión lineal beta tal como se presentó en (2.6). Si se considera la distancia valor absoluto,

$$d_{ii'}^2 = \sum_{j=1}^{p_1} |v_{ij} - v_{i'j}| \quad \delta_{ii'}^2 = \sum_{j=1}^{p_2} |u_{ij} - u_{i'j}|, \quad (2.42)$$

entonces el modelo DBBR es no lineal. En el modelo de regresión clásico, Cuadras et al. (1996) probaron que si se considera una sola variable explicativa con valores equidistantes, el enfoque DB es equivalente a una regresión ordinaria de polinomios ortogonales; específicamente, éstos son los polinomios de Chebyshev. El mismo resultado se mantiene para el DBBR cuando  $p_1 = 1$  en el modelo de media y  $p_2 = 1$  en el modelo de dispersión variable.

En el momento, no existen resultados teóricos cuando hay dos o más variables explicativas (Cuadras et al. 1996). Sin embargo, el modelo DB con la distancia (2.42) en regresión beta no lineal podría ser útil en algunas aplicaciones reales.

## 2.5 Aplicaciones

En esta sección se presentan dos aplicaciones que motivaron la propuesta de la metodología presentada en este capítulo: la primera es un estudio presentado en Prater (1956) sobre la proporción de petróleo crudo convertido a gasolina, el cual fue también estudiado por Ferrari & Cribari-Neto (2004), Ospina et al. (2006), Espinheira et al. (2008a) y Ferrari et al. (2011). La segunda aplicación es un estudio sobre el rendimiento en fondos de inversión que no ha sido estudiado en ningún otro trabajo. En los dos estudios es de gran interés la predicción de la variable respuesta involucrada, así que el modelo DBBR es muy útil.

### 2.5.1 Proporción de petróleo crudo convertido a gasolina

A continuación se considera el conjunto de datos recolectados por Prater (1956); los datos contienen 32 observaciones. En este estudio se desea modelar la proporción de crudo convertido en gasolina tras un proceso de destilación, y relacionarla, con las covariables: temperatura en la cual el 10 % se convierte en vapor y la temperatura ( $^{\circ}F$ ) en la cual toda la gasolina se evapora. Los datos están ordenados en forma ascendente de acuerdo con la covariable que mide la temperatura a la cual el 10 % del petróleo pasa a ser vapor. Esta variable asume diez diferentes valores y se utilizan para definir diez lotes de petróleo. La relación final está determinada por nueve variables dummy para los primeros nueve lotes de petróleo y la covariable temperatura ( $^{\circ}F$ ) en la cual la gasolina se evapora.

Este conjunto de datos fue analizada por Atkinson (1985), quien utilizó el modelo de regresión lineal y encontró que la distribución de los errores no era completamente simétrica, lo cual genera residuos demasiado grandes. Luego, Ferrari & Cribari-Neto (2004), Ospina et al. (2006) y Simas et al. (2010) utilizaron estos datos como una ilustración del modelo de regresión beta en los esquemas de corrección del sesgo y en el modelamiento de la dispersión variable. Debido a que en este trabajo no se realizan las correcciones de sesgo y a que se está en otro sistema de coordenadas al utilizar el método DB, no se pretende con esta aplicación hacer comparaciones con el análisis obtenido por

Ospina et al. (2006) y Simas et al. (2010), pero si con respecto al de Ferrari & Cribari-Neto (2004).

A partir del enfoque planteado en esta aplicación, se ajusta el modelo de DBBR con dispersión variable. En el modelo de medias se utiliza la distancia de Gower porque se consideran los diez lotes de petróleo y la temperatura, y en el modelo de dispersión variable se utiliza la distancia euclidiana clásica ya que solo se considera la variable temperatura para modelar la heterocedasticidad.

En primer lugar se construyen las distancias a partir de las variables explicativas para cada modelo y luego las matrices  $\mathbf{B}_v = \mathbf{H}\mathbf{A}_v\mathbf{H}$  y  $\mathbf{B}_u = \mathbf{H}\mathbf{A}_u\mathbf{H}$  con  $\mathbf{H} = (\mathbf{I}_n - \frac{1}{32}\mathbf{1}\mathbf{1}^t)$ . Para elegir las variables que se deben incluir en los modelos, se utilizaron las diferentes funciones de enlace en el modelo de medias (logit, cloglog, loglog y logit) y en el modelo de dispersión variable (identity, log y sqrt) que se encuentran implementadas en el paquete *betareg* del R (R Development Core Team 2013). El anterior proceso se realizó utilizando los  $w_\beta$ 's y  $w_\alpha$ 's más altos en cada caso al realizar el modelo de DBBR, pero se puede utilizar cualquier otro procedimiento de los presentado en la subsección 2.3.2.

En base a los resultados encontrados, se seleccionan 10 coordenadas principales en el modelo de medias ( $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6, \mathbf{X}_9, \mathbf{X}_{10}, \mathbf{X}_{12}$  y  $\mathbf{X}_{16}$ ) y una en el modelo de dispersión variable ( $\mathbf{Z}_1$ ). Esto se hace con el objetivo de mantener el mismo número de variables explicativas que en los modelos propuestos por Ferrari & Cribari-Neto (2004) y Cribari-Neto & Zeileis (2010); sin embargo, en los modelos con DB no es necesario hacerlo porque incluso se pueden incluir más componentes en cada modelo, las cuales pueden tener aún más relación con la variable respuesta que las seleccionadas inicialmente al combinar entre modelos de medias y de dispersión.

En el modelo de dispersión variable se incluye inicialmente la coordenada principal  $\mathbf{Z}_1$  con el propósito de elegir la función de enlace apropiada en la DBBR, y en el modelo de regresión beta clásico, se elige la variable temperatura. Además se encuentra que las variables temperatura y  $\mathbf{Z}_1$  tienen correlación 1, es decir que la componente  $\mathbf{Z}_1$  es directamente proporcional a la temperatura en el conjunto de datos originales. Por lo tanto, la función de enlace dejando el modelo de medias constante la misma para los dos modelos de dispersión variable. De este modo, los valores obtenidos al realizar los diferentes modelos son: AIC=-54.94 y loglik=30.47 para el enlace identidad, AIC=-55.13 y loglik=30.56 para el enlace log, y AIC=-55.22 y loglik=30.61 para el enlace raíz cuadrada. Estos resultados son prácticamente idénticos, lo que significa que se puede elegir cualquier función de enlace para el modelo de dispersión variable en los dos casos. Por fines prácticos, se eligió la función de enlace log ya que fue optimizada en menos pasos que los otras dos.

En la Tabla 2.1, se muestran los resultados del coeficiente de pseudo-

correlación ( $R_{10}^2 = 98.5\%$  en el modelo clásico y  $R_{10}^2 = 98.0\%$  en el modelo DB), el AIC (-166.58 en el caso clásico y -168.54 en DB) y loglik (96.29 en el caso clásico y 97.27 en DB). A partir de los resultados presentado en esta tabla, se eligió la función de enlace loglog para el modelo de media en ambos casos (clásico y DB). En general, al comparar las estadísticas de los dos modelos son muy parecidas, sin una notable ganancia en cualquiera de los dos modelos.

TABLA 2.1: Funciones de enlace utilizando los modelos de regresión beta clásico y DBBR con dispersión variable para la proporción de petróleo crudo convertido a gasolina.

Enlace	Modelo clásico			DBBR		
	$R^2$	AIC	loglik	$R^2$	AIC	loglik
logit	0.952	-147.95	86.98	0.955	-149.39	87.69
probit	0.973	-154.51	90.25	0.971	-157.68	91.84
cloglog	0.934	-142.12	84.06	0.941	-141.84	83.92
loglog	0.985	-166.58	96.29	0.980	-168.54	97.27

En este caso para mejorar el modelo, se puede incluir en el modelo de dispersión variable la componente  $\mathbf{Z}_2$  con la finalidad de modelar mejor la heterocedasticidad. Si la finalidad es predecir, el anterior cambio ocasiona varias mejorías en el modelo como se puede verificar con las estadísticas: pseudo- $R_{10}^2 = 0.974$ , AIC=-209.03 y loglik=118.5, con lo cual se mejora en comparación al mismo caso de DBBR, planteado anteriormente. Esto se hace de nuevo teniendo en cuenta otra de las ventajas del método DB.

Después de seleccionar los modelos, la hipótesis  $H_0 : \alpha_1 = \alpha_2 = 0$  (dispersión constante o precisión constante) se juzga utilizando la función de enlace log en el modelo de dispersión variable. Para ello se compara el modelo DBBR con dispersión variable con el modelo de regresión beta con dispersión constante incluyendo las mismas componentes en el modelo de media. Entonces, la hipótesis de dispersión constante se rechaza porque  $\chi_c^2 = 49.67 > \chi_{(2,0.05)}^2$  (valor  $p = 1.635e - 11$ ); por lo tanto, el modelo de dispersión variable es apropiado.

Usando la función de enlace loglog en el modelo de medias, la hipótesis  $H_0 : \beta_1 = \beta_2 = \dots = \beta_{10} = 0$  es rechazada a un nivel de significancia del 5% porque la prueba de razón de verosimilitud es  $\chi_c^2 = 173.833 > \chi_{(10,0.05)}^2$  (valor  $p < 2.2e - 16$ ). Por lo tanto, existe al menos un término diferente de cero y existe una relación entre las coordenadas principales y la proporción del petróleo crudo convertido a gasolina. En la Tabla 2.2, se muestran los coeficientes con su respectivas pruebas de significancia tanto para el modelo de media como para el modelo de dispersión variable.

En la Tabla 2.2, se puede observar que todas las coordenadas principales en los modelos de medias y de dispersión variable son significantes a un nivel del 5%, por consiguiente el modelo de medias obtenido en términos de coordenadas

TABLA 2.2: Estimación de parámetros para el modelo DBBR con dispersión variable que relaciona la proporción de petróleo crudo convertido en gasolina con las coordenadas principales.

Efecto	Estimación	Desv. Error	$w_\beta$	$Pr(>  w_\beta )$
Coeficientes (modelo de media con enlace loglog):				
(Intercepto)	-0.530	0.006	-85.23	0.000
$X_1$	-0.255	0.019	-13.47	0.000
$X_2$	-0.563	0.036	-15.44	0.000
$X_3$	-0.922	0.039	-23.90	0.000
$X_4$	-1.662	0.039	-42.74	0.000
$X_5$	-0.606	0.044	-13.86	0.000
$X_6$	0.654	0.062	10.48	0.000
$X_9$	-0.989	0.084	-11.74	0.000
$X_{10}$	-1.545	0.079	-19.55	0.000
$X_{12}$	1.025	0.035	29.39	0.000
$X_{16}$	-2.158	0.230	-9.38	0.000
Phi coeficientes (modelo de precisión con enlace log):				
(Intercepto)	6.516	0.256	25.474	0.000
$Z_1$	-6.567	1.401	-4.688	0.000
$Z_2$	-36.168	1.698	-21.301	0.000

principales se puede escribir como

$$-\log(-\log(\hat{\mu})) = -0.53\mathbf{1}_{32} - 0.26\mathbf{X}_1 - 0.56\mathbf{X}_2 - 0.92\mathbf{X}_3 - 1.66\mathbf{X}_4 - 0.61\mathbf{X}_5 \\ + 0.65\mathbf{X}_6 - 0.99\mathbf{X}_9 - 1.55\mathbf{X}_{10} + 1.03\mathbf{X}_{12} - 2.16\mathbf{X}_{16},$$

y el modelo de dispersión variable como

$$\log(\hat{\phi}) = 6.52\mathbf{1}_{32} - 6.57\mathbf{Z}_1 - 36.17\mathbf{Z}_2$$

Una vez ajustado el modelo es muy importante analizar la bondad de ajuste del modelo estimado. Como se mencionó anteriormente el modelo cuenta con un pseudo- $R_{10}^2 = 97.4\%$ . En la Figura 2.1 se observa un ajuste adecuado del presente modelo, ya que no hay observaciones influyentes según el gráfico de los Cook's y tampoco hay observaciones atípicas que se salgan del intervalo -3 y 3 de acuerdo a los residuos de Pearson y de deviance, aunque la observación 19 tiene un valor un poco alto. Adicionalmente, en el gráfico de datos observados versus predichos se ve que la respuesta original se parece mucho a los valores predichos.

## 2.5.2 Rendimiento en fondos de inversión

El conjunto de datos contiene información de 44 fondos de inversión que fueron observados en el 2008 y forman parte del Morningstar Funds500 (Morningstar

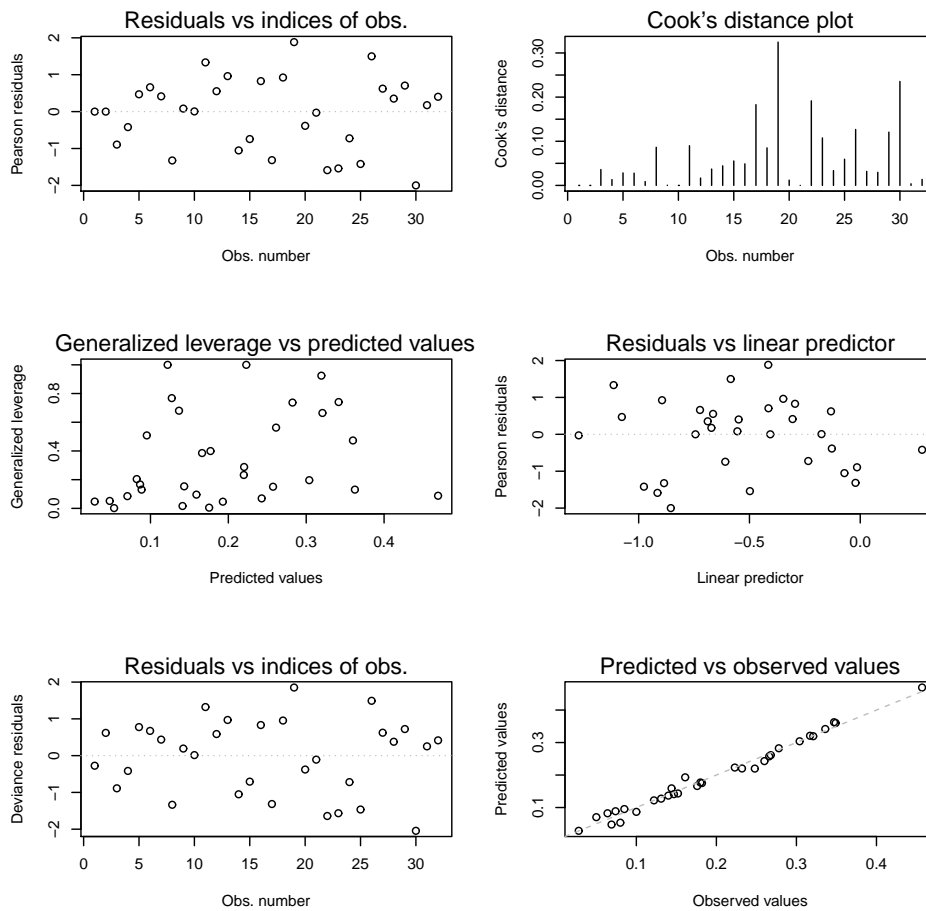


FIGURA 2.1: Diagnóstico del modelo con dispersión variable para la proporción de petróleo crudo convertido a gasolina.

Inc. et al. 2008, Anderson et al. 2011). El conjunto de datos contiene las siguientes cinco variables: i) el tipo de fondo (fund type, FT) el cual está conformado por la renta variable nacional (domestic equity, DE), la renta variable internacional (international equity, IE) y la renta fija (Fixed Income, FI); ii) el patrimonio neto (net asset value, NAV, \$), que es el precio de cierre por acción a diciembre 31 de 2007; iii) rendimiento promedio durante 5 años (5-year average return, FYAR, %), que es la rentabilidad media anual del fondo en los últimos 5 años; iv) índice de gasto (Expense Ratio, ER, %), que es el porcentaje de los activos deducidos cada año fiscal para los gastos del fondo; y v) clasificación de Morningstar (Morningstar rank, MR) que es la calificación ajustada con estrellas del riesgo para cada fondo, estas calificaciones van de bajo (1 estrella) a alto (5 estrellas). El interés se centra en la predicción del rendimiento promedio durante los 5 años relacionadas con las demás variables descritas.

Es imposible seleccionar valores basados solamente en sus patrones de desempeños pasados; por ello mirando los fondos con mejores resultados en cualquier período de tiempo, se pueden encontrar pistas importantes sobre ciertos capitales y tendencias del mercado. Un buen modelo para predecir el rendimiento promedio durante 5 años puede ser un primer paso esencial para determinar si un fondo es apropiado o no. El analista resume todos los puntos sobresalientes de un fondo mediante la evaluación de su perfil de riesgo histórica en relación a su grupo paritario, discutiendo la estrategia al gerente y tomando nota de los cambios recientes en el portafolio.

Dado el enfoque de esta aplicación y la naturaleza de las variables involucradas, se emplean los modelos DBBR con dispersión variable. En el modelo de media se utiliza la distancia de Gower porque se emplean las variables FT, NAV, ER y MR. En el modelo de dispersión variable se utilizó también la distancia de Gower, pero en este caso, sólo se consideran las variables FT y MR para el modelamiento de la heterocedasticidad.

Una vez definidas las variables explicativas que intervienen en los modelos de medias y de dispersión variable para construir el modelo DBBR; se construyen las distancias a partir de las variables explicativas para cada modelo y las matrices  $\mathbf{B}_v = \mathbf{H}\mathbf{A}_v\mathbf{H}$  y  $\mathbf{B}_u = \mathbf{H}\mathbf{A}_u\mathbf{H}$  con  $\mathbf{H} = (\mathbf{I}_{44} - \frac{1}{44}\mathbf{1}\mathbf{1}^t)$ . Para elegir las coordenadas principales a ser incluidas en ambos modelos, fueron utilizadas diferentes funciones de enlace en el modelo de media (logit, cloglog, loglog y logit) y en el modelo de dispersión variable (identity, log y sqrt). Estas funciones de enlace fueron implementadas en el paquete de R *betareg* (R Development Core Team 2013). El proceso anterior se llevó a cabo utilizando los mayores  $w_\beta$ 's y  $w_\alpha$ 's en cada modelo.

Se seleccionaron siete coordenadas principales en el modelo de medias ( $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6$  y  $\mathbf{X}_8$ ) y cuatro en el modelo de dispersión variable ( $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_4$  y  $\mathbf{Z}_5$ ). Al igual que en la primera aplicación, esto se hizo con el objetivo de mantener el mismo número de variables explicativas que en el modelo de regresión beta clásico con dispersión variable.

Una vez realizado el procedimiento anterior, con la finalidad de elegir la función de enlace para la parte del modelo de dispersión variable, un modelo DBBR con dispersión variable se ajustó utilizando diferentes funciones de enlace dejando constante el modelo de media. Para ello, las coordenadas principales  $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_4$  y  $\mathbf{Z}_5$  se incluyeron en la parte del modelo de dispersión variable. Los valores obtenidos en los diferentes modelos fueron: AIC=-106.98 y loglik=59.49 para la función de enlace identidad, AIC=-108.36 y loglik=60.18 para la función de enlace log, y AIC=-107.44 y loglik=59.72 para la función de enlace raíz cuadrada. Así, se escogió la función enlace log para el modelo de dispersión variable sin incluir ninguna covariable en la parte del modelo de media.

Por otra parte, con el objetivo de hacer una comparación con el modelo de regresión beta clásico con dispersión variable, las variables FT y MR se seleccionaron para la parte del modelo de dispersión variable. En este caso, también se eligió la función de enlace log porque el AIC=-144.77 fue el más pequeño y la loglik=80.39 fue la más alta.

Una vez seleccionada la función de enlace log en ambos casos (el modelo de regresión beta clásico y el modelo DBBR) para la parte del modelo de dispersión variable, se consideraron los modelos DBBR y regresión beta con dispersión variable utilizando la parte del modelo de medias. La Tabla 2.3 muestra los resultados que permiten seleccionar la función enlace apropiada para el modelo de media dado que la función enlace log fue seleccionada en el modelo de dispersión variable. Los resultados basados en: (a) el coeficiente de pseudo-correlación ( $R_7^2 = 83.4\%$  en el caso clásico y  $R_7^2 = 79.3\%$  en DB), (b) el AIC (-197.32 en el caso clásico y -208.95 en DB) y (c) la loglik (111.66 en el caso clásico y 117.47 en DB), muestran que se puede seleccionar la función de enlace loglog para el modelo de media en ambos casos (clásico y DB). En general, el modelo DBBR se comportó mejor que la regresión beta clásica de acuerdo a las estadísticas AIC y loglik, pero la regresión beta clásica es mejor si se utiliza el coeficiente de pseudo- $R^2$ .

TABLA 2.3: Funciones de enlace utilizando los modelos de regresión beta clásico y DBBR con dispersión variable para el retorno promedio de 5 años (%).

Enlace	Clásico			DB		
	$R^2$	AIC	loglik	$R^2$	AIC	loglik
logit	0.851	-195.82	110.91	0.820	-204.94	115.47
probit	0.846	-196.53	111.26	0.810	-207.35	116.67
cloglog	0.853	-195.73	110.86	0.824	-204.08	115.04
loglog	0.834	-197.32	111.66	0.793	-208.95	117.47

Después de seleccionar los dos modelos con sus respectivas funciones de enlace, fue contrastada la hipótesis  $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$  (dispersión constante) utilizando la función de enlace log en el modelo de dispersión variable. Para hacer esto, el modelo DBBR con dispersión variable se comparó con el modelo DBBR con dispersión constante incluyendo las mismas coordenadas en el modelo de media. Así, la hipótesis de dispersión constante se rechaza porque  $\chi_c^2 = 44.57 > \chi_{(4,0.05)}^2$  (valor p < 0.0001), por tanto, el modelo de dispersión variable es apropiado.

Ahora, utilizando la función de enlace loglog en el modelo de media, se rechaza la hipótesis  $H_0 : \beta_1 = \beta_2 = \dots = \beta_7 = 0$  a un nivel de 5% de significancia ya que la prueba de razón de verosimilitud fue  $\chi_c^2 = 80.57 > \chi_{(7,0.05)}^2$  (valor p < 0.0001), por lo tanto, hay al menos un término diferente de cero y existe relación significativa entre las coordenadas principales y los 5-años



de rentabilidad promedio. Entonces, empleando el modelo DBBR con dispersión variable que relaciona el rendimiento promedio de 5 años con coordenadas principales, se encuentra que todas las coordenadas en los modelos de media y de dispersión variable fueron significativas a un nivel de 5 %. Así, el modelo de medias se pueda escribir en términos de las coordenadas principales como

$$\begin{aligned} -\log(-\log(\widehat{\boldsymbol{\mu}})) = & -0.721_{44} + 0.62\mathbf{X}_1 - 0.55\mathbf{X}_2 - 0.51\mathbf{X}_3 - 0.34\mathbf{X}_4 \\ & - 0.27\mathbf{X}_5 - 0.88\mathbf{X}_6 - 0.24\mathbf{X}_8 \end{aligned}$$

y el modelo de dispersión variable como

$$\log(\widehat{\boldsymbol{\phi}}) = 5.941_{44} + 3.62\mathbf{Z}_1 + 7.16\mathbf{Z}_2 - 3.08\mathbf{Z}_4 - 4.81\mathbf{Z}_5$$

Las conclusiones sobre las coordenadas principales pueden no ser revelantes para el objetivo general debido a que el investigador podría estar interesado en información sobre las variables originales; sin embargo, éstas se presentan con la finalidad de destacar que estas coordenadas principales son significativas y deben ser consideradas en el proceso de predicción. No obstante, podemos decir que el tipo de fondo, NAVE, coeficientes de gastos y el rango Morningstar afecta a la rentabilidad promedio de 5 años, ya que a partir de estas variables explicativas se construyeron las coordenadas principales empleando la distancia de Gower y debido a que estas coordenadas fueron significativas al 5 %.

El modelo DBBR con dispersión variable ajustado tiene una pseudo  $R_7^2 = 0.793$ , lo cual indica que se tiene un buen modelo. La Figura 2.2 muestra un buen ajuste porque no hay observaciones influyentes de acuerdo a la gráfica de Cook's y no hay valores extremos que estén afuera del intervalo de -3 a 3 de acuerdo con los residuales de Pearson y de deviance. Aunque la distribución de los residuales no es conocida, el gráfico half-normal no muestra aparentemente problemas porque los residuos de deviance están dentro de los intervalos de confianza de 95 %. Adicionalmente, los datos observados y los valores de predicción son muy similares, lo cual es una buena indicación de un adecuado ajuste del modelo DBBR.

### Manejo de datos faltantes

Para hacer frente al efecto de los datos faltantes, se generaron algunos valores faltantes en la matriz de variables explicativas suponiendo que los valores son faltantes de forma completamente aleatoria (missing completely at random, MCAR). Se consideraron pérdidas de información del 5 %, 10 % y 20 % aproximadamente. Luego se calculan  $\lambda_{x_{k_1}}$  y  $\lambda_{z_{k_2}}$ , las cuales son 0 si las dos matrices de distancias son euclidianas, mientras que si  $\lambda_{x_{k_1}} < 0$  y/o  $\lambda_{z_{k_2}} < 0$  se realiza la transformación (2.38). Los resultados obtenidos se presentan en la Tabla 2.4, donde también se considera el caso de datos no faltantes para fines

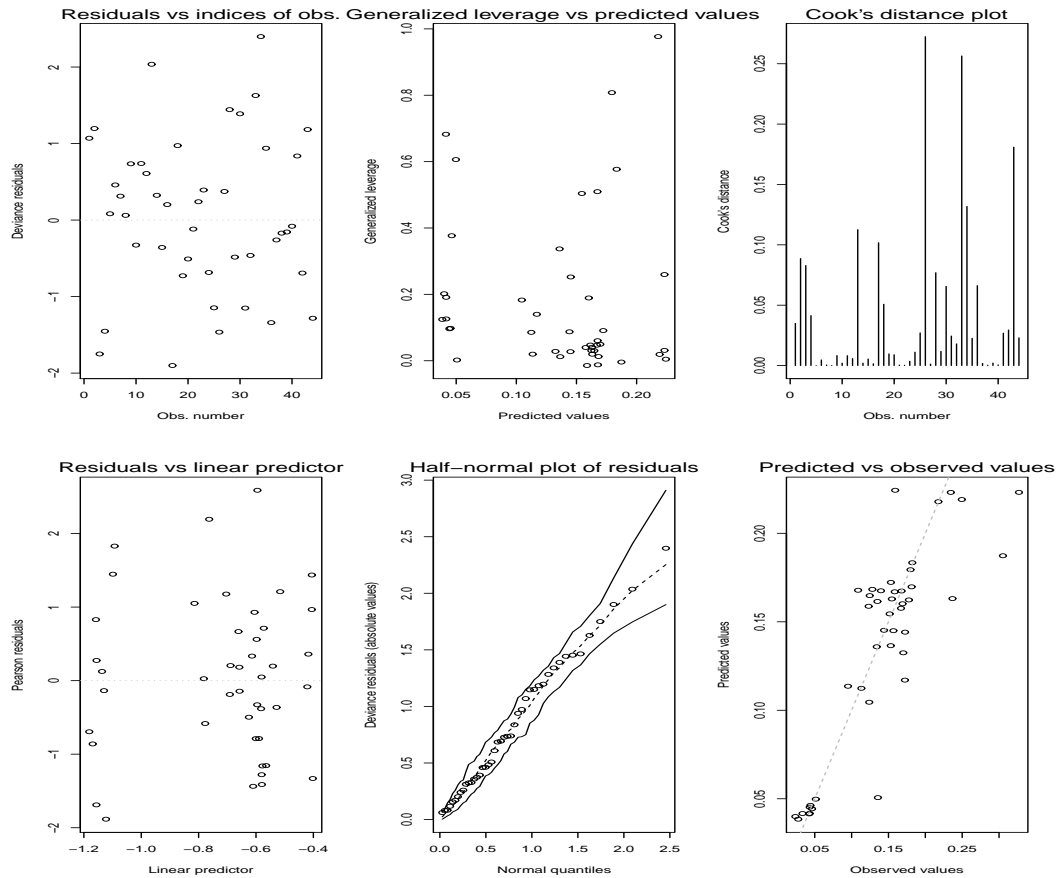


FIGURA 2.2: Diagnósticos del modelo de dispersión variable para el rendimiento de fondos de inversión.

de comparación. Las funciones de enlace loglog y log fueron elegidas para los modelos de media y de dispersión variable, respectivamente.

Se utilizan los métodos descritos en la subsección 2.3.2 para ajustar el modelo DBBR en todos los casos presentados en la Tabla 2.4; el modelo DBBR claramente funciona bien a pesar de los valores faltantes. De acuerdo a las estadísticas pseudo- $R^2$  y error cuadrático medio (mean squared error,  $MSE = \sum_{i=1}^n (\hat{\eta}_i - \hat{\eta}_{(i)})^2 / (n - k_1 - k_2)$  donde cada  $\hat{\eta}_{(i)}$  es la predicción por validación cruzada dejando fuera la  $i$ -ésima observación (leave-one out procedure)), las predicciones son realizadas mediante el modelo DBBR utilizando cualquiera de los métodos del modelo de media (M1, M2 y M3) independientemente de los métodos S1, S2 y S3 para el modelo de dispersión variable. En general, hay una tendencia a aumentar el AIC (o a disminuir la loglik) a medida que el porcentaje de valores faltantes crece para los métodos M1, M2 y M3 en el modelo de media independientemente de los métodos S1, S2 y S3 para el modelo de dispersión variable. Esto significa que la calidad de la predicción

TABLA 2.4: DBBR con dispersión variable sobre el conjunto de datos Morningstar con: 0 % (sin datos faltantes), 5 %, 10 % y 20 % de las observaciones con datos faltantes, y elección de  $\kappa_v = |\lambda_{x_{k_1}}|$  y  $\kappa_u = |\lambda_{z_{k_2}}|$ .

		Faltantes				
		0 %	5 %	10 %	20 %	
Métodos		$\lambda_m$	0.00	-0.24	-0.46	-0.75
		$\gamma_m$	0.00	-0.49	-0.64	-1.05
M1	S1	AIC	-183.42	-182.39	-161.22	-156.08
		loglik	104.71	104.20	93.61	91.03
		$R^2$	0.80	0.83	0.79	0.78
		$MSE(\%)$	0.26	0.18	0.42	0.31
	S2	AIC	-208.95	-180.33	-162.45	-161.13
		loglik	117.47	103.16	94.22	93.57
		$R^2$	0.79	0.84	0.75	0.77
		$MSE(\%)$	0.16	0.17	0.27	0.28
	S3	AIC	-197.07	-216.68	-177.46	-159.19
		loglik	111.53	121.34	101.73	92.60
		$R^2$	0.82	0.83	0.72	0.79
		$MSE(\%)$	0.18	0.16	0.25	0.26
M2	S1	AIC	-183.42	-182.40	-161.60	-156.08
		loglik	104.71	104.20	93.80	91.03
		$R^2$	0.80	0.83	0.78	0.78
		$MSE(\%)$	0.26	0.18	0.33	0.31
	S2	AIC	-208.95	-180.33	-164.17	-161.13
		loglik	117.47	103.16	95.09	93.57
		$R^2$	0.79	0.84	0.73	0.77
		$MSE(\%)$	0.16	0.17	0.40	0.28
	S3	AIC	-197.07	-216.68	-176.93	-159.19
		loglik	111.53	121.34	101.47	92.60
		$R^2$	0.82	0.83	0.73	0.79
		$MSE(\%)$	0.18	0.16	0.24	0.26
M3	S1	AIC	-184.44	-177.68	-166.55	-151.67
		loglik	105.22	101.84	96.28	88.83
		$R^2$	0.79	0.82	0.77	0.74
		$MSE(\%)$	0.23	0.22	0.25	0.30
	S2	AIC	-166.77	-175.49	-170.37	-150.58
		loglik	96.38	100.75	98.18	88.29
		$R^2$	0.83	0.82	0.75	0.73
		$MSE(\%)$	0.30	0.19	0.18	0.29
	S3	AIC	-196.45	-207.94	-176.46	-154.71
		loglik	111.53	116.97	101.23	90.35
		$R^2$	0.81	0.82	0.76	0.73
		$MSE(\%)$	0.19	0.17	0.20	0.25

---

del modelo empeora a medida que el porcentaje de valores faltantes crece. Los métodos M1 y M2 producen resultados casi idénticos, y en general, estos métodos muestran mejores resultados que el método M3, independientemente de la selección del método empleado en el modelo de dispersión.



## Capítulo 3

# Modelos lineales generalizados espaciales mixtos basados en distancias

### 3.1 Introduccción

Los modelos de riesgo a partir de datos ambientales han demostrado su efectividad ampliamente en la delimitación del riesgo en áreas geográficas porque son intuitivamente fáciles de entender. En particular, el *Loa loa* se ha convertido en un gusano filaria de gran importancia en la salud pública en Camerún. Altas densidades de *Loa loa* microfilaria (embriones que deposita la hembra) están asociadas con altas prevalencias de la infección *Loa loa* (Boussinesq et al. 2001); se considera que las personas que viven en un área de alta prevalencia están en riesgo relativamente alto de experimentar reacciones encefalopáticas a la ivermectina (Diggle et al. 2007). Para abordar estos problemas, que se aplican a *Loa loa*, datos detallados epidemiológicos de Camerún han sido analizados utilizando un modelado de regresión logística estándar (Kamgno et al. 1997, Thomson et al. 1999, Boussinesq et al. 2001). Mas aún, estos datos fueron analizados por Diggle et al. (2007), ellos realizaron un modelamiento estadístico espacial e hicieron inferencia Bayesiana para cuantificar la incertidumbre en las predicciones de una región presentando un mapa de riesgo ambiental para el *Loa loa*.

Para resolver el anterior problema, los modelos lineales generalizados (GLMs) introducidos por Nelder & Wedderburn (1972) proveen una estructura unificada para el análisis. Se han propuesto diferentes maneras de ampliar el GLM clásico para datos dependientes, entre los cuales quizás los más ampliamente utilizados son: el modelo marginal (Liang & Zeger 1986) y el modelo mixto (Breslow & Clayton 1993). Después de la primera contribución de Nelder

& Wedderburn (1972), McCullagh & Nelder (1989) propusieron los GLMs con la finalidad de unificar los modelos y técnicas de modelamiento para analizar datos más generales (datos de conteo y datos politómicos).

Algunos autores (Laird & Ware 1982, Stiratelli et al. 1984, Schall 1991) consideraron una generalización natural de los GLMs para modelar datos correlacionados no-normales incorporando términos aleatorios en el predictor lineal. Los modelos resultantes se denominan modelos lineales generalizados mixtos (generalised linear mixed models, GLMMs), los cuales proporcionan una conveniente y flexible manera para modelar datos multivariados no-normales. De acuerdo a Diggle & Ribeiro (2007), un GLM geoestadístico es un GLM mixto orientado específicamente a datos geoestadísticos. En particular, los GLMMs constituyen un unificado marco para modelar datos geoestadísticos no normales, usando términos mixtos para modelar procesos espaciales subyacentes; la aplicación particular de los GLMMs a datos geoestadísticos es conocida como GLMMs espaciales (Diggle et al. 1998, Zhang 2002).

Por otra parte, en muchas disciplinas relacionadas con análisis de datos espaciales (epidemiología, minería, hidrogeología, ecología, ciencias de la tierra y ambiental, entre otras), los investigadores a menudo tienen que enfrentarse con variables explicativas de distinta naturaleza que están asociadas con la variable respuesta: variables categóricas y binaria tales como el tipo de suelo o roca, y variables continuas (por ejemplo, las coordenadas espaciales o la elevación del terreno). De acuerdo a lo visto en el Capítulo 2, en estos casos, el concepto geométrico de distancia entre individuos o poblaciones se puede utilizar (Cuadras & Arenas 1990, Cuadras et al. 1996, Arenas & Cuadras 2002).

Para modelos geoestadísticos, el interés principal es a menudo la predicción de la variable respuesta; sin embargo, los parámetros del modelo no tienen el mismo interés (Christensen 2004). En una primera etapa de un análisis geoestadístico es común investigar la estructura del modelo, esto es, si los datos se deben transformar, si se debe considerar el problema de la anisotropía, además cuál función de correlación se debe utilizar?, y así sucesivamente. En geoestadística clásica, esta investigación inicial se hace utilizando gráficas, pero como sugiere Christensen et al. (2001), se podría además estudiar el perfil de la función de verosimilitud. Adicionalmente, solamente se pueden hacer predicciones sobre la variable respuesta en localizaciones dentro de la región geográfica de estudio en donde las variables explicativas han sido observadas.

Las anteriores consideraciones motivan la estimación de parámetros por máxima verosimilitud en un GLM espacial. La máxima verosimilitud vía Monte Carlo (Monte Carlo maximum likelihood, MCML) (Christensen 2004, Geyer & Thompson 1992, Geyer 1994, Steinsland 2007) es un enfoque alternativo al enfoque de gradiente Monte Carlo utilizado por Zhang (2002).

En este capítulo se propone una solución alterna para resolver problemas

como el de Loa loa utilizando distancias euclidianas entre individuos; se describe un modelo lineal generalizado espacial mixto incorporando medidas generales de distancia o disimilaridad que se pueden aplicar a variables explicativas: numéricas, categóricas o una mezcla de ellas. Por lo tanto, se utiliza el método DB para el modelamiento de variables respuesta aleatorias continuas y/o categóricas a través de modelos lineales generalizados espaciales mixtos basados en distancias (distance-based spatial generalised linear mixed models, DBSGLMM). Los parámetros involucrados en el modelo propuesto se estiman utilizando máxima verosimilitud mediante el método de Monte Carlo vía cadenas de Markov (Markov chain Monte Carlo, MCMC), la cual es una técnica muy útil para el análisis de este tipo de información. Esta estrategia permite una inferencia más completa que la del gradiente Monte Carlo esperanza-maximización (Monte Carlo expectation-maximization gradient, MCEMG) porque se pueden obtener las funciones de verosimilitud y el perfil de las funciones de verosimilitud, y no solamente, las estimaciones de máxima verosimilitud. Por consiguiente, el DBSGLMM se utiliza tanto para predecir la tendencia como para estimar la estructura de covarianza. Además, se presentan los métodos de inferencia para el modelo propuesto y las pruebas de bondad de ajuste.

El método se ilustra a través del análisis de la variación en la prevalencia de Loa loa en una muestra de residentes en Camerún. Esta prevalencia se relaciona con las variables explicativas: elevación o altura (elevation, ELE), el máximo índice de vegetación de la diferencia normalizada (normalised difference vegetation index, NDVI) y la desviación estándar del NDVI calculada a partir de escaneos satelitales repetidos sobre el tiempo. Como se mencionó anteriormente, estos datos fueron previamente estudiados por Thomson et al. (2004) y Diggle et al. (2007). Se utiliza DBSGLMM para predecir la tendencia y estimar la estructura de covarianza empleando el MCML, el cual ayuda a seleccionar el mejor modelo y hacer inferencia sobre las coordenadas principales obtenidas del método DB. En este estudio, también se consideró adicionalmente la variación no espacial que no podía ser atribuido a la distribución binomial del error, pero no fue significativa. El gráfico de prevalencia observada de Loa loa microfilaria versus la prevalencia pronosticada utilizando el enfoque DBSGLMM muestra sustancialmente menor dispersión que los ajustado por Thomson et al. (2004) y Diggle et al. (2007), lo cual sugiere que el método DBSGLMM funciona muy bien.

Este capítulo está dividido de la siguiente manera: en la Sección 3.2 se desarrolla la metodología propuesta, el DBSGLMM, la representación espectral, el algoritmo MCMC para el DBSGLMM, algunas medidas de bondad de ajuste y la selección de las coordenadas principales. En la Sección 3.3 se presenta la predicción espacial de un nuevo individuo junto con su valor esperado y varianza, en la Sección 3.4 se compara el método GLMM espacial clásico con el



método DBSGLMM y en la Sección 3.5 se desarrolla una aplicación que ilustra la metodología propuesta.

## 3.2 Modelos espaciales mixtos basados en distancias

Sea  $\mathbf{s} \in \mathbb{R}^d$  una localización cualquiera del espacio Euclídiano  $d$ -dimensional (en general  $d = 2$ , aunque no necesariamente), y suponga que se está interesado en analizar un determinado fenómeno de interés que toma un valor aleatorio  $y(\mathbf{s})$  en cada localización espacial  $\mathbf{s}$  (Cressie 1993). Si  $\mathbf{s}$  varía sobre un determinado conjunto  $D \subseteq \mathbb{R}^d$ , se tendrá el proceso aleatorio  $\{y(\mathbf{s}), \mathbf{s} \in D\}$ , el cual 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.

Por otra parte, una distribución pertenece a la familia exponencial si tiene una función de densidad dada por

$$f(y(\mathbf{s}); \alpha_s) = h_1(y(\mathbf{s})) \exp\{\eta(\alpha_s)h_2(y(\mathbf{s})) - b(\alpha_s)\}$$

donde  $\eta(\alpha_s)$ ,  $b(\alpha_s)$ ,  $h_1(y(\mathbf{s}))$  y  $h_2(y(\mathbf{s}))$  son funciones que toman valores en la recta real.

La interpolación propuesta se construye para un modelo en un campo aleatorio no gaussiano considerando específicamente variables explicativas categóricas y continuas. Los datos se generan mediante un mecanismo condicional sobre que el modelo sigue un GLM clásico como el descrito por McCullagh & Nelder (1989). Explícitamente, al condicionar sobre  $z(\mathbf{s}_i)$  y  $e(\mathbf{s}_i)$ , las respuestas  $y(\mathbf{s}_i)$  en las localizaciones  $\mathbf{s}_i$  con  $i = 1, \dots, n$  son variables aleatorias mutuamente independientes cuya esperanza condicional,  $\mu(\mathbf{s}_i) = E[y(\mathbf{s}_i)|\mathbf{v}(\mathbf{s}_i), z(\mathbf{s}_i), e(\mathbf{s}_i)]$ , está determinada como

$$\eta_i = g(\mu(\mathbf{s}_i)) = \gamma_0 + \mathbf{v}^t(\mathbf{s}_i)\boldsymbol{\gamma} + z(\mathbf{s}_i) + e(\mathbf{s}_i) \quad (3.1)$$

donde  $g(\cdot)$  es una función de enlace que es invertible y continua,  $\gamma_0 + \mathbf{v}^t(\mathbf{s}_i)\boldsymbol{\gamma}$  es la tendencia,  $\gamma_0$  es un parámetro desconocido asociado al intercepto,  $\mathbf{v}(\mathbf{s}_i) = (v_1(\mathbf{s}_i), \dots, v_p(\mathbf{s}_i))^t$  es un vector conformado por las variables explicativas asociadas a la localización espacial  $\mathbf{s}_i$  y  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^t$  es un vector de parámetros desconocidos asociados a la regresión espacial. Se llama la distribución del error a la función de distribución condicional de cada  $y(\mathbf{s}_i)$  dados  $z(\mathbf{s}_i)$  y  $e(\mathbf{s}_i)$ .

Se supone que los  $z(\mathbf{s}_i)$ 's tienen una estructura de campo aleatorio, cuya función de covarianza está caracterizada por un parámetro dimensional finito,  $\boldsymbol{\theta}$ , denominado el componente de varianza espacial. Además, el término de ruido no observado  $e(\mathbf{s}_i)$  puede o no ser incluido, el cual depende de la aplicación. Nominalmente, tal término de error asimila la sobre-dispersión relativa a la varianza media, relación implícita por la función de distribución bajo consideración. Los componentes del error se asumen mutuamente independientes con media 0 y varianza  $\tau^2$ .

Sin significativa pérdida de generalidad, en este capítulo se trabaja con el modelo (3.1) sin el término de ruido  $e(\mathbf{s}_i)$ ; así, el modelo (3.1) puede ser expresado como

$$\eta_i = g(\mu(\mathbf{s}_i)) = \gamma_0 + \mathbf{v}^t(\mathbf{s}_i)\boldsymbol{\gamma} + z(\mathbf{s}_i) \quad (3.2)$$

Dentro del campo de modelos lineales es usual trabajar con el modelo en la forma canónica ( $\eta(\alpha_{\mathbf{s}_i}) = \alpha_{\mathbf{s}_i}$ ,  $h_2(y(\mathbf{s}_i)) = y(\mathbf{s}_i)$ ), que incluye un parámetro de dispersión  $\phi > 0$ . Específicamente, condicionando sobre las variables aleatorias espaciales no observadas  $z(\mathbf{s}_i)$ ,  $y(\mathbf{s}_i)$  sigue una función de distribución de la familia exponencial, es decir,

$$y(\mathbf{s}_i) \mid (\mathbf{v}(\mathbf{s}_i), z(\mathbf{s}_i)) \stackrel{ind}{\sim} f_i(y(\mathbf{s}_i) \mid \mathbf{v}(\mathbf{s}_i), z(\mathbf{s}_i))$$

$$f_i(y(\mathbf{s}_i) \mid \mathbf{v}(\mathbf{s}_i), z(\mathbf{s}_i)) = \exp \left\{ \frac{1}{a(\phi)} [y(\mathbf{s}_i)\alpha_{\mathbf{s}_i} - b(\alpha_{\mathbf{s}_i})] + c(y(\mathbf{s}_i), \phi) \right\}$$

donde  $\phi$  es un parámetro de extra variación,  $a(\phi)$  y  $c(\cdot)$  son algunas funciones específicas. La media condicional,  $\mu(\mathbf{s}_i)$ , está relacionada con  $\alpha_{\mathbf{s}_i}$  a través de la identidad  $\mu(\mathbf{s}_i) = \frac{\partial b(\alpha_{\mathbf{s}_i})}{\partial \alpha_{\mathbf{s}_i}} = b'(\alpha_{\mathbf{s}_i})$ . Después de una transformación apropiada, esta media se modelará como el GLMM dado en (3.2) tanto en la parte fija de  $\mu(\mathbf{s}_i)$  como en los efectos espaciales aleatorios ( $z(\mathbf{s}_i)$ ).

El modelo (3.2) se puede reformular en forma vectorial como

$$g\{E(\mathbf{y}_s \mid \mathbf{V}, \mathbf{z}_s)\} = g(\boldsymbol{\mu}_s) = \gamma_0 \mathbf{1}_n + \mathbf{V}\boldsymbol{\gamma} + \mathbf{E}\mathbf{z}_s \quad (3.3)$$

donde  $\boldsymbol{\mu}_s = E(\mathbf{y}_s \mid \mathbf{V}, \mathbf{z}_s)$ ,  $\mathbf{y}_s = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))^t$ ,  $\mathbf{V} = \mathbf{H}\mathbf{V}^*$  con  $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$  una matriz centrada,  $\mathbf{I}_n$  la matriz identidad de tamaño  $n \times n$  y  $\mathbf{1}_n$  el vector de unos de tamaño  $n \times 1$ .  $\mathbf{V}^*$  es una matriz de variables explicativas; observe que  $\mathbf{V}^*$  puede incluir variables continuas, categóricas y binarias, o incluso una mezcla de ellas. Además,  $\mathbf{z}_s = (z(\mathbf{s}_1), \dots, z(\mathbf{s}_n))^t$ ,  $\boldsymbol{\mu}_s = (\mu(\mathbf{s}_1), \dots, \mu(\mathbf{s}_n))^t$  y  $\mathbf{E}$  es una matriz de diseño no aleatoria que es compatible con los efectos aleatorios  $\mathbf{z}_s$ .

### 3.2.1 Modelo mixto basado en distancias

Para construir el modelo propuesto en este capítulo se incorporan algunas medidas de distancia/disimilitud presentadas en los capítulos 1 y 2 que se pueden

aplicar a variables explicativas: continuas, categóricas o una mezcla de ellas. Para este objetivo, se necesitan definir algunas medidas de semejanza (o distancia euclidiana) que dependan de las características de las variables explicativas y que son adaptadas al caso espacial.

De acuerdo a Cuadras (1989) y Cuadras & Arenas (1990), sea  $\Omega = \{\omega_1, \dots, \omega_n\}$  un conjunto de  $n$  individuos. Sea  $d_{ii'} = d(\omega_i, \omega_{i'}) = d(\omega_{i'}, \omega_i) \geq d(\omega_i, \omega_i) = 0$  una función distancia (o disimilaridad) definida sobre  $\Omega$ . Suponga que la matriz de distancia con dimensión  $n \times n$ ,  $\mathbf{D}_v = (d_{ii'})$  es euclidiana. Entonces, existe una configuración de puntos  $\mathbf{v}(\mathbf{s}_1), \dots, \mathbf{v}(\mathbf{s}_n) \in \mathbb{R}^p$  (donde  $\mathbf{v}(\mathbf{s}_i) = (v_1(\mathbf{s}_i), \dots, v_p(\mathbf{s}_i))^t$ ,  $i = 1, \dots, n$ , está formado por variables binarias, categóricas y continuas), de tal manera que la similaridad de Gower (1968) se puede expresar para variables mixtas como

$$m_{ii'} = \frac{\sum_{j=1}^{p_c} \left(1 - \frac{|v_j(\mathbf{s}_i) - v_j(\mathbf{s}_{i'})|}{G_j}\right) + c_{1ii'} + v_{ii'}}{p_c + (p_b - c_{4ii'}) + p_q} \quad (3.4)$$

donde para los efectos fijos,  $p_c$  es el número de variables continuas,  $c_{1ii'} = c_1(\mathbf{s}_i, \mathbf{s}_{i'})$  y  $c_{4ii'} = c_4(\mathbf{s}_i, \mathbf{s}_{i'})$  son el número de coincidencias positivas y negativas, respectivamente, para las  $p_b$  variables binarias.  $v_{ii'} = v(\mathbf{s}_i, \mathbf{s}_{i'})$  es el número de coincidencias para las  $p_q$  variables multiestado y  $G_j$  es el rango (o distancia) de la  $j$ -ésima variable continua.

En el caso en el que las variables explicativas en (3.2) sean binarias o categóricas, la semejanza se puede definir por

$$m_{ii'} = \frac{c_{1ii'} + c_{4ii'}}{c_{1ii'} + c_{2ii'} + c_{3ii'} + c_{4ii'}} \quad (\text{Sokal-Michener})$$

$$m_{ii'} = \frac{c_{1ii'}}{c_{1ii'} + c_{2ii'} + c_{3ii'}} \quad (\text{Jaccard})$$

donde  $c_{1ii'}$ ,  $c_{2ii'} = c_2(\mathbf{s}_i, \mathbf{s}_{i'})$ ,  $c_{3ii'} = c_3(\mathbf{s}_i, \mathbf{s}_{i'})$ ,  $c_{4ii'}$  son las frecuencias de (1,1), (1,0), (0,1) y (0,0), respectivamente. A través de la transformación

$$d_{ii'} = \sqrt{1 - m_{ii'}}$$

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

$$d_{ii'} = \sqrt{(\mathbf{v}(\mathbf{s}_i) - \mathbf{v}(\mathbf{s}_{i'}))^t (\mathbf{v}(\mathbf{s}_i) - \mathbf{v}(\mathbf{s}_{i'}))} \quad (3.5)$$

o alternativamente por la distancia absoluta  $d_{ii'} = \sqrt{\sum_{j=1}^p |v_j(\mathbf{s}_i) - v_j(\mathbf{s}_{i'})|}$ . Expresiones para la similaridad de Gower como la presentada en la ecuación (3.4) son útiles en la medida que se tenga información asociada con variables mixtas, no sólo para las localizaciones de muestreo sino también para las localizaciones no muestreadas, lo cual limita su uso en áreas no muestreadas.

Estas distancias satisfacen que la distancia es cercana a 0 si las mediciones  $\mathbf{v}$  sobre  $\mathbf{s}_i$  y  $\mathbf{s}_{i'}$  son muy similares, es decir,  $d_{ii'} \cong 0$  si  $\mathbf{v}(\mathbf{s}_i) \cong \mathbf{v}(\mathbf{s}_{i'})$ . Después de seleccionar una de las anteriores distancias, se define  $\mathbf{A}_v = (-d_{ii'}^2/2)$  una matriz de tamaño  $n \times n$  y  $\mathbf{B}_v = \mathbf{H}\mathbf{A}_v\mathbf{H}$ . Entonces,  $\mathbf{B}$  es una matriz semi definida positiva (Mardia et al. 2002) de rango  $n-1$  y la matriz de coordenadas principales,  $\mathbf{X}$ , se obtiene de la siguiente descomposición espectral

$$\mathbf{B}_v = \mathbf{H}\mathbf{A}_v\mathbf{H} = \mathbf{L}_x\mathbf{\Lambda}_x\mathbf{L}_x^t = \mathbf{X}\mathbf{X}^t$$

donde  $\mathbf{\Lambda}_x$  es una matriz diagonal que contiene los valores propios de  $\mathbf{B}_v$  y  $\mathbf{X} = \mathbf{L}_x\mathbf{\Lambda}_x^{1/2}$  es una matriz de tamaño  $n \times n$  de rango  $n-1$  ya que tiene un valor propio igual a  $\mathbf{1}_n$  y  $\mathbf{L}_x$  contiene las coordenadas estandarizadas.

Además, las filas  $\mathbf{x}^t(\mathbf{s}_1), \dots, \mathbf{x}^t(\mathbf{s}_n)$  de  $\mathbf{X}$  son las coordenadas principales de  $\mathbf{B}_v$  con respecto a la matriz de distancia  $\mathbf{D}_v$ . Como  $\mathbf{v}(\mathbf{s}_i) \cong \mathbf{v}(\mathbf{s}_{i'})$  cuando un individuo  $i$  es similar a otro individuo  $i'$  en (3.3), es claro que  $\mathbf{x}(\mathbf{s}_i) \cong \mathbf{x}(\mathbf{s}_{i'})$ .

Uno de los peligros potenciales en la predicción usando DB es la enorme sobreparametrización ya que el rango de  $\mathbf{B}_v$  puede ser tan grande como  $n-1$ . Entonces, el número de coordenadas principales (columnas de  $\mathbf{X}$ ) puede ser excesivo, lo que permite un arbitrario sobre ajuste en el modelo. Con el fin de evitar tales problemas, se selecciona únicamente las coordenadas principales más significativas utilizando cualquier método de la sección 2.4. Por lo tanto, el DBSGLMM en forma matricial reducida se puede expresar por

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}_s) = \beta_0\mathbf{1} + \mathbf{X}_{(k)}\boldsymbol{\beta}_{(k)} + \mathbf{E}\mathbf{z}_s \quad (3.6)$$

donde  $\boldsymbol{\mu}_s = \mathbf{E}(\mathbf{y} \mid \mathbf{X}_{(k)}, \mathbf{z}_s)$ ,  $\beta_0$  y  $\boldsymbol{\beta}_{(k)}^t = (\beta_1, \dots, \beta_k)$  son parámetros desconocidos,  $\boldsymbol{\beta}_{(k)} \in \mathbb{R}^k$ ,  $k \leq n-1$ ,  $\mathbf{X}_{(k)} = (\mathbf{X}_1, \dots, \mathbf{X}_k)$  y contiene un subconjunto de  $k$  columnas correspondientes a  $\mathbf{X}$ , las cuales están significativamente correlacionadas con  $\mathbf{y}_s$ . Mas aún, cada  $\mathbf{X}_j$  ( $j = 1, \dots, k$ ) es una coordenada principal, es decir, un vector columna de  $\mathbf{X}$ .

El modelo propuesto en (3.6) se puede escribir como

$$\boldsymbol{\eta} = \beta_0\mathbf{1} + \sum_{j=1}^k \beta_j\mathbf{X}_j + \mathbf{E}\mathbf{z}_s \quad (3.7)$$

o alternativamente como

$$\eta_i = g(\mu(\mathbf{s}_i)) = \sum_{j=0}^k x_{ij}\beta_j + z(\mathbf{s}_i) = \mathbf{x}_i^t\boldsymbol{\beta} + z(\mathbf{s}_i) \quad (3.8)$$

con  $i = 1, \dots, n$ , y donde  $x_{i0} = 1$ ,  $\mathbf{x}_i^t = (x_{i0}, \dots, x_{ik})$  y  $\boldsymbol{\beta}^t = (\beta_0, \boldsymbol{\beta}_{(k)}^t) = (\beta_0, \beta_1, \dots, \beta_k)$ . Note que para el modelo (3.7) se tiene  $\mathbf{X}_j^t\mathbf{1} = 0$ ,  $\mathbf{X}_j^t\mathbf{X}_j = \lambda_j$  y  $\mathbf{X}_j^t\mathbf{X}_{j'} = 0$  para  $j \neq j'$ , con  $j, j' = 1, \dots, k$ .

Por otra parte, el modelo (3.8) es amplio y abarca una variedad de modelos, incluyendo el modelo de regresión espacial logístico como un caso especial. Específicamente, para respuestas binarias se tiene que

$$b(\alpha_{\mathbf{s}_i}) = \ln(1 + e^{\alpha_{\mathbf{s}_i}}), \quad a(\phi) = \phi \equiv 1, \quad c(y(\mathbf{s}_i), \phi) = \ln \binom{n}{y(\mathbf{s}_i)},$$

y para respuestas continuas se sigue que

$$b(\alpha_{\mathbf{s}_i}) = -\frac{1}{2}\alpha_{\mathbf{s}_i}^2, \quad \phi = \sigma^2, \quad c(y(\mathbf{s}_i), \sigma^2) = -\frac{1}{2\sigma^2}y(\mathbf{s}_i)^2 - \frac{1}{2}\ln(2\pi\sigma^2),$$

y  $g(\cdot)$  una función identidad, así el modelo (3.8) se reduce al modelo lineal mixto clásico basado en distancias.

Para datos de conteo se tiene

$$b(\alpha_{\mathbf{s}_i}) = e^{\alpha_{\mathbf{s}_i}}, \quad a(\phi) = \phi \equiv 1, \quad c(y(\mathbf{s}_i), \phi) = \ln y(\mathbf{s}_i)!$$

y haciendo  $g(\mu(\mathbf{s}_i)) = \ln(\mu(\mathbf{s}_i))$ , la ecuación (3.8) da como resultado un modelo de regresión Poisson mixto.

Algunos de los casos anteriores se pueden considerar en una clase general de funciones de enlace, Aranda-Ordaz (1981) propuso una familia de funciones de enlace para analizar datos en forma de proporciones dada por:

$$g_\nu(\mu(\mathbf{s}_i)) = \log \left[ \frac{(1 - \mu(\mathbf{s}_i))^{-\nu} - 1}{\nu} \right]$$

siendo  $\nu$  una constante desconocida que tiene como casos particulares el modelo logístico para  $\nu = 1$  y el complemento log-log para  $\nu \rightarrow 0$ .

Otra forma general de funciones de enlace propuesta por Box & Cox (1964) y utilizada principalmente para datos con media positiva, trabajada en el caso espacial por Christensen (2004), es la transformación Box-Cox que esta especificada por

$$g_\nu(\mu(\mathbf{s}_i)) = \begin{cases} (\mu^\nu(\mathbf{s}_i))/\nu & \text{si } \nu > 0 \\ \log(\mu(\mathbf{s}_i)) & \text{si } \nu = 0 \end{cases}$$

Al trabajar con esta última transformación para modelar el estudio de Loa loa, se obtiene lo siguiente: como la función enlace  $g_\nu(\mu(\mathbf{s}_i))$  relaciona  $\boldsymbol{\mu}_s$  con  $\mathbf{X}$  y  $\mathbf{z}_s$ , se tiene que

$$\boldsymbol{\mu}_s = m_s g_\nu^{-1}(\mathbf{X}, \mathbf{z}_s)$$

donde  $m_s$  es una función determinística. Si  $\nu > 0$  entonces  $g_\nu(\mathbb{R}) = (-1/\nu, \infty)$ , para lo cual es necesario definir  $g_\nu^{-1}(\mu(\mathbf{s}_i)) = 0$  cuando  $\mu(\mathbf{s}_i) \notin g_\nu(\mathbb{R})$  y  $f(y(\mathbf{s}_i) | \mu(\mathbf{s}_i)) = 1_{\{y(\mathbf{s}_i)=0\}}$  cuando  $\mu(\mathbf{s}_i) = 0$  (ver detalles en Christensen (2004)).

El modelo (3.6) no es un simple cambio de formato ya que éste se acomoda a una estructura de datos más complejo; más allá de los datos espaciales. Por ejemplo, definiendo apropiadamente  $\mathbf{E}$  y los efectos aleatorios  $\mathbf{z}_s$ , el modelo (3.6) abarca datos agrupados o de cluster no-normales y datos de factores cruzados (Breslow & Clayton 1993). Cuando  $\mathbf{E}$  se define como una matriz que indica la pertenencia a una determinada región espacial (por ejemplo, condado o distrito censal), el modelo (3.6) se puede utilizar para ajustar datos de áreas.

La base de muchos procedimientos, incluyendo el descrito anteriormente, es una función de distribución normal multivariada. En este caso, considere el proceso espacial  $\mathbf{z}_s$  y asume que éste tiene función de distribución normal multivariada (multivariate normal, MN)  $\mathbf{z}_s \sim MN(\mathbf{0}, \Sigma_z)$ . La clave para el eficiente modelamiento espacial en problemas de grandes muestras es la eficiente parametrización de esta normal multivariada. Reescribiendo  $\mathbf{z}_s$  en términos de la descomposición espectral,

$$\mathbf{z}_s = \Psi \boldsymbol{\delta} \quad (3.9)$$

donde  $\Psi_{n \times n}$  es una matriz de funciones base  $\Psi = (\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_n)$  con  $\boldsymbol{\psi}_i = (\psi_i(\mathbf{s}_1), \dots, \psi_i(\mathbf{s}_n))^t$  y  $\boldsymbol{\delta}$  es un vector  $n \times 1$  de coeficientes espectrales (proyecciones del proceso  $\mathbf{z}_s$  sobre las funciones base), con función de distribución  $\boldsymbol{\delta} \sim MN(\mathbf{0}, \Sigma_\delta)$ . Observe que si las funciones base son ortogonales, entonces  $\boldsymbol{\delta} = \Psi^t \mathbf{z}_s$  y  $\Sigma_\delta = \Psi^t \Sigma_z \Psi$  porque  $\Psi^t \Psi = \Psi \Psi^t = \mathbf{I}$ .

Según Royle & Wikle (2005), el proceso  $\mathbf{z}_s$  se puede expandir en términos de funciones base de Fourier (es decir, senos y cosenos). Además, asumiendo que el proceso espacial está definido en un retículo regular, para las localizaciones  $\mathbf{s}_i$ ,  $i = 1, \dots, n$  y frecuencias espaciales  $w_q = q/n$  para  $q = 0, \dots, n/2$  ( $n$  par), se tiene

$$\begin{aligned} z(\mathbf{s}_i) &= \sum_{q=0}^{n/2} \delta^{(1)}(q) \cos(2\pi \mathbf{s}_i w_q) + \sum_{q=1}^{n/2-1} \delta^{(2)}(q) \sin(2\pi \mathbf{s}_i w_q) \\ &= (\boldsymbol{\psi}_i^{(1)})^t \boldsymbol{\delta}^{(1)} + (\boldsymbol{\psi}_i^{(2)})^t \boldsymbol{\delta}^{(2)} = \boldsymbol{\psi}_i^t \boldsymbol{\delta} \end{aligned}$$

donde  $\boldsymbol{\psi}_i^{(1)} = (\psi_i^{(1)}(w_0), \dots, \psi_i^{(1)}(w_{n/2}))^t$ ,  $\boldsymbol{\psi}_i^{(2)} = (\psi_i^{(2)}(w_1), \dots, \psi_i^{(2)}(w_{n/2-1}))^t$ ,  $\psi_i^{(1)}(w_q) = \cos(2\pi \mathbf{s}_i w_q)$ ,  $\psi_i^{(2)}(w_q) = \sin(2\pi \mathbf{s}_i w_q)$ ,  $\boldsymbol{\psi}_i = \left( (\boldsymbol{\psi}_i^{(1)})^t, (\boldsymbol{\psi}_i^{(2)})^t \right)^t$ ,  $\boldsymbol{\delta}^{(1)} = (\delta^{(1)}(0), \dots, \delta^{(1)}(n/2))^t$ ,  $\boldsymbol{\delta}^{(2)} = (\delta^{(2)}(1), \dots, \delta^{(2)}(n/2 - 1))^t$  y  $\boldsymbol{\delta} = \left( (\boldsymbol{\delta}^{(1)})^t, (\boldsymbol{\delta}^{(2)})^t \right)^t$ .

Es bien conocido que para procesos aleatorios estacionarios de segundo orden, los coeficientes  $\delta_{(q)}$  son casi no correlacionados y sus varianzas a una frecuencia dada son aproximadamente iguales a la mitad de la densidad espectral de potencia a esa frecuencia, excepto para frecuencias  $w_0$  y  $w_{n/2}$  en el que la varianza es igual a la densidad de potencia espectral asociada (Shumway

& Stoffer 2000). Así, asumiendo que,  $\mathbf{z}_s$  es estacionario de segundo orden con matriz de covarianza  $\Sigma_z = \sigma^2 \mathbf{R}$  donde  $\mathbf{R}$  es la matriz de correlación, entonces  $\text{Cov}(\boldsymbol{\delta}) \approx \sigma^2 \mathbf{C}$  donde  $\mathbf{C}$  es una matriz diagonal con los elementos de la diagonal dados por  $[f(w_0), \frac{1}{2}f(w_1), \dots, \frac{1}{2}f(w_{n/2}), \frac{1}{2}f(w_1), \dots, \frac{1}{2}f(w_{n/2-1})]$ , con  $f(w_q)$  la densidad espectral en la frecuencia  $w_q$  correspondiente a la función de correlación utilizada para construir  $\mathbf{R}$ .

En este caso, se parametriza la matriz de correlación  $\mathbf{R}(\boldsymbol{\theta})$  en términos de un vector de parámetros de dependencia espacial  $\boldsymbol{\theta}$ . Así, la matriz diagonal  $\mathbf{D}(\boldsymbol{\theta})$  es también una función de  $\boldsymbol{\theta}$ . En el análisis presentado en esta sección, la matriz de covarianza de Matérn con la función de densidad espectral asociada a la frecuencia  $w$  está dada por Royle & Wikle (2005)

$$f(w) = \frac{2^{\kappa-1} \sigma_w \Gamma(\kappa + d_w/2) \vartheta_w^{2\kappa}}{\pi^{d_w/2} (\vartheta_w^2 + w^2)^{\kappa + d_w/2}}, \quad \sigma_w > 0, \vartheta_w > 0, \kappa > 0 \quad (3.10)$$

donde  $d_w$  es la dimensionalidad del proceso espacial (Stein 1999, p. 49),  $\kappa$  está relacionado con el grado de suavidad del proceso espacial,  $\vartheta_w$  está relacionado con el rango de correlación y  $\sigma_w$  es proporcional a la varianza del proceso (Stein 1999, p. 48). Así, si se elige  $\boldsymbol{\Psi}$  como funciones base de Fourier, entonces (3.10) sugiere la forma de  $\Sigma_\delta(\boldsymbol{\theta})$  (matriz diagonal, con los elementos de la diagonal correspondientes a la frecuencia  $w$  dada por (3.10)).

Hay algunas ventajas al escribir el proceso espacial  $\mathbf{z}_s$  en términos del proceso espectral,  $\boldsymbol{\delta}$ . Primero, el operador espectral a menudo actúa como un operador que elimina la correlación, de tal forma que en  $\Sigma_\delta$  las correlaciones entre los individuos son relativamente bajas en comparación con las de  $\Sigma_z$ . Además, el cálculo computacional es eficiente, y en algunos casos, se logra una reducción de dimensión (Royle & Wikle 2005).

Por último, sustituyendo (3.9) en el modelo (3.6) se obtiene la siguiente versión del DBSGLMM reducido

$$\begin{aligned} \boldsymbol{\eta} &= g(\boldsymbol{\mu}_s) = g\{\mathbb{E}(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta})\} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}\boldsymbol{\Psi}\boldsymbol{\delta} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{E}_1\boldsymbol{\delta} \end{aligned} \quad (3.11)$$

donde  $\mathbf{X} = (\mathbf{1} \quad \mathbf{X}_{(k)})$  y  $\mathbf{E}_1 = \mathbf{E}\boldsymbol{\Psi}$ .

### 3.2.2 Algoritmo de máxima verosimilitud de Monte Carlo para DBSGLMM

La aplicación de los métodos basados en verosimilitud a DBSGLMM no gaussianos está obstaculizado por dificultades computacionales que surgen debido a la gran dimensionalidad del vector aleatorio no observado  $\boldsymbol{\delta} = \{\delta(\mathbf{s}_1), \dots, \delta(\mathbf{s}_n)\}$ . En esta sección se estiman los parámetros del DBSGLMM

propuesto utilizando máxima verosimilitud MCMC (Geyer & Thompson 1992, Geyer 1994, Højbjerg 2003, Christensen 2004).

Asumiendo que en el modelo (3.11) cada  $Y(\mathbf{s}_i)$ , ( $i = 1, \dots, n$ ) tiene una distribución de la familia exponencial y por independencia de los  $Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n)$  dados  $\mathbf{X}$ ,  $\boldsymbol{\delta}$  y  $g_\nu^{-1}(\cdot)$ , la función de densidad condicional de  $\mathbf{Y}_s = \mathbf{y}_s$  dadas las covariables observadas  $\mathbf{X}$  y  $\boldsymbol{\delta}$  es

$$f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \nu) = \prod_{i=1}^n f\{y(\mathbf{s}_i); m_i g_\nu^{-1}[\mathbf{x}(\mathbf{s}_i), \delta(\mathbf{s}_i); \boldsymbol{\beta}]\} \quad (3.12)$$

donde  $m_i = m(\mathbf{s}_i)$  y  $\nu$  es la función de enlace.

Ahora, desde la perspectiva clásica, la función de verosimilitud basada en las variables aleatorias observadas  $\mathbf{y}_s$  se obtiene marginalizando con respecto a las variables no observadas  $\boldsymbol{\delta}$ , llevando a la verosimilitud del modelo mixto. Entonces, la función de verosimilitud para el DBSGLMM no es expresable en forma cerrada, sino solo como una integral de alta dimensión

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \nu) &= f(\mathbf{y}_s | \boldsymbol{\beta}, \boldsymbol{\theta}, \nu) = \int_{\mathbb{R}^n} f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \nu) f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta}) d\boldsymbol{\delta} \\ &= \int_{\mathbb{R}^n} \prod_{i=1}^n f\{y(\mathbf{s}_i); m_i g_\nu^{-1}[\mathbf{x}(\mathbf{s}_i), \delta(\mathbf{s}_i); \boldsymbol{\beta}]\} f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta}) d\delta(\mathbf{s}_1) \cdots d\delta(\mathbf{s}_n) \end{aligned} \quad (3.13)$$

donde  $f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta})$  es la función de distribución conjunta de  $\boldsymbol{\delta}$  dadas las covariables observadas  $\mathbf{X}$ , con  $\boldsymbol{\theta}$  el vector de parámetros asociado a  $\boldsymbol{\delta}$ . La integral anterior es la constante normalizada en la función de densidad condicional de  $\boldsymbol{\delta}$  dado  $\mathbf{y}_s$ ,

$$f(\boldsymbol{\delta} | \mathbf{y}_s, \boldsymbol{\beta}, \boldsymbol{\theta}, \nu) \propto \prod_{i=1}^n f\{y(\mathbf{s}_i); m_i g_\nu^{-1}[\mathbf{x}(\mathbf{s}_i), \delta(\mathbf{s}_i); \boldsymbol{\beta}]\} f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta}) \quad (3.14)$$

MCMC provee un método para la simulación de (3.14) y la aproximación (3.13).

La integral en (3.13) tiene una dimensión alta, y por consiguiente, no se puede encontrar las estimaciones de máxima verosimilitud (maximum likelihood estimates, MLEs) vía maximización directa. Entonces, si se toma el caso donde  $\nu$  es fija, este término se suprime. Así, la función de verosimilitud (3.13) puede expresarse como

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \int_{\mathbb{R}^n} f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}) f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta}) d\boldsymbol{\delta} \\ &= \int_{\mathbb{R}^n} \frac{f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}) f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta})}{\tilde{f}(\mathbf{y}_s, \boldsymbol{\delta})} \tilde{f}(\mathbf{y}_s, \boldsymbol{\delta}) d\boldsymbol{\delta} \\ &\propto \int_{\mathbb{R}^n} \frac{f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}) f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta})}{f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}_0) \tilde{f}(\boldsymbol{\delta})} \tilde{f}(\boldsymbol{\delta} | \mathbf{y}_s) d\boldsymbol{\delta} \\ &= \tilde{\mathbb{E}} \left[ \frac{f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}) f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta})}{f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}_0) \tilde{f}(\boldsymbol{\delta})} \middle| \mathbf{y}_s \right] \end{aligned} \quad (3.15)$$



donde  $\tilde{f}(\mathbf{y}_s, \boldsymbol{\delta}) = f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}_0) \tilde{f}(\boldsymbol{\delta})$  y  $\tilde{f}(\boldsymbol{\delta})$  es alguna función de densidad con soporte en  $\mathbb{R}^n$ ,  $\tilde{f}(\boldsymbol{\delta} | \mathbf{y}_s) \propto f(\mathbf{y}_s | \boldsymbol{\delta}) \tilde{f}(\boldsymbol{\delta})$  es la función de densidad condicional y  $\tilde{\mathbb{E}}(\cdot | \mathbf{y}_s)$  denota la esperanza con respecto a  $\tilde{f}(\cdot | \mathbf{y}_s)$ , la cual depende de un valor inicial de  $\boldsymbol{\beta}$ ,  $\boldsymbol{\beta}_0$ . Los MLEs se pueden calcular mediante la maximización de la aproximación de Monte Carlo de la verosimilitud (3.15), mediante la siguiente expresión

$$L_r(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{r} \sum_{j=1}^r \frac{f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}(j); \boldsymbol{\beta}) f(\boldsymbol{\delta}(j) | \mathbf{X}; \boldsymbol{\theta})}{f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}(j); \boldsymbol{\beta}_0) \tilde{f}(\boldsymbol{\delta}(j))} \quad (3.16)$$

donde los  $\boldsymbol{\delta}(j)$ 's ( $j = 1, \dots, r$ ) son muestreados por MCMC de la función de distribución  $\tilde{f}(\cdot | \mathbf{y}_s)$ . Como se observa de (3.15) se podría seleccionar  $\tilde{f}(\cdot)$  cercano a  $f(\cdot | \boldsymbol{\theta})$ , donde  $\hat{\boldsymbol{\theta}}$  es el MLE de  $\boldsymbol{\theta}$ , porque de lo contrario uno o muy pocos de los términos  $f(\boldsymbol{\delta}(j) | \boldsymbol{\beta}, \boldsymbol{\theta}) / \tilde{f}(\boldsymbol{\delta}(j))$ ,  $j = 1, \dots, r$  pueden llegar a dominar a los otros en  $L_r(\boldsymbol{\beta}, \boldsymbol{\theta})$ , los cuales hacen la aproximación menos útil.

A continuación, se presenta un procedimiento numérico para maximizar la aproximación de Monte Carlo (3.16). Se reparametriza por conveniencia computacional el efecto pepita,  $\tau^2$ , utilizando el efecto pepita relativo  $\tau_R = \tau^2 / \sigma^2$ ; por consiguiente, la matriz de covarianza de  $\boldsymbol{\eta}$  es  $\sigma^2(\mathbf{R}(\vartheta) + \tau_R^2 \mathbf{I}_n)$ . De esta manera, sea  $(\boldsymbol{\beta}, \boldsymbol{\theta}) = (\boldsymbol{\beta}, \sigma^2, \vartheta, \tau_R^2)$ , la maximización de  $L_r$  con respecto a  $\boldsymbol{\beta}$  y  $\sigma^2$  dado  $\vartheta$  y  $\tau_R^2$  es bastante sencilla porque la primera y segunda derivadas de la función de densidad normal  $f(\boldsymbol{\eta}(j) | \mathbf{X}, \boldsymbol{\beta}, \sigma^2, \vartheta, \tau_R^2)$ ,  $j = 1, \dots, r$ , con respecto a esos parámetros son simples. Lo anterior se hace utilizando un algoritmo iterativo como el de Newton-Raphson que es computacionalmente rápido; para realizar este algoritmo, se puede utilizar un procedimiento iterativo apropiado para los valores iniciales como el siguiente

$$\begin{aligned} \boldsymbol{\beta}(j) &= [\mathbf{X}^t (R(\vartheta) + \tau_R^2 \mathbf{I}_n)^{-1} \mathbf{X}]^{-1} \mathbf{X}^t (R(\vartheta) + \tau_R^2 \mathbf{I}_n)^{-1} \boldsymbol{\eta}(j) \\ \sigma^2(j) &= \frac{1}{n} [\boldsymbol{\eta}(j) - \mathbf{X}\boldsymbol{\beta}(j)]^t (R(\vartheta) + \tau_R^2 \mathbf{I}_n)^{-1} [\boldsymbol{\eta}(j) - \mathbf{X}\boldsymbol{\beta}(j)] \end{aligned}$$

con  $j = 1, \dots, r$ , y donde esta estimación de los parámetros corresponden a los estimadores de máxima verosimilitud para la función de densidad normal  $f(\boldsymbol{\eta}(j) | \mathbf{X}; \boldsymbol{\beta}, \sigma^2, \vartheta, \tau_R^2)$ .

Los valores de  $\boldsymbol{\beta}$  y  $\sigma^2$  que maximizan  $L_r(\boldsymbol{\beta}, \boldsymbol{\theta})$  para un valor fijo tanto de  $\vartheta$  como de  $\tau_R^2$ ;  $\hat{\boldsymbol{\beta}}(\vartheta, \tau_R^2)$  y  $\hat{\sigma}^2(\vartheta, \tau_R^2)$  están conectados dentro de  $L_r$ , y se obtiene  $\tilde{L}_r(\vartheta, \tau_R^2) = L_r(\hat{\boldsymbol{\beta}}(\vartheta, \tau_R^2), \hat{\sigma}^2(\vartheta, \tau_R^2), \vartheta, \tau_R^2)$ . Esta función es maximizada con respecto a  $\vartheta$  y  $\tau_R^2$  para una función de correlación dada utilizando optimización numérica. Los parámetros  $\vartheta$  y  $\tau_R^2$  entran en  $\tilde{L}_r$  vía la matriz  $R(\vartheta) + \tau_R^2 \mathbf{I}_n$  y debido a que se necesita la invertibilidad de esta matriz, la maximización podría ser relativamente lenta. La maximización puede ser también sensible a los valores iniciales en este proceso porque la aproximación de  $L_r$  puede ser multimodal. Por ello, se debe investigar cuidadosamente el resultado considerando una variedad de valores iniciales.

Por otra parte, cuando el interés es investigar cuáles funciones de enlace son apropiadas se debe integrar con respecto a  $\boldsymbol{\mu}_s = (\mu(\mathbf{s}_1), \dots, \mu(\mathbf{s}_n))^t$ , donde  $\mu(\mathbf{s}_i) = m_i g_\nu^{-1}(\mathbf{x}(\mathbf{s}_i), \delta(\mathbf{s}_i))$ ,  $i = 1, \dots, n$ . El determinante del Jacobiano para esta transformación es

$$J_\nu(\boldsymbol{\mu}_s) = \prod_{i=1}^n \frac{g'_\nu(\mu(\mathbf{s}_i)/m_i)}{m_i}$$

Suponiendo que la función de enlace satisface  $g_\nu(\mathbb{R}) = \mathbb{R}$  para todo  $\nu$  de interés y definiendo

$$f_\nu(\boldsymbol{\mu}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\theta}) = J_\nu(\boldsymbol{\mu}_s) f(g_\nu^{-1}(\boldsymbol{\mu}_s) \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\theta})$$

se encuentra que

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \nu) &= \int_{\mathbb{R}^n} f(\mathbf{y}_s \mid \boldsymbol{\mu}_s) f_\nu(\boldsymbol{\mu}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\theta}) d\boldsymbol{\mu}_s \\ &\propto \int_{\mathbb{R}^n} \frac{f(\mathbf{y}_s \mid \boldsymbol{\mu}_s) f_\nu(\boldsymbol{\mu}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\theta})}{f(\mathbf{y}_s \mid \boldsymbol{\mu}_s) \tilde{f}_{\nu_0}(\boldsymbol{\mu}_s)} \tilde{f}_{\nu_0}(\boldsymbol{\mu}_s \mid \mathbf{y}_s) d\boldsymbol{\mu}_s \\ &= \tilde{\mathbb{E}} \left[ \frac{f_\nu(\boldsymbol{\mu}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\theta})}{\tilde{f}_{\nu_0}(\boldsymbol{\mu}_s)} \mid \mathbf{y}_s \right] \end{aligned} \quad (3.17)$$

donde  $\tilde{f}_{\nu_0}(\boldsymbol{\mu}_s) = J_{\nu_0}(\boldsymbol{\mu}_s) \tilde{f}(g_{\nu_0}^{-1}(\boldsymbol{\mu}_s))$ ,  $\tilde{f}_{\nu_0}(\boldsymbol{\mu}_s \mid \mathbf{y}_s) \propto f(\mathbf{y}_s \mid \boldsymbol{\mu}_s) \tilde{f}_{\nu_0}(\boldsymbol{\mu}_s)$  es la función de densidad condicional y  $\tilde{\mathbb{E}}(\cdot \mid \mathbf{y}_s)$  denota la esperanza con respecto a  $\tilde{f}_{\nu_0}(\cdot \mid \mathbf{y}_s)$ . El MLE se puede calcular maximizando la aproximación de Monte Carlo a (3.17),

$$L_r(\boldsymbol{\beta}, \boldsymbol{\theta}, \nu) = \frac{1}{r} \sum_{j=1}^r \frac{f_\nu(\boldsymbol{\mu}_s(j) \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \boldsymbol{\theta})}{\tilde{f}_{\nu_0}(\boldsymbol{\mu}_s(j))} \quad (3.18)$$

donde los  $\boldsymbol{\mu}_s(j)$ 's ( $j = 1, \dots, r$ ) son muestreados por MCMC a partir de la función de distribución  $\tilde{f}(\cdot \mid \mathbf{y}_s)$ . Si  $g_\nu(\mathbb{R}) \neq \mathbb{R}$  entonces la ecuación (3.17) sigue siendo válida.

Por último, se puede seleccionar un valor inicial para  $\boldsymbol{\beta}$  ajustando un DBSGLMM con efectos aleatorios independientes e idénticamente distribuidos. A partir de las estimaciones resultantes de los efectos aleatorios, se puede calcular el variograma empírico a través de

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (\hat{\delta}(\mathbf{s}_i) - \hat{\delta}(\mathbf{s}_j))^2, \quad h > 0 \quad (3.19)$$

donde  $N(h) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| = h\}$  y  $|N(h)|$  es el número de parejas distintas  $N(h)$ . Luego, se gráfica este variograma empírico, lo que puede ayudar a hacerse una idea de la forma paramétrica del variograma y a elegir los valores iniciales de los parámetros asociados al variograma.

Otro método que se puede utilizar para estimar los parámetros involucrados en el modelo (3.11) es la MLE utilizando la versión del MCEMG, la cual se presenta en la siguiente subsección.

### 3.2.3 Versión Monte Carlo del algoritmo gradiente EM para la MLE de los parámetros del DBSGLMM

En este caso, al utilizar la verosimilitud condicional dada las covariables presentada en la ecuación (3.12) y tomando ln en la ecuación (3.13), se obtiene

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\theta}, \nu) &= \ln \int_{\mathbb{R}^n} \prod_{i=1}^n f\{y(\mathbf{s}_i); m_i g_\nu^{-1}[\mathbf{x}(\mathbf{s}_i), \boldsymbol{\delta}(\mathbf{s}_i); \boldsymbol{\beta}]\} f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta}) d\boldsymbol{\delta}(\mathbf{s}_1) \cdots d\boldsymbol{\delta}(\mathbf{s}_n) \\ &= \ln L(\boldsymbol{\beta}, \boldsymbol{\theta}, \nu) = \ln \int_{\mathbb{R}^n} f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \nu) f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta}) d\boldsymbol{\delta} \end{aligned} \quad (3.20)$$

donde  $f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}, \nu)$  y  $f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta})$  fueron definidas en la ecuación (3.13).

Aquí, el problema de la alta dimensión de la integral para la estimación de los parámetros del DBSGLMM se resuelve utilizando el algoritmo de esperanza-maximización (expectation-maximization, EM) y procedimientos de aproximación. El algoritmo EM ha sido un procedimiento estándar para la estimación en GLMM desde el trabajo de Dempster et al. (1977). La utilidad del algoritmo EM en un entorno espacial reside en el tratamiento como datos faltantes, de los términos espaciales aleatorios no observados espaciales y en la imputación de la información faltante basada en los datos observados, con el objetivo de maximizar la probabilidad marginal de los datos observados. En concreto, si el efecto aleatorio  $\mathbf{z}_s$  o  $\boldsymbol{\delta}$  es observado y si se toma el caso donde  $\nu$  es fija, entonces  $\nu$  se suprime y se pueden expresar los datos completos como  $(\mathbf{y}_s, \boldsymbol{\delta})$ , cuya log-verosimilitud conjunta es

$$l_1 = l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \ln f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}) + \ln f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta}) \quad (3.21)$$

Sin embargo, debido a que  $\boldsymbol{\delta}$  es no observable, el calculo directo de (3.21) no es posible. Por consiguiente, es procedente el algoritmo iterativo EM mediante la maximización de la esperanza condicional de la log verosimilitud de datos completos,  $E(l_1 | \mathbf{y}_s)$ , en cada iteración (paso M), donde se toma la esperanza bajo el actual valor (paso E). Algunos algoritmos han sido desarrollados para acelerar la convergencia del EM, uno de los cuales es el algoritmo del gradiente EM (EM gradient, EMG) que sustituye algoritmos de un paso como el Newton-Raphson para el paso M (Lange (1995a, 1995b)), el cual no es similar al presentado en la subsección 3.2.2.

Luego, la primera y segunda derivadas parciales de (3.21) con respecto a  $\boldsymbol{\beta}$  y  $\boldsymbol{\theta}$  son, respectivamente,

$$\frac{\partial l_1}{\partial \boldsymbol{\beta}} = \frac{\partial \ln f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \quad \frac{\partial l_1}{\partial \boldsymbol{\theta}} = \frac{\partial \ln f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (3.22)$$

$$\frac{\partial^2 l_1}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} = \frac{\partial^2 \ln f(\mathbf{y}_s | \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \quad \frac{\partial^2 l_1}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} = \frac{\partial^2 \ln f(\boldsymbol{\delta} | \mathbf{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} \quad (3.23)$$

La matriz de información de Fisher está dada por las siguientes expresiones en cada caso, respectivamente,

$$I(\boldsymbol{\beta}) = -E \left( \frac{\partial^2 \ln l_1}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \mid \mathbf{y}_s \right) \quad I(\boldsymbol{\theta}) = -E \left( \frac{\partial^2 \ln l_1}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} \mid \mathbf{y}_s \right)$$

y el valor esperado de (3.22) está dado por

$$S(\boldsymbol{\beta}) = E \left( \frac{\partial \ln l_1}{\partial \boldsymbol{\beta}} \mid \mathbf{y}_s \right) \quad S(\boldsymbol{\theta}) = E \left( \frac{\partial \ln l_1}{\partial \boldsymbol{\theta}} \mid \mathbf{y}_s \right)$$

El algoritmo EMG actualiza las estimaciones mediante las siguientes expresiones

$$\begin{aligned} \boldsymbol{\beta}^{(m+1)} = & \boldsymbol{\beta}^{(m)} - \left[ E \left\{ \frac{\partial^2 \ln f(\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \mid \mathbf{y}_s \right\} \right]^{-1} \\ & \times E \left\{ \frac{\partial \ln f(\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta}} \mid \mathbf{y}_s \right\} \end{aligned} \quad (3.24)$$

$$\begin{aligned} \boldsymbol{\theta}^{(m+1)} = & \boldsymbol{\theta}^{(m)} - \left[ E \left\{ \frac{\partial^2 \ln f(\boldsymbol{\delta} \mid \mathbf{X}, \boldsymbol{\theta}^{(m)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} \mid \mathbf{y}_s \right\} \right]^{-1} \\ & \times E \left\{ \frac{\partial \ln f(\boldsymbol{\delta} \mid \mathbf{X}, \boldsymbol{\theta}^{(m)})}{\partial \boldsymbol{\theta}} \mid \mathbf{y}_s \right\} \end{aligned} \quad (3.25)$$

donde todas las esperanzas condicionales son evaluadas bajo los valores actuales de los parámetros  $\boldsymbol{\beta}^{(m)}$  y  $\boldsymbol{\theta}^{(m)}$ . Este procedimiento iterativo continúa hasta lograr la convergencia de los parámetros estimados. Zhang (2002) encontró estos parámetros reduciendo a la mitad el tamaño de los pasos, utilizando una técnica comúnmente empleada en la práctica (Zimmerman & Zimmerman 1991, Breslow & Clayton 1993).

Si la distribución condicional de  $\mathbf{y}_s$  dado  $\boldsymbol{\delta}$  pertenece a la familia exponencial y el enlace es canónico, se pueden calcular las derivadas en (3.24) en forma cerrada (McCullagh & Nelder 1989). En particular, si  $\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta}$  pertenece a la familia binomial o Poisson con función de enlace canónico, se encuentra que

$$\begin{aligned} \frac{\partial \ln f(\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \mathbf{X}^t (\mathbf{y}_s - E(\mathbf{y}_s \mid \mathbf{X}; \boldsymbol{\delta})) \\ \frac{\partial^2 \ln f(\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} &= -\mathbf{X}^t \mathbf{V} \mathbf{X} \end{aligned}$$

donde  $\mathbf{V} = \text{Var}(\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta})$  y esta es la matriz diagonal de la matriz de varianzas condicional de  $\mathbf{y}_s$  dados  $\mathbf{X}$  y  $\boldsymbol{\delta}$ . También las derivadas en (3.25) se pueden expresar en forma cerrada ya que  $\boldsymbol{\delta}$  es una normal multivariante (ver Mardia &

Marshall (1984) o los resultados relevantes de la teoría de matrices presentados en Graybill (1983, Capítulo 10)). De esta forma,

$$\begin{aligned}\frac{\partial \ln f(\boldsymbol{\delta} \mid \mathbf{X}; \boldsymbol{\theta})}{\partial \theta_i} &= -\frac{1}{2} \text{tr}(\mathbf{V}^{-1} \mathbf{V}_i) + \frac{1}{2} \boldsymbol{\delta}^t (\mathbf{V}^{-1} \mathbf{V}_i \mathbf{V}^{-1}) \boldsymbol{\delta} \\ \frac{\partial^2 \ln f(\boldsymbol{\delta} \mid \mathbf{X}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} &= -\frac{1}{2} \text{tr}(\mathbf{V}^{-1} \mathbf{V}_{ij} - \mathbf{V}^{-1} \mathbf{V}_i \mathbf{V}^{-1} \mathbf{V}_j) - \frac{1}{2} \boldsymbol{\delta}^t \mathbf{V}^{ij} \boldsymbol{\delta}\end{aligned}$$

donde  $\mathbf{V}^{(-1)}$  es la inversa de la matriz de covarianzas  $\mathbf{V}(\boldsymbol{\theta})$  de  $\boldsymbol{\delta}$  y

$$\begin{aligned}\mathbf{V}_i &= \frac{\partial \mathbf{V}}{\partial \theta_i}, & \mathbf{V}_{ij} &= \frac{\partial^2 \mathbf{V}}{\partial \theta_i \partial \theta_j}, \\ \mathbf{V}^{ij} &= \frac{\partial^2 \mathbf{V}^{-1}}{\partial \theta_i \partial \theta_j} = \mathbf{V}^{-1} (\mathbf{V}_i \mathbf{V}^{-1} \mathbf{V}_j + \mathbf{V}_j \mathbf{V}^{-1} \mathbf{V}_i - \mathbf{V}_{ij}) \mathbf{V}^{-1}\end{aligned}$$

Las esperanzas condicionales en (3.24) y (3.25) no se pueden calcular en forma cerrada, pero se pueden aproximar utilizando muestreos de Monte Carlo  $\boldsymbol{\delta}^{(1)}, \dots, \boldsymbol{\delta}^{(r)}$  a partir del algoritmo de Metropolis-Hastings bajo las estimaciones actuales  $\boldsymbol{\beta}^{(m)}$  y  $\boldsymbol{\theta}^{(m)}$ . Por ejemplo,

$$\begin{aligned}-\mathbf{E} \left\{ \frac{\partial^2 \ln f(\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta}; \boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \mid \mathbf{y}_s \right\} &= \mathbf{X}^t \mathbf{E} \{ \text{Var}(\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta}) \mid \mathbf{y}_s \} \mathbf{X} \\ &\approx \frac{1}{r} \sum_{j=1}^r \mathbf{X}^t \text{Var}(\mathbf{y}_s \mid \mathbf{X}, \boldsymbol{\delta}^{(j)}) \mathbf{X}\end{aligned}$$

Incorporando esta técnica de Monte Carlo en el algoritmo EMG se obtiene el algoritmo MCEMG. Al igual que en el método presentado en la subsección 3.2.2, un valor inicial para  $\boldsymbol{\beta}$  se puede escoger haciendo un primer ajuste de un DBSGLMM con efectos aleatorios bajo el supuesto de i.i.d. A partir de las estimaciones resultantes de los efectos aleatorios, se puede construir el variograma empírico a través de la ecuación (3.19) con la finalidad de seleccionar los parámetros iniciales asociados al variograma. Una vez estimados los vectores de parámetros  $\boldsymbol{\beta}$  y  $\boldsymbol{\theta}$ , se estima  $\nu$ , para lo cual se utiliza la misma estrategia utilizada en la subsección 3.2.2.

Cuando se han estimado el vector de parámetros de tendencia  $\boldsymbol{\beta}$  y el vector de parámetros de correlación espacial  $\boldsymbol{\theta}$ , se está preparado para discutir las medidas espaciales de bondad de ajuste.

### 3.2.4 Medidas de bondad de ajuste

Después de ajustar el DBSGLMM, es importante llevar a cabo un análisis de diagnóstico para verificar la bondad de ajuste del modelo estimado. Una

medida global de la variabilidad explicada se obtiene calculando el pseudo  $R_k^2$  definido como

$$R_k^2 = \frac{l(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}{l(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})}, \quad 0 \leq R_k^2 \leq 1$$

donde  $l(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$  es la función de log-verosimilitud para el modelo saturado evaluado en  $\tilde{\boldsymbol{\beta}}$  y  $\tilde{\boldsymbol{\theta}}$ , y  $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$  es la función log-verosimilitud para el modelo de interés. Observe que  $l(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$  será más grande que cualquier otra función de verosimilitud para las observaciones, asumiendo la misma función de distribución y función de enlace.

El buen ajuste del DBSGLMM se puede determinar por medio de qué tan diferente es el modelo ajustado del modelo saturado, el cual contiene tantos parámetros como observaciones tiene el modelo. Para esto, sea

$$D(\mathbf{y}_s, \boldsymbol{\beta}, \boldsymbol{\theta}) = 2 [l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})]$$

Una aproximación a esta cantidad esta dada por

$$D(\mathbf{y}_s, \boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^n (r_D(\mathbf{s}_i))^2 \quad (3.26)$$

que se conoce como la *deviance*, y sea

$$r_{D_i} = r_D(\mathbf{s}_i) = \text{sign}[y(\mathbf{s}_i) - \hat{\mu}(\mathbf{s}_i)] \left\{ 2 [l(y(\mathbf{s}_i), \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - l(y(\mathbf{s}_i), \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})] \right\}^{1/2}$$

donde  $l(y(\mathbf{s}_i), \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$  es la máxima log-verosimilitud para el modelo saturado asociada a la  $i$ -ésima observación y  $l(y(\mathbf{s}_i), \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$  es el valor máximo de la función de log-verosimilitud para el modelo de interés asociado a la  $i$ -ésima observación.  $r_D(\mathbf{s}_i)$  es la  $i$ -ésima *deviance residual* ya que una observación con un valor absoluto grande de  $r_D(\mathbf{s}_i)$  se puede ver como un dato atípico. Tal como se esperaba, la log-verosimilitud asociada con el modelo saturado debe ser mayor que la de un modelo con  $k < n$  parámetros.

Como se sabe, el análisis de residuales tiene como objetivo identificar observaciones atípicas y/o mala especificación del modelo. Este análisis se puede basar en los residuales ordinarios o en los residuales de deviance. Como es sabido, los residuales son medidas de concordancia entre los datos y el modelo ajustado; la mayoría de los residuales se basan en la diferencia entre la respuesta observada y la media condicional ajustada. Otra medida que mide esta discrepancia y que son ampliamente utilizados, son los residuales de Pearson dados por la siguiente expresión

$$r_{P_i} = r_P(\mathbf{s}_i) = \frac{y(\mathbf{s}_i) - \hat{\mu}(\mathbf{s}_i)}{\sqrt{\text{Var}(\hat{\mu}(\mathbf{s}_i))}}$$

donde  $\text{Var}(\hat{\mu}(\mathbf{s}_i))$  es la varianza de  $\hat{\mu}(\mathbf{s}_i)$ .

Para evitar el problema que tiene un pseudo  $R_k^2 \simeq 1$  cuando el rango de  $\mathbf{X}$  es  $n - 1$ , es necesario considerar solamente los vectores propios más correlacionados de  $\mathbf{B}_v$ , dados por  $\mathbf{1}, \mathbf{X}_1, \dots, \mathbf{X}_{n-1}$ , con la variable regionalizada  $\mathbf{y}_s$ , es decir, las coordenadas principales significativamente más correlacionadas con  $\mathbf{y}_s$ .

### 3.2.5 Selección de las coordenadas principales para el DBSGLMM reducido

Inicialmente, se pueden elegir  $k$  variables explicativas, una buena aproximación para seleccionar las columnas de  $\mathbf{X}$  consiste en clasificarlas según su coeficiente de pseudo correlación con respecto a  $\mathbf{y}_s$ , es decir,

$$R_1^2(\mathbf{X}_1) > \dots > R_1^2(\mathbf{X}_k) > \dots > R_1^2(\mathbf{X}_{n-1})$$

donde  $R_1^2(\mathbf{X}_j)$  es el coeficiente de pseudo correlación entre la  $j$ -ésima coordenada principal ( $\mathbf{X}_j$ ,  $j = 1, \dots, k, \dots, n - 1$ ) y  $\mathbf{y}_s$ . Estos coeficientes de pseudo correlación se obtienen dejando la misma función de enlace en  $g$  para los diferentes modelos ajustados. Con este proceso, las variables menos correlacionadas con  $\mathbf{y}_s$  en la matriz de coordenadas principales  $\mathbf{X}$  se eliminan, es decir, no se consideran  $n - k - 1$  coordenadas principales en el modelo final.

Un procedimiento similar se obtiene utilizando (3.26), pero con  $r_i(\mathbf{X}_j)$ , que es el residuo obtenido a partir de la inclusión en el modelo de la coordenada principal  $\mathbf{X}_j$ ,  $j = 1, \dots, k, \dots, n - 1$ . De esta manera, las pseudo deviances se ordenan de menor a mayor; las primeras  $k$  pseudo deviances son

$$D(\mathbf{y}_s; \mathbf{X}_1) < D(\mathbf{y}_s; \mathbf{X}_2) < \dots < D(\mathbf{y}_s; \mathbf{X}_k)$$

Otra opción para la selección de las coordenadas principales, se hace realizando una gráfica que represente los puntos  $(j, 1 - c(j))$   $j = 1, \dots, k, k+1, \dots, n-1$ , y luego, se determinan los puntos con un descenso significativo en la falta de predictibilidad, dada por  $1 - c(j)$  (Cuadras 1989, Cuadras et al. 1996). La predictibilidad  $c(j)$  está dada por

$$c(0) = 0, \quad c(j) = \frac{\sum_{l=1}^j R_1^2(\mathbf{X}_l) \lambda_l}{\sum_{l=1}^{n-1} R_1^2(\mathbf{X}_l) \lambda_l}, \quad j = 1, \dots, k, \dots, n - 1$$

donde  $\lambda_l$  es el  $l$ -ésimo valor propio asociado a  $\mathbf{X}_l$ ,  $l = 1, \dots, k, \dots, n - 1$  (ver mayores detalles en Cuadras et al. (1996)). Por lo tanto, se eliminan las coordenadas principales  $\mathbf{X}_{k+1}, \dots, \mathbf{X}_{n-1}$ .

Una vez elegidas las coordenadas principales mas significativas ( $k$ ), se discuten las técnicas espaciales para predecir el valor de un campo aleatorio en una localización dada de observaciones cercanas.

### 3.3 Predicción espacial de un nuevo individuo

Las coordenadas  $\mathbf{x}_{(k)}(\mathbf{s}_0)$  se obtienen asumiendo que las observaciones de las variables explicativas mixtas son observadas para un nuevo individuo, esto es,  $\mathbf{v}(\mathbf{s}_0) = (v_1(\mathbf{s}_0), \dots, v_p(\mathbf{s}_0))^t$  es conocido. Entonces, se deben calcular las distancias entre el nuevo individuo y cada uno de los individuos involucrados en el modelo propuesto en (3.2), es decir,  $d_{0i} = d(v(\mathbf{s}_0), v(\mathbf{s}_i))$ ,  $i = 1, \dots, n$ . A partir de estas distancias, se puede realizar una predicción utilizando un resultado propuesto por Gower (1968) y Cuadras & Arenas (1990, Section 3.3), que relaciona al vector  $\mathbf{d}_0 = (d_{01}^2, \dots, d_{0n}^2)^t$  de distancias al cuadrado con el vector  $\mathbf{x}_{(k)}(\mathbf{s}_0) = (x_1(\mathbf{s}_0), \dots, x_k(\mathbf{s}_0))^t$  de coordenadas principales asociadas al nuevo individuo, como se presenta a continuación

$$d_{0i}^2 = (\mathbf{x}_{(k)}(\mathbf{s}_0) - \mathbf{x}_{(k)}(\mathbf{s}_i))^t (\mathbf{x}_{(k)}(\mathbf{s}_0) - \mathbf{x}_{(k)}(\mathbf{s}_i))$$

donde  $\mathbf{x}_{(k)}(\mathbf{s}_i) = (x_1(\mathbf{s}_i), \dots, x_k(\mathbf{s}_i))^t$  con  $i = 1, \dots, n$ . Entonces, se tiene que

$$\mathbf{x}_{(k)}(\mathbf{s}_0) = \frac{1}{2} \Lambda^{-1} \mathbf{X}_{(k)}^t (\mathbf{b} - \mathbf{d}_0) \quad (3.27)$$

donde  $\mathbf{b} = (b_{11}, \dots, b_{nn})^t$  es un vector conformado por los elementos de la diagonal de  $\mathbf{B}_v$ , con  $b_{ii} = \mathbf{x}_{(k)}^t(\mathbf{s}_i) \mathbf{x}_{(k)}(\mathbf{s}_i)$ ,  $i = 1, \dots, n$ .

#### 3.3.1 Predicción espacial

En este caso se utiliza la técnica de kriging que es una técnica mediante la cual se puede interpolar el valor  $y(\mathbf{s}_0)$  de un campo aleatorio  $Y(\mathbf{s}_0)$  en una localización predefinida,  $\mathbf{s}_0$ , a partir de las observaciones  $y_i = y(\mathbf{s}_i)$ ,  $i = 1, \dots, n$ .

#### Interpolación de efectos aleatorios sobre una área espacial continua

En esta subsección el enfoque se centra en la interpolación de efectos aleatorios sobre un área espacial continua donde las observaciones son no gaussianas. Sea  $\mathbf{s}_0 \neq \mathbf{s}_i$  para todo  $i = 1, 2, \dots, n$ . Se sabe que la predicción de mínimo error cuadrático medio (minimum mean square error, MMSE) para el efecto aleatorio  $\delta_0 = \delta(\mathbf{s}_0)$  en una localización  $\mathbf{s}_0$  es la esperanza condicional  $E(\delta_0 | \mathbf{y}_s)$ . Denótese la función de densidad de probabilidad conjunta de  $(\delta_0, \boldsymbol{\delta}, \mathbf{y}_s)$  por  $f(\delta_0, \boldsymbol{\delta}, \mathbf{y}_s)$ ,  $\delta_0 \in \mathbb{R}$ ,  $\boldsymbol{\delta}, \mathbf{y}_s \in \mathbb{R}^n$ . Debido a la formulación del modelo, la



función de distribución condicional de  $\mathbf{y}_s$  dado  $\{\delta(\mathbf{s}_0), \mathbf{s}_0 \in \mathbb{R}^d\}$  es la función de distribución de  $\mathbf{y}_s$  dado  $\boldsymbol{\delta} = (\delta(\mathbf{s}_1), \delta(\mathbf{s}_2), \dots, \delta(\mathbf{s}_n))^t$ . Esto implica que  $f(\mathbf{y}_s | \delta_0, \boldsymbol{\delta}) = f(\mathbf{y}_s | \boldsymbol{\delta})$ , por lo tanto,

$$\begin{aligned} f(\delta_0, \boldsymbol{\delta}, \mathbf{y}_s) &= f(\mathbf{y}_s | \delta_0, \boldsymbol{\delta})f(\delta_0, \boldsymbol{\delta}) = f(\mathbf{y}_s | \boldsymbol{\delta})f(\delta_0, \boldsymbol{\delta}) \\ &= \frac{f(\mathbf{y}_s, \boldsymbol{\delta})}{f(\boldsymbol{\delta})}f(\delta_0, \boldsymbol{\delta}) = f(\mathbf{y}_s, \boldsymbol{\delta})f(\delta_0 | \boldsymbol{\delta}) \end{aligned}$$

Dividiendo a ambos lados la anterior expresión por  $f(\mathbf{y}_s, \boldsymbol{\delta})$ , se obtiene

$$f(\delta_0 | \boldsymbol{\delta}, \mathbf{y}_s) = f(\delta_0 | \boldsymbol{\delta})$$

y en consecuencia

$$E(\delta_0 | \boldsymbol{\delta}, \mathbf{y}_s) = E(\delta_0 | \boldsymbol{\delta}) = \sum_{i=1}^n c_i \delta(\mathbf{s}_i)$$

para alguna constante apropiada  $c_i$ . Entonces,

$$\begin{aligned} E(\delta_0 | \boldsymbol{\delta}) &= E\{E(\delta_0 | \boldsymbol{\delta}, \mathbf{y}_s) | \mathbf{y}_s\} \\ &= E\{E(\delta_0 | \boldsymbol{\delta}) | \mathbf{y}_s\} = E\left\{\sum_{i=1}^n c_i \delta(\mathbf{s}_i) | \mathbf{y}_s\right\} \\ &= \sum_{i=1}^n c_i E(\delta(\mathbf{s}_i) | \mathbf{y}_s) \end{aligned}$$

donde  $E(\delta(\mathbf{s}_i) | \mathbf{y}_s)$  es la estimación de MMSE del efecto aleatorio  $\delta(\mathbf{s}_i)$  y los coeficientes  $c_i$  son tales que  $\sum_{i=1}^n c_i \delta(\mathbf{s}_i)$  se iguala a  $E(\delta_0 | \boldsymbol{\delta})$ , la predicción MMSE de  $\delta_0$  dado  $\boldsymbol{\delta}$ . En otras palabras, estos coeficientes son los mismos de la predicción MMSE de  $\delta(\mathbf{s}_0)$  dado  $\boldsymbol{\delta}$ .

Por lo tanto, una vez que se obtienen las estimaciones de los efectos aleatorios en los sitios muestreados, se puede obtener la predicción MMSE para el efecto aleatorio en cualquier sitio no muestreado como si los efectos aleatorios fueran observables en los sitios muestreados. La predicción MMSE es particularmente apropiada para el DBSGLMM debido a la propiedad lineal anterior que es análoga a la del kriging lineal.

Por otro lado, en el clásico kriging se utiliza la combinación lineal de los valores observados para aproximarse a una nueva localización, con pesos grandes asignados a las localizaciones más cercanas. Sin embargo por ejemplo, en muchas situaciones para datos no normales, el supuesto de linealidad es demasiado estricto y no es plausible. Por lo tanto, se considera un predictor óptimo que minimiza la siguiente función condicional del error cuadrático medio

$$E[\{p(y; \mathbf{s}_0) - y(\mathbf{s}_0)\}^2 | \mathbf{y}_s] \quad (3.28)$$

donde  $p(y; \mathbf{s}_0)$  es el predictor en  $\mathbf{s}_0$  basado en la  $\mathbf{y}_s$  observada. Por lo que se obtiene,

$$\begin{aligned} E[\{p(y; \mathbf{s}_0) - y(\mathbf{s}_0)\}^2 | \mathbf{y}_s] &= \text{Var}[\{p(y; \mathbf{s}_0) - y(\mathbf{s}_0) | \mathbf{y}_s\}] + [E\{p(y; \mathbf{s}_0) - y(\mathbf{s}_0) | \mathbf{y}_s\}]^2 \\ &= \text{Var}\{y(\mathbf{s}_0) | \mathbf{y}_s\} + [p(y; \mathbf{s}_0) - E\{y(\mathbf{s}_0) | \mathbf{y}_s\}]^2 \end{aligned}$$

Es obvio por consiguiente que el predictor óptimo está dado por

$$\hat{y}(\mathbf{s}_0) = E\{y(\mathbf{s}_0) | \mathbf{y}_s\}$$

Entonces, el mejor predictor para los efectos aleatorios que MMSE condicional (3.28) es  $E\{y(\mathbf{s}_0) | \mathbf{y}_s\}$ , el cual no es necesariamente lineal en  $\mathbf{y}_s$ .

Cuando  $y(\cdot)$  es un proceso gaussiano, el predictor óptimo  $\hat{y}(\mathbf{s}_0)$  coincide con el obtenido bajo kriging clásico y está dado por

$$\hat{y}_K(\mathbf{s}_0) = \mathbf{x}^t(\mathbf{s}_0)\hat{\boldsymbol{\beta}} + \mathbf{c}^t \boldsymbol{\Sigma}_\delta^{-1} (\mathbf{y}_s - \mathbf{X}\hat{\boldsymbol{\beta}})$$

donde  $\mathbf{x}(\mathbf{s}_0) = (1, x_{(k)}^t(\mathbf{s}_0))^t = (1, x_1(\mathbf{s}_0), \dots, x_k(\mathbf{s}_0))^t$ ,  $\mathbf{c} = (\mathbb{C}(\mathbf{s}_0 - \mathbf{s}_1; \boldsymbol{\theta}), \dots, \mathbb{C}(\mathbf{s}_0 - \mathbf{s}_n; \boldsymbol{\theta}))^t$  y  $\boldsymbol{\Sigma}_\delta = [\mathbb{C}(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta})]_{n \times n}$ , con  $\mathbb{C}(\cdot)$  la función de covarianza la cual es especificada fácilmente del semivariograma.

La matriz de covarianza para la predicción está dada por

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{s_0} - \mathbf{c}^t \boldsymbol{\Sigma}_\delta^{-1} \mathbf{c}$$

donde  $\boldsymbol{\Sigma}_{s_0}$  es la matriz de covarianza entre los nuevos individuos,  $\mathbf{s}_0$ .

### Interpolación de efectos aleatorios utilizando predictores lineales insesgados

Sea  $\boldsymbol{\eta}^0 = (\eta(\mathbf{s}_{n+1}), \dots, \eta(\mathbf{s}_{n+n'}))^t$  la predicción funcional y sea  $f(\boldsymbol{\eta}^0, \boldsymbol{\eta})$  la función de densidad conjunta de  $\boldsymbol{\eta}$  y  $\boldsymbol{\eta}^0$ . Limitando el interés a los predictores lineales pseudo insesgados de la forma

$$\tilde{\boldsymbol{\eta}} = \mathbf{p} + \mathbf{Q}\boldsymbol{\eta} \quad (3.29)$$

para algún vector conformable  $\mathbf{p}$  y una matriz  $\mathbf{Q}$  (McCulloch et al. 2008), la MMSE de la predicción se realiza por medio de

$$E [(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^0)^t \mathbf{A} (\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^0)] = \int \int (\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^0)^t \mathbf{A} (\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^0) f(\boldsymbol{\eta}^0, \boldsymbol{\eta}) d\boldsymbol{\eta} d\boldsymbol{\eta}^0 \quad (3.30)$$

donde  $f(\boldsymbol{\eta}^0, \boldsymbol{\eta})$  es la función de distribución conjunta de  $\boldsymbol{\eta}^0$  y  $\boldsymbol{\eta}$ , y  $\mathbf{A}$  es una matriz simétrica definida positiva.

Utilizando (3.29), la parte izquierda de (3.30) se puede expresar como

$$\begin{aligned} \mathbf{q} &= \mathbb{E} \left[ (\mathbf{p} + \mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0)^t \mathbf{A} (\mathbf{p} + \mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0) \right] \\ &= \mathbf{p}^t \mathbf{A} \mathbf{p} + 2\mathbf{p}^t \mathbf{A} \mathbb{E} (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0) + \mathbb{E} \left[ (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0)^t \mathbf{A} (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0) \right] \end{aligned} \quad (3.31)$$

Derivando parcialmente (3.31) con respecto a  $\mathbf{p}$  e igualando a  $\mathbf{0}$ , se obtiene

$$\begin{aligned} \frac{\partial \mathbf{q}}{\partial \mathbf{p}} &= 2\mathbf{A} [\mathbf{p} + \mathbb{E} (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0)] = \mathbf{0} \\ \mathbf{p} &= -\mathbb{E} (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0) \end{aligned} \quad (3.32)$$

Sustituyendo (3.32) en (3.31) se llega a

$$\begin{aligned} \mathbf{q} &= -\left[ \mathbb{E} (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0) \right]^t \mathbf{A} \mathbb{E} [(\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0)] + \mathbb{E} \left[ (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0)^t \mathbf{A} (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0) \right] \\ &= \text{tr} \{ \mathbf{A} \text{Var} (\mathbf{Q}\boldsymbol{\eta} - \boldsymbol{\eta}^0) \} \\ &= \text{tr} \{ \mathbf{A} [\mathbf{Q}\boldsymbol{\Sigma}_\eta \mathbf{Q}^t + \text{Var} (\boldsymbol{\eta}^0) - \mathbf{Q} \text{Cov} (\boldsymbol{\eta}, \boldsymbol{\eta}^0) - \text{Cov}^t (\boldsymbol{\eta}, \boldsymbol{\eta}^0) \mathbf{Q}^t] \} \end{aligned} \quad (3.33)$$

donde  $\boldsymbol{\Sigma}_\eta = \sigma^2 (\mathbf{R}(\vartheta) + \tau_R^2 \mathbf{I}_n)$ ,  $\text{Var} (\boldsymbol{\eta}^0)$  es la matriz de covarianza de  $\boldsymbol{\eta}^0$  y  $\text{Cov} (\boldsymbol{\eta}, \boldsymbol{\eta}^0)$  es la matriz de covarianzas cruzadas entre  $\boldsymbol{\eta}$  y  $\boldsymbol{\eta}^0$ .

Ahora se desea minimizar (3.33) con respecto a  $\mathbf{Q}$ . Para hacer esto, se ignoran  $\mathbf{A}$  y  $\text{Var} (\boldsymbol{\eta}^0)$  porque no involucran  $\mathbf{Q}$ . Entonces, derivando parcialmente (3.33) con respecto a  $\mathbf{Q}$  e igualando a  $\mathbf{0}$ , se tiene que

$$\begin{aligned} \frac{\partial \mathbf{q}}{\partial \mathbf{Q}} &= 2\mathbf{Q}\boldsymbol{\Sigma}_\eta - 2\text{Cov}^t (\boldsymbol{\eta}, \boldsymbol{\eta}^0) = \mathbf{0} \\ \mathbf{Q} &= \text{Cov}^t (\boldsymbol{\eta}, \boldsymbol{\eta}^0) \boldsymbol{\Sigma}_\eta^{-1} \end{aligned} \quad (3.34)$$

Así, sustituyendo (3.32) y (3.34) en (3.29), el mejor pseudo predictor lineal insesgado (best pseudo linear unbiased predictor, BPLUP) está dado por

$$\begin{aligned} \tilde{\boldsymbol{\eta}} &= \mathbb{E} (\boldsymbol{\eta}^0) + \text{Cov}^t (\boldsymbol{\eta}, \boldsymbol{\eta}^0) \boldsymbol{\Sigma}_\eta^{-1} [\boldsymbol{\eta} - \mathbb{E} (\boldsymbol{\eta})] \\ &= \mathbf{X}^0 \boldsymbol{\beta} + \text{Cov}^t (\boldsymbol{\eta}, \boldsymbol{\eta}^0) \boldsymbol{\Sigma}_\eta^{-1} [\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}] \end{aligned} \quad (3.35)$$

donde  $\mathbf{X}^0$  es una matriz de  $k$  coordenadas principales para los nuevos datos espaciales  $n'$  que incluyen un vector de unos, es decir,  $\mathbf{1}_{n'}$  de tamaño  $n' \times 1$ .

Tomando valor esperado en (3.35), se llega a

$$\begin{aligned} \mathbb{E}(\tilde{\boldsymbol{\eta}} \mid \mathbf{y}_s) &= \mathbf{X}^0 \boldsymbol{\beta} + \text{Cov}^t (\boldsymbol{\eta}, \boldsymbol{\eta}^0) \boldsymbol{\Sigma}_\eta^{-1} [\mathbb{E}(\boldsymbol{\eta} \mid \mathbf{y}_s) - \mathbf{X}\boldsymbol{\beta}] \\ &\approx \mathbf{X}^0 \boldsymbol{\beta} + \text{Cov}^t (\boldsymbol{\eta}, \boldsymbol{\eta}^0) \boldsymbol{\Sigma}_\eta^{-1} [\tilde{\mathbb{E}}_r(\boldsymbol{\eta} \mid \mathbf{y}_s) - \mathbf{X}\boldsymbol{\beta}] \end{aligned}$$

donde  $\tilde{\mathbb{E}}_r$  es el vector de medias empíricas basadas en las muestras  $\boldsymbol{\eta}(1), \dots, \boldsymbol{\eta}(r)$ .

La matriz de covarianza para la predicción presentada en (3.35) está dada por

$$\begin{aligned}\text{Var}(\tilde{\boldsymbol{\eta}} \mid \mathbf{y}_s) &= \boldsymbol{\Sigma}_0 + \text{Cov}^t(\boldsymbol{\eta}, \boldsymbol{\eta}^0) \boldsymbol{\Sigma}_\eta^{-1} [\text{Var}(\boldsymbol{\eta} \mid \mathbf{y}_s)] \boldsymbol{\Sigma}_\eta^{-1} \text{Cov}(\boldsymbol{\eta}, \boldsymbol{\eta}^0) \\ &\approx \boldsymbol{\Sigma}_0 + \text{Cov}^t(\boldsymbol{\eta}, \boldsymbol{\eta}^0) \boldsymbol{\Sigma}_\eta^{-1} [\widetilde{\text{Var}}_r(\boldsymbol{\eta} \mid \mathbf{y}_s)] \boldsymbol{\Sigma}_\eta^{-1} \text{Cov}(\boldsymbol{\eta}, \boldsymbol{\eta}^0)\end{aligned}$$

donde  $\boldsymbol{\Sigma}_0 = \text{Var}(\boldsymbol{\eta}^0) - \text{Cov}^t(\boldsymbol{\eta}, \boldsymbol{\eta}^0) \boldsymbol{\Sigma}_\eta^{-1} \text{Cov}(\boldsymbol{\eta}, \boldsymbol{\eta}^0)$  y  $\widetilde{\text{Var}}_r$  es la matriz de covarianza basada en las muestras  $\boldsymbol{\eta}(1), \dots, \boldsymbol{\eta}(r)$ .

**Ejemplo 3.1.** *Considérese un modelo Dirichlet para respuestas binarias espaciales de tal manera que*

$$y(\mathbf{s}_i) \mid \eta(\mathbf{s}_i) \sim \text{Bernoulli}(\eta(\mathbf{s}_i)), \quad i = 1, \dots, n$$

donde  $\eta(\mathbf{s}_i) = \delta(\mathbf{s}_i)$  y el efecto aleatorio  $\boldsymbol{\eta} = (\eta(\mathbf{s}_1), \dots, \eta(\mathbf{s}_n)) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_n)$  con  $\alpha_i > 0$ ,  $i = 1, \dots, n$ , y  $\eta(\mathbf{s}_n) = 1 - \sum_{l=1}^{n-1} \eta(\mathbf{s}_l)$ .

Es sencillo calcular los momentos de la  $y(\mathbf{s})$  bajo el modelo (3.11). Así se obtiene

$$\begin{aligned}\text{E}(y(\mathbf{s}_i)) &= \text{E}\{\text{E}[y(\mathbf{s}_i) \mid \eta(\mathbf{s}_i)]\} = \text{E}[\eta(\mathbf{s}_i)] = \frac{\alpha_i}{\alpha_0} \\ \text{Var}(y(\mathbf{s}_i)) &= \text{E}\{\text{Var}[y(\mathbf{s}_i) \mid \eta(\mathbf{s}_i)]\} + \text{Var}\{\text{E}[y(\mathbf{s}_i) \mid \eta(\mathbf{s}_i)]\} \\ &= \text{E}\{\eta(\mathbf{s}_i)[1 - \eta(\mathbf{s}_i)]\} + \text{Var}[\eta(\mathbf{s}_i)] \\ &= \text{E}\{[\eta(\mathbf{s}_i)][1 - \eta(\mathbf{s}_i)]\} \\ &= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2} = c_{ii}\end{aligned}$$

donde  $\alpha_0 = \sum_{l=1}^n \alpha_l$ . La matriz de covarianza se puede obtener similarmente como

$$\begin{aligned}\text{Cov}[y(\mathbf{s}_i), y(\mathbf{s}_j)] &= \text{Cov}\{\text{E}[y(\mathbf{s}_i) \mid \eta(\mathbf{s}_i)], \text{E}[y(\mathbf{s}_j) \mid \eta(\mathbf{s}_j)]\} \\ &\quad + \text{E}\{\text{Cov}[y(\mathbf{s}_i), y(\mathbf{s}_j) \mid \eta(\mathbf{s}_i), \eta(\mathbf{s}_j)]\}, \quad i \neq j \\ &= \text{Cov}[\eta(\mathbf{s}_i), \eta(\mathbf{s}_j)] + 0 \\ &= -\frac{\alpha_i \alpha_j}{\alpha_0^2(1 + \alpha_0)} = c_{ij}\end{aligned}$$

De otro lado, se puede calcular la covarianza entre  $y(\mathbf{s}_i)$  y  $\boldsymbol{\eta}(\mathbf{s}_i)$  utilizando

las siguientes expresiones

$$\begin{aligned}
\text{Cov}[y(\mathbf{s}_i), \eta(\mathbf{s}_i)] &= \text{Cov}\{E[y(\mathbf{s}_i) \mid \eta(\mathbf{s}_i)], E[\eta(\mathbf{s}_i) \mid \eta(\mathbf{s}_i)]\} \\
&\quad + E\{\text{Cov}[y(\mathbf{s}_i), \eta(\mathbf{s}_i) \mid \eta(\mathbf{s}_i)]\} \\
&= \text{Var}[\eta(\mathbf{s}_i)] + 0 \\
&= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(1 + \alpha_0)} = \varphi_{ii} \\
\text{Cov}[\eta(\mathbf{s}_i), y(\mathbf{s}_j)] &= \text{Cov}\{E[\eta(\mathbf{s}_i) \mid \eta(\mathbf{s}_i)], E[y(\mathbf{s}_j) \mid \eta(\mathbf{s}_j)]\} \\
&\quad + E\{\text{Cov}[\eta(\mathbf{s}_i), y(\mathbf{s}_j) \mid \eta(\mathbf{s}_i), \eta(\mathbf{s}_j)]\}, \quad i \neq j \\
&= \text{Cov}[\eta(\mathbf{s}_i), \eta(\mathbf{s}_j)] + 0 \\
&= -\frac{\alpha_i \alpha_j}{\alpha_0^2(1 + \alpha_0)} = \varphi_{ij}
\end{aligned}$$

Utilizando (3.35), se obtiene el BPLUP para  $\eta(\mathbf{s}_i)$ ,

$$\tilde{\eta}(\mathbf{s}_i) = \frac{\alpha_i}{\alpha_0} + \boldsymbol{\varphi}_i^t \boldsymbol{\Sigma}_0^{-1} [\boldsymbol{\eta} - E(\boldsymbol{\eta})]$$

donde  $E(\boldsymbol{\eta}) = (\alpha_1/\alpha_0, \dots, \alpha_n/\alpha_0)^t$ ,  $\boldsymbol{\Sigma}_0 = (c_{ij})_{n \times n}$  y  $\boldsymbol{\varphi}_i = (\varphi_{1i}, \dots, \varphi_{ii}, \dots, \varphi_{ni})^t$ .

Como un simple ejemplo tómesese  $n = 2$ , entonces

$$\begin{aligned}
\tilde{\eta}(\mathbf{s}_i) &= \frac{\alpha_i}{\alpha_1 + \alpha_2} + \frac{\alpha_i(\alpha_1 + \alpha_2 - \alpha_i)}{(\alpha_1 + \alpha_2)^2(1 + \alpha_1 + \alpha_2)} \frac{(\alpha_1 + \alpha_2)^2}{\alpha_i(\alpha_1 + \alpha_2 - \alpha_i)} \left( \bar{\eta}_2 - \frac{\alpha_i}{\alpha_1 + \alpha_2} \right) \\
&= \frac{\alpha_i + \bar{\eta}_2}{1 + \alpha_1 + \alpha_2}
\end{aligned}$$

donde  $\bar{\eta}_2 = [\eta(\mathbf{s}_1) + \eta(\mathbf{s}_2)]/2$ .

### 3.4 Relación con el GLMM espacial clásico

El modelo (3.11) depende de la distancia seleccionada,  $d_{ii'}$  ( $i, i' = 1, \dots, n$ ), lo cual es particularmente interesante bajo casos de variables explicativas mixtas y con un comportamiento no lineal en su relación con la variable respuesta. Cuando las variables explicativas son continuas y se utiliza la distancia euclidiana, el DBSGLMM es compatible con el GLMM espacial; esta equivalencia también se mantiene para variables cualitativas cuando se utiliza una distancia que se basa en la disimilaridad. Además, se puede mostrar lo mismo cuando hay una mezcla de variables explicativas continuas, categóricas y binarias. En esta subsección se demuestran estas equivalencias entre el DBSGLMM y el GLMM espacial.

### 3.4.1 Variables continuas

Si todas las variables en (3.3),  $\mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_p)$ , son continuas, la distancia euclidiana está dada por (3.5). Luego, la matriz de distancias  $\mathbf{D}_v = (d_{ii'})$  se obtiene, y

$$\mathbf{A}_C = -\frac{1}{2}\{\text{diag}(\mathbf{V}\mathbf{V}^t)\mathbf{1}^t + \mathbf{1}[\text{diag}(\mathbf{V}\mathbf{V}^t)]^t - 2\mathbf{V}\mathbf{V}^t\}$$

donde  $\text{diag}(\mathbf{V}\mathbf{V}^t)$  es un vector que contiene los términos de la matriz diagonal  $\mathbf{V}\mathbf{V}^t$ . Por lo tanto,

$$\mathbf{B}_C = \mathbf{H}\mathbf{A}_C\mathbf{H} = \mathbf{H}\mathbf{V}\mathbf{V}^t\mathbf{H} = \mathbf{X}\mathbf{X}^t$$

ya que  $\mathbf{H}[\text{diag}(\mathbf{V}\mathbf{V}^t)\mathbf{1}^t]\mathbf{H} = \mathbf{0}$ ,  $\mathbf{H}\mathbf{1}[\text{diag}(\mathbf{V}\mathbf{V}^t)]^t\mathbf{H} = \mathbf{0}$  y donde  $\mathbf{B}_C$  se definió en la Sección 3.2.1. Entonces, el DBSGLMM en (3.11) es un modelo lineal generalizado centrado espacial mixto (3.3), es decir, el DBSGLMM produce las mismas predicciones utilizando  $p$  coordenadas principales que el modelo presentado en (3.3).

Sin embargo, no es necesario considerar una distancia euclidiana  $p$ -dimensional como la presentada en la ecuación (3.5) ya que cualquier otra distancia se podría utilizar, por ejemplo, la distancia absoluta que satisface las mismas propiedades de una distancia euclidiana. Si  $E_l$  ( $k \leq l \leq n-1$ ) es el espacio generado por las columnas de  $\mathbf{X}$ , donde  $\mathbf{X}$  es una solución métrica escalada obtenida a partir de una distancia aplicada a los datos. Entonces tomando  $k > p$ , es decir, las columnas más significativas de  $\mathbf{X}$ , el DBSGLMM supera al GLMM espacial clásico cuando  $(\eta - \hat{\beta}_0\mathbf{1}) \in E_l$ . Observe que esto es siempre cierto para  $l = n-1$ . Para ilustrar lo anterior, suponga  $p = 1$  y observe que aplicando el DBSGLMM, el pseudo  $R_l^2$  se puede escribir como

$$R_l^2 = \sum_{j=1}^l R_1^2(\mathbf{X}_j), \quad 1 \leq l \leq n-1$$

ya que las coordenadas principales  $(\mathbf{X}_1, \dots, \mathbf{X}_{n-1})$  no están correlacionadas. Por otro lado, se puede representar como  $R^2(\boldsymbol{\eta}, \mathbf{V}_1)$  el pseudo- $R^2$  entre el  $\boldsymbol{\eta}$  y la variable  $\mathbf{V}_1$  utilizando GLMM espacial clásico. Por lo tanto, si se toma  $l = k > 1$  (por ejemplo,  $k = 2$ ) entonces se encuentra que

$$R_l^2 = \sum_{j=1}^l R_1^2(\mathbf{X}_j) \geq R^2(\boldsymbol{\eta}, \mathbf{V}_1)$$

Así, el DBSGLMM supera al GLMM espacial clásico.

### 3.4.2 Variables cualitativas

Ahora suponga que todas las variables explicativas  $\mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_p)$  son cualitativas en (3.3), donde ahora cada  $\mathbf{V}_j$  es una variable con  $q_j$  estados,  $j = 1, \dots, p$ . Una medida de semejanza entre los individuos  $i$  y  $i'$  es el número de coincidencias  $m_{ii'}$  para las variables cualitativas involucradas en el modelo. Observe que  $0 \leq m_{ii'} \leq p$ , y cuando las variables explicativas son binarias, los  $m_{ii'}/p$ 's son los coeficientes de coincidencias (ver detalles en Cuadras & Arenas (1990) y Cuadras et al. (1996)).

La distancia entre los individuos  $i$  e  $i'$ , se define como:

$$d_{ii'}^2 = m_{ii} + m_{i'i'} - 2m_{ii'} = 2(p - m_{ii'})$$

donde  $d_{ij}^2$  se convierte en un distancia euclidiana al cuadrado porque  $V_j$  se puede representar por  $q_j$  ( $j = 1, \dots, p$ ) variables binarias codificadas (0, 1).

De una manera similar, como se mostró para variables continuas, haciendo  $\mathbf{A}_Q = -\frac{1}{2}[\mathbf{M}_r + \mathbf{M}_r^t - 2\mathbf{M}]$ , se encuentra que

$$\mathbf{B}_Q = \mathbf{H}\mathbf{A}_Q\mathbf{H} = \mathbf{H}\mathbf{M}\mathbf{H} = \mathbf{H}\mathbf{V}\mathbf{V}^t\mathbf{H}$$

donde todas las filas de  $\mathbf{M}_r$  son iguales,  $\mathbf{M} = (m_{ii'})$ , y  $\mathbf{H}\mathbf{M}_r = \mathbf{M}_r^t\mathbf{H} = \mathbf{0}$ .

Por lo tanto, no hay ventajas sobre el GLMM espacial clásico, excepto que el problema de multicolinealidad se resuelve automáticamente usando distancias (Cuadras et al. 1996). Por consiguiente, el DBSGLMM se reduce al GLMM espacial clásico para variables cualitativas codificando los estados como 0 (ausente) y 1 (presente).

### 3.4.3 Variables mixtas

Ahora suponga que se tiene en (3.3) la matriz  $\mathbf{V} = (\mathbf{V}_c \ \mathbf{V}_q)$ , donde  $\mathbf{V}_c$  y  $\mathbf{V}_q$  son submatrices de variables cuantitativas y cualitativas de  $\mathbf{V}$ , respectivamente. Según (Cuadras et al. 1996), es apropiado utilizar la similitud  $m_{ii'}$  dada por (3.4) (Gower 1971) entre los individuos  $i$  e  $i'$ . Así, el cuadrado de la diferencia entre los individuos  $i$  e  $i'$  es  $d_{ii'}^2 = 1 - m_{ii'}$  y  $\mathbf{D}_v = (d_{ij})$  es una matriz de distancias euclidianas sobre el conjunto de  $n$  individuos (Cuadras et al. 1996). Por consiguiente, el DBSGLMM se demuestra que es equivalente al GLMM espacial clásico considerando las variables cuantitativas como de costumbre y codificando las variables cualitativas como se describe en la subsección anterior 3.4.2.

### 3.4.4 DBSGLMM no lineal

El enfoque basado en distancias es también útil para realizar un GLMM espacial no lineal. Si  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_k)$  son las coordenadas principales, el DBSGLMM no lineal está dado por

$$\eta_i = g(\mu(\mathbf{s}_i)) = \sum_{j=1}^k f_j(x_j(\mathbf{s}_i))\beta_j + z(\mathbf{s}_i, \boldsymbol{\theta}) \quad (3.36)$$

donde  $f_j$  ( $j = 1, \dots, k$ ) es una función no lineal (posiblemente desconocida) en función del parámetro  $\beta_j$ .

## 3.5 Aplicación

En esta Sección se consideran los datos epidemiológicos analizados por Thomson et al. (2004) y Diggle et al. (2007). La información se obtuvo de una serie de mediciones en aldeas que fueron tomadas conjuntamente por el *Institut de Recherche pour le Développement (IRD)-Centre Pasteur du Cameroun* en Camerún entre 1991 y 2001 (Kamgno et al. 1997, Boussinesq et al. 2001), y posteriormente, por otros institutos (Takougang et al. 2002, Wanji et al. 2003). Con el objetivo de modelar la variación espacial en la prevalencia de Loa loa (presencia/ausencia de Loa-loa microfilarias en muestras de sangre), se estudió un conjunto de datos de 21938 individuos de 168 aldeas (Diggle & Ribeiro 2007). Las variables explicativas incluyen altura, junto con el máximo NDVI (máx(NDVI)) y la desviación estándar de NDVI (S.D.(NDVI)) calculada a partir de los análisis satélite repetidas en el tiempo. Según Thomson et al. (2004) y Diggle & Ribeiro (2007) la inclusión de los valores máximos y las desviaciones estándar del índice de vegetación permiten la discriminación entre localizaciones con diferentes grados de la variación estacional en el verdor.

En una primera etapa, se emplea la distancia euclidiana entre individuos dada en (3.5) porque todas las variables explicativas son continuas. Luego, las distancias empleando las variables explicativas centradas y la matriz  $\mathbf{B}_v = \mathbf{H}\mathbf{A}_v\mathbf{H}$  fueron obtenidas. Así, se construyeron las coordenadas principales utilizando las variables explicativas: altura, el máximo NDVI y la desviación estándar de NDVI. Para seleccionar las coordenadas principales a incluir en el modelo, se escogieron los mayores  $c(j)$ 's, pero cualquier otro procedimiento propuesto en la subsección 3.2.5 se puede utilizar. Para este estudio se seleccionaron tres coordenadas principales ( $\mathbf{X}_1$ ,  $\mathbf{X}_2$  y  $\mathbf{X}_3$ ), lo cual fue realizado con la finalidad de mantener el mismo número de variables explicativas. Sin embargo, esto no es necesario bajo el enfoque basado en distancias, ya que más coordenadas principales se podrían incluir en el modelo, lo que puede mejorar la bondad de ajuste del modelo propuesto como se vio en la Sección 3.4.



La dependencia residual espacial se maneja de la siguiente manera: sea  $y(\mathbf{s}_i)$  que denota el número de muestras positivas ( $n_i$ ) en la aldea  $\mathbf{s}_i$ . Además, se supone un conjunto de efectos aleatorios debido a los individuos dentro de cada aldea específica,  $e(\mathbf{s}_i)$ , los cuales son mutuamente independientes y gaussianos, con media 0 y varianza  $\tau_R^2 = \tau^2/\sigma^2$ ; este es el efecto pepita relativo en vez de  $\tau^2$ , lo cual se asume por conveniencia computacional. En el modelo se supone que los  $y(\mathbf{s}_i)$ 's son variables binomiales independientes dado el proceso estocástico espacial no observado  $z(\mathbf{s}_i)$  y el efecto aleatorio  $e(\mathbf{s}_i)$ , y además, que la respuesta media en  $\mathbf{s}_i$  depende de las variables explicativas observadas en la ubicación  $\mathbf{s}_i$ . Específicamente, si  $p(\mathbf{s}_i)$  es la probabilidad de que para una persona seleccionada aleatoriamente en la ubicación  $\mathbf{s}_i$  la prueba de Loa sea positiva, entonces

$$\log\left(\frac{p(\mathbf{s}_i)}{1-p(\mathbf{s}_i)}\right) = \beta_0 + x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2 + x_3(\mathbf{s}_i)\beta_3 + z(\mathbf{s}_i) + e(\mathbf{s}_i) \quad (3.37)$$

donde  $z(\mathbf{s}_i)$  se modela como un proceso gaussiano con media cero y varianza  $\sigma^2$  y estructura de correlación descrita por  $\text{Corr}(z(\mathbf{s}_i), z(\mathbf{s}_{i'})) = \rho(h; \vartheta)$ ,  $i, i' = 1, \dots, n$ .

En el modelo seleccionado no se encontró alguna evidencia de anisotropía, así que se omitió ésta en el modelo ajustado. La función de correlación  $\rho(h; \vartheta)$  depende únicamente de la distancia euclidiana  $h = \|\mathbf{s}_i - \mathbf{s}_{i'}\|$  entre localizaciones. Los papeles de  $z(\mathbf{s}_i)$  y  $e(\mathbf{s}_i)$  en el modelo, después de ajustarlo por las tres coordenadas principales, es capturar las variaciones residuales espaciales y no espaciales, respectivamente.

Para el ajuste de estos datos, Diggle et al. (2007) utilizaron una función de correlación exponencial,  $\rho(h; \vartheta) = \exp(-h/\vartheta)$ . En este análisis, se consideran las funciones de correlación de: la clase Matérn como se presentó en la ecuación (3.10) (obsérvese que el modelo exponencial puede ser considerado como un caso particular de la función de Matérn) y la función de correlación esférica (ver Cressie (1993)). El algoritmo MCMC fue utilizado en este problema para generar muestras de la función de distribución predictiva de la superficie completa  $z(\mathbf{s}_i)$  con una resolución de 1 Km, dados los valores observados de la variable respuesta  $y(\mathbf{s}_i)$  en cada aldea muestreada y las tres coordenadas principales también con una resolución de 1 km en toda la región del estudio.

A continuación, primero se ajustan los datos utilizando un modelo binomial mixto DB empleando la función de enlace logit con efectos aleatorios independientes e igualmente distribuidos. Aplicando el GLM clásico y tomado las coordenadas principales  $\mathbf{X}_1$ ,  $\mathbf{X}_2$  y  $\mathbf{X}_3$ , se obtienen las estimaciones  $\hat{\beta}_0^0 = -0.713$ ,  $\hat{\beta}_1^0 = -0.497$ ,  $\hat{\beta}_2^0 = 0.216$  y  $\hat{\beta}_3^0 = -0.251$ . Luego, el panel izquierdo de la Figura 1.1 muestra la nube de puntos del variograma; su apariencia difusa es totalmente típica. Una variante más estable de la nube de puntos del variograma es el variograma empírico como se ilustra en el panel derecho

de la Figura 3.1. Utilizando éste, se obtuvieron las estimaciones de los efectos aleatorios y a partir de éstos se obtiene la estimación del vector de parámetros  $\boldsymbol{\theta}$ , dada por  $\hat{\boldsymbol{\theta}}_0 = (\hat{\sigma}_0^2, \hat{\vartheta}_0, \hat{\tau}_{R0}^2) = (0.5, 5.6, 0.001)$ , el cual es obtenido utilizando el método de mínimos cuadrados (Cressie 1993). Luego se utilizan las anteriores estimaciones como valores iniciales para ejecutar el algoritmo MCMC empleando el DBSGLMM. El tamaño de la muestra de Monte Carlo fue 1000 y se descartaron las primeras 100000 iteraciones ya que fue tomado como un periodo de entrenamiento de la cadena de Markov. Un total de 1000000 de simulaciones se ejecutaron, tomando muestras cada 1000 iteraciones. Se estudiaron los siguientes modelos

$H_1$  : No correlación espacial.

$H_2$  : Función de correlación exponencial,  $\vartheta > 0$  y  $\tau_R^2 = 0$ .

$H_3$  : Función de correlación exponencial,  $\vartheta > 0$  y  $\tau_R^2 \geq 0$ .

$H_4$  : Función de correlación de Matérn,  $\kappa = 1, \vartheta > 0$  y  $\tau_R^2 \geq 0$ .

$H_5$  : Función de correlación esférica,  $\vartheta > 0$  y  $\tau_R^2 \geq 0$ .

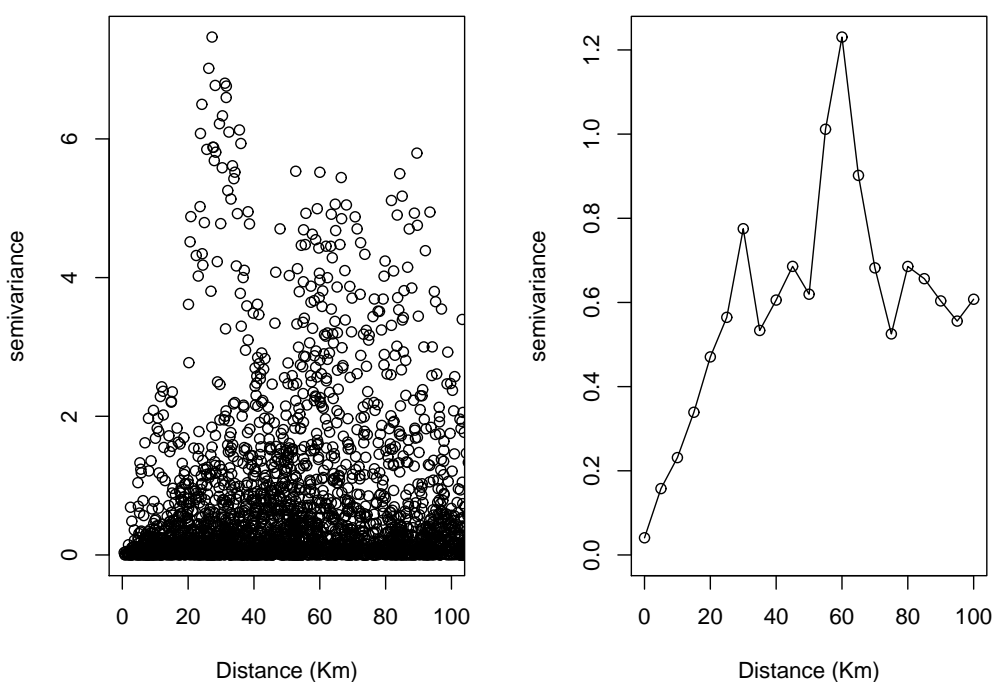


FIGURA 3.1: Nube de puntos del variograma (panel izquierdo) y variograma empírico (panel derecho) para la prevalencia de *Loa loa* utilizando DB.

Los resultados obtenidos después de ejecutar el algoritmo MCMC para el DBSGLMM se presentan en la Tabla 3.1. Al comparar la diferencia entre los  $\log \hat{L}$ 's para  $H_3$  y  $H_1$  con distribución  $0.5\chi_{(2)}^2$  (cuyo cuantil 95 % es 2.996), se encuentra que hay una evidente correlación espacial. Además, como en Diggle et al. (2007), el efecto pepita no es significativo ( $\tau^2 = \sigma^2\tau_R^2$ ) al comparar las diferencias de  $\log \hat{L}$ 's entre  $H_2$  y  $H_3$  con distribución  $0.5\chi_{(1)}^2$ , cuyo cuantil 95 % es 1.921. Observe que la estimación de  $\vartheta$  en el modelo  $H_2$  es más pequeña que en el modelo  $H_3$ ; resultado que es consistente con el modelo  $H_2$  el cual no tiene efecto pepita y el modelo  $H_3$  que tiene un efecto pepita pequeño. Los resultados de la Tabla 3.1 también muestran que la elección de la función de correlación no es importante porque todos los  $\log \hat{L}$ 's con correlación espacial se parecen, aunque el modelo exponencial con un pequeño efecto pepita fue ligeramente mejor que los otros modelos estudiados.

Algunas ejecuciones del algoritmo MCMC para el enfoque DBSGLMM se realizaron con diferentes valores iniciales ( $\beta_0^0, \beta_1^0, \beta_2^0, \beta_3^0, \sigma_0^2, \vartheta_0, \tau_{R0}^2$ ) y diferentes funciones de correlación, pero los resultados no difirieron mucho ya que el patrón fue el mismo al de la Tabla 3.1.

TABLA 3.1: Estimaciones de máxima verosimilitud utilizando MCMC para los modelos  $H_1, H_2, H_3, H_4$  y  $H_5$ .

Modelo	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}^2$	$\hat{\vartheta}$	$\hat{\tau}_R^2$	$\log \hat{L}$
$H_1$	-1.870	-0.595	0.193	-0.335	0.795	*	0.00	94.0
$H_2$	-1.862	-0.489	0.254	-0.273	0.801	21.609	0.00	148.9
$H_3$	-1.863	-0.484	0.209	-0.288	0.769	26.195	0.04	149.5
$H_4$	-1.836	-0.482	0.190	-0.278	0.684	12.697	0.11	147.7
$H_5$	-1.820	-0.500	0.205	-0.277	0.699	35.566	0.04	145.4

La comparación entre modelos es utilizada de una manera informal porque algunos de los modelos comparados no son anidados. En el análisis también se asume que la razón de log-verosimilitud se puede razonablemente aproximar a la distribución  $\chi^2$ , pero esto es difícil de demostrar rigurosamente. Sin embargo, el modelo exponencial con un efecto pepita tiene ligeramente la log-verosimilitud más alta, lo cual es un criterio útil cuando no están anidados. Por lo tanto, este modelo será investigado aún más mediante simulación en la siguiente subsección.

### 3.5.1 Inferencia sobre los parámetros, diagnóstico y predicción utilizando DBSGLMM

Después de haber realizado una ejecución inicial del algoritmo MCMC empleando el DBSGLMM y con el objetivo de hacer inferencia estadística sobre los parámetros involucrados en el modelo (3.37), se repite 200 veces el proceso de ejecución de 1000000 simulaciones, obteniendo muestras cada 1000 iteraciones luego de la eliminación de las primeras 100000 iteraciones.

La deviance fue de 112.23 la cual se compara con una  $\chi^2_{(0.05,161)} = 132.66$ , por este criterio la función logit en el DBSGLMM parece ajustar bien. Por lo tanto, utilizando la función de enlace logit, se encuentra que el pseudo  $R_k^2 = 0.93$ , lo que sugiere un buen ajuste del modelo propuesto. La Tabla 3.2 muestra la media, desviación estándar (S.D.) y el intervalo de confianza del 95 % para cada uno de los parámetros del modelo (3.37). Con respecto a los parámetros de regresión, la Tabla 3.2 muestra que las coordenadas principales  $\mathbf{X}_1$  y  $\mathbf{X}_3$  reducen la prevalencia de Loa loa, mientras que esta prevalencia crece con la coordenada principal  $\mathbf{X}_2$ . En algunos casos, estas conclusiones sobre las coordenadas principales podrían no ser muy relevantes para el objetivo general, ya que el investigador podría estar más interesado en la información de las variables originales; sin embargo, éstas son presentadas con el objetivo de destacar que son significativas y deben ser consideradas en la predicción de la prevalencia de Loa loa. La Figura 3.2 confirma estos resultados, observe que los parámetros  $\beta_1$ ,  $\beta_2$  y  $\beta_3$  son diferentes de cero, y por lo tanto, el verdor de la vegetación circundante y la altura afectan notablemente la prevalencia de Loa loa, ya que ellas juegan un papel importante en las distancias a partir de las cuales las coordenadas principales  $\mathbf{X}_1$ ,  $\mathbf{X}_2$  y  $\mathbf{X}_3$  fueron construidas.

TABLA 3.2: Estimaciones e intervalos de confianza para los parámetros involucrados en el DBSGLMM (modelo (3.37)) para ajustar la prevalencia de Loa loa.

Parámetro	Media	S.D.	2.5 % cuantil	97.5 % cuantil
$\beta_0$	-1.870	0.031	-1.928	-1.818
$\beta_1(X_1)$	-0.393	0.040	-0.459	-0.314
$\beta_2(X_2)$	0.158	0.052	0.051	0.255
$\beta_3(X_3)$	-0.286	0.072	-0.421	-0.142
$\sigma^2$	0.801	0.059	0.696	0.927
$\vartheta$	37.465	8.041	23.486	53.472
$\tau_R^2$	0.089	0.054	0.000	0.202

Para interpretar los resultados para el parámetro de escala con respecto a la distancia,  $\vartheta$ , la distancia mínima entre las localizaciones o ubicaciones muestreadas es de 0.56 km y la máxima de 715.7 km. En particular, el intervalo

de confianza del 95 % para  $\sigma^2$  es (0.70; 0.93) en la escala logit, el intervalo para  $\vartheta$  es (23.49; 53.47) km lo cual indica que esta variación residual espacial no es tan alta ni tan baja y el intervalo para  $\tau_R^2$  de (0; 0.20) muestra que este efecto es prácticamente cero, y por lo tanto, el efecto pepita es poco relevante. Estos resultados se confirman también por medio de la Figura 3.2. Observe que los parámetros estimados del modelo  $H_3$  presentados en la Tabla 3.1 se encuentran dentro de los intervalos de confianza presentados en la Tabla 3.2; esto era de esperarse ya que el resultado de la Tabla 3.1 para el modelo  $H_3$  es una realización obtenida de la aleatorización del proceso. Por ejemplo, la estimación puntual de  $\tau_R^2$  difiere entre la Tabla 3.1 y la Tabla 3.2, pero el intervalo de confianza contiene ambas estimaciones y en los dos casos,  $\tau_R^2$  no es significativo.

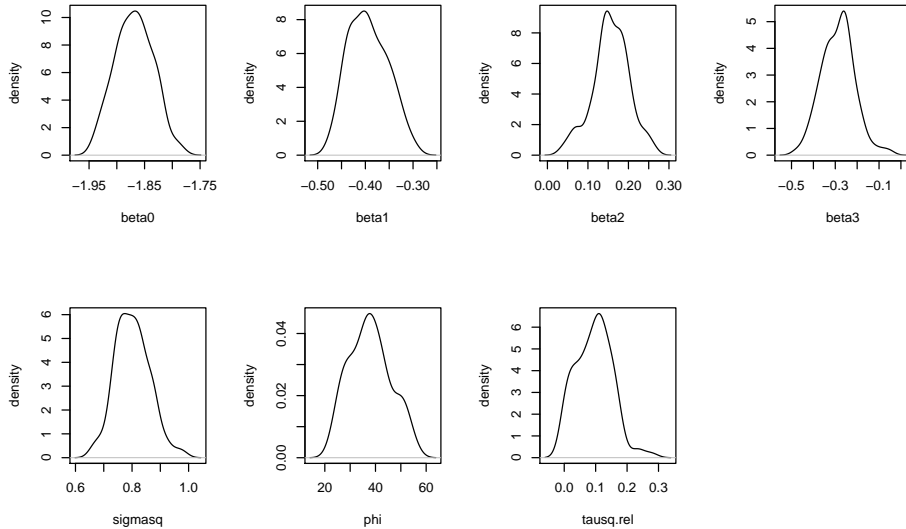


FIGURA 3.2: Distribución de los parámetros involucrados en el DBSGLMM.

Además, la Figura 3.3(c) grafica los residuales de Pearson de la aldea ( $r_{P_i}$ ) contra los valores pronosticados y muestra la ausencia deseada de cualquier relación, indicando un adecuado ajuste de primer orden. Los paneles izquierdos (a) y (b) de la Figura 3.3 muestran la ausencia de cualquier relación entre los residuales de Pearson (o los de Deviance) y la aldea. El panel de la derecha (d) de la Figura 3.3 grafica la prevalencia observada de *Loa loa* microfilaria contra las prevalencias pronosticadas utilizando el DBSGLMM; este panel muestra sustancialmente puntos menos dispersos que los ajustados por Thomson et al. (2004) y Diggle et al. (2007).

Una vez ajustado y validado el DBSGLMM, se obtuvieron el mapa de pronósticos de la prevalencia de *Loa loa* y su correspondiente desviación

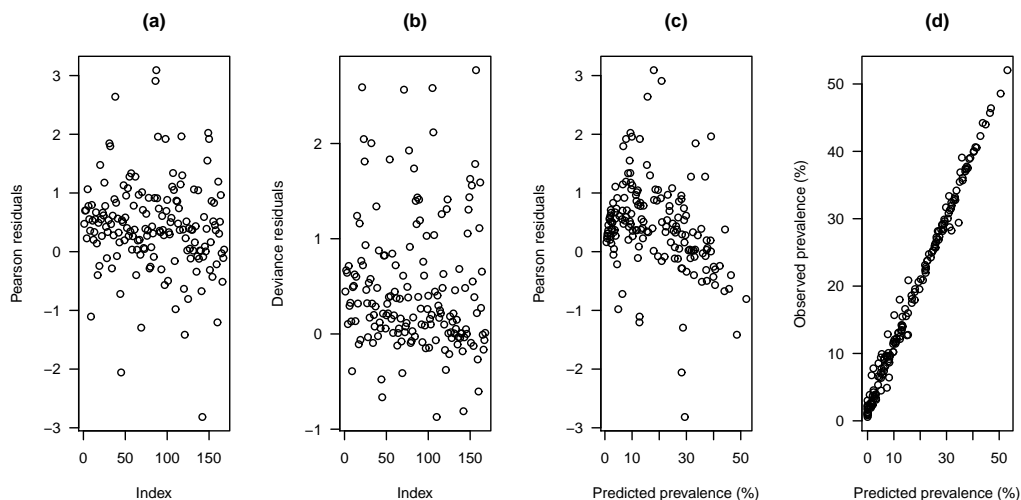


FIGURA 3.3: (a) Residuales de Pearson contra la aldea, (b) residuales de deviance contra la aldea, (c) residuales de Pearson contra valores pronosticados para el DBSGLMM, modelo (3.37), y (d) relación entre prevalencia observada de *Loa loa* microfilaria y prevalencia pronosticada usando DBSGLMM.

estándar para la predicción del error; los resultados se presentan en la Figura 3.4. Obsérvese que en las ubicaciones de las aldeas observadas en la muestra, la distribución de las predicciones muestran una varianza reducida alrededor de estas aldeas. En el panel izquierdo de la Figura 3.4, las áreas dentro de los límites de color naranja pálido y amarillo representan altas prevalencias de *Loa loa*, los cuales exceden el umbral del 20% a partir del cual hay una intervención de las entidades públicas que manejan el sistema de salubridad en Camerún. Igualmente, las áreas en el rango de color rojo y naranja son aquellas en las cuales hay una baja prevalencia de *Loa loa*, que no excede este umbral del 20% .

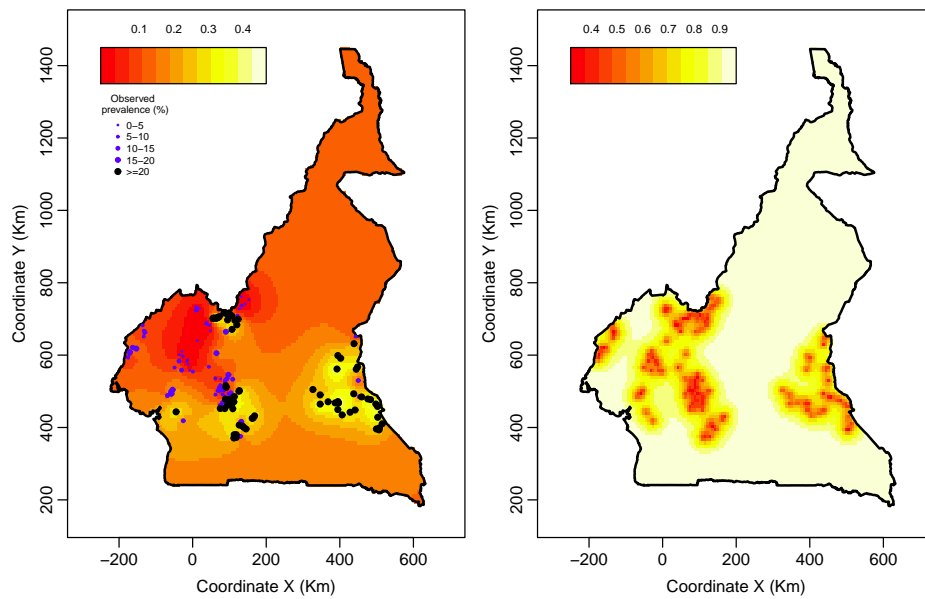


FIGURA 3.4: Prevalencias estimadas para el *Loa loa* microfilaria utilizando la aproximación DBSGLMM, sobrepuesta con la prevalencia observada en el campo de estudio (panel izquierdo). Raíz cuadrada de las varianzas de la predicción (panel derecho).

# Capítulo 4

## Modelo lineal beta espacial mixto con dispersión variable

### 4.1 Introducción

La caracterización de la variabilidad espacial de los atributos del suelo es esencial para lograr un mejor entendimiento de relaciones complejas entre propiedades del suelo y factores ambientales (Goovaerts 1998) y para determinar prácticas de gestión adecuadas para utilizar los recursos del suelo (Bouma et al. 1999). También tiene implicaciones prácticas en el diseño de muestreo en estudios ecológicos, ambientales y de agricultura (Stein & Ettema 2003). Además, ha aumentado en el último tiempo la demanda de información más precisa sobre la distribución espacial de los suelos con la inclusión de la dependencia espacial y de escala en los modelos ecológicos y sistemas de gestión ambiental (Godwin & Miller 2003). En estas áreas, se puede tener una variable respuesta espacial dada por una tasa, una proporción o partes por millón que se limita a un valor en el intervalo  $(0,1)$ . En general, también se tiene la presencia de un conjunto de covariables espaciales asociadas a esta clase de respuesta.

Particularmente, en este capítulo se estudia el conjunto de datos que contiene 147 observaciones del perfil del suelo que se tomaron en el área de investigación del programa Tropenbos Camerún (Tropenbos Cameroon Programme, TCP), el cual es un representante de la región de la selva húmeda del sudoeste de Camerún y áreas adyacentes de Guinea Ecuatorial y Gabon (Yemefack et al. 2005). Específicamente, se estudiaron las mediciones que se hicieron en muestras de la capa de suelo fijo 0-10 cm. El conjunto de datos es de dos fuentes: la primera, 45 perfiles de suelo representativos se describieron y muestrearon por horizonte genético, se calcularon las características del suelo para cada una de las tres capas fijas como promedios ponderados utilizando el es-



pesor del horizonte genético. La segunda, 102 parcelas de los distintos tipos de cobertura del suelo de uso/tierra se muestrearon en las tres profundidades fijas. Cada una de estas muestras era un granel compuesto de cinco sub muestras tomadas con una barrena en una parcela diagonal. Para los dos conjuntos de datos, el muestreo se hizo intencionalmente y subjetivamente en ciertas localizaciones para representar los tipos de suelo y usos de la tierra. Los análisis de laboratorio se hicieron por métodos estándares locales (Pauwels et al. 1992). En este estudio, se desea investigar la relación entre el porcentaje de arcilla y las variables explicativas: altura en metros sobre el nivel del mar (elevation, ELE), zona agro-ecológica (agro-ecological zone, ZONE) y grupo de referencia del suelo (World Reference Base for Soil Resources, WRB). Para cada observación, se registraron las coordenadas este y norte (UTM Zone 32N, WGS84 datum, en metros).

También se estudió en este capítulo las muestras de suelo que fueron tomadas de 0-20 cm de profundidad de la capa en cada uno de 178 localizaciones. El conjunto de datos tiene información acerca del contenido de magnesio, la ubicación espacial, altitud (altitude, ALT) y código de la sub-región (sub-region, SR) en cada muestra; las sub-regiones están asociadas con tres periodos de fertilización en diferentes áreas. El diseño muestral es un retículo regular incompleto con un espacio aproximadamente de 50 metros. Los datos fueron recolectados por investigadores de PESEAGRO y EMBRAPA-Solos, Rio de Janeiro, Brasil (Capeche et al. 1997, Diggle & Ribeiro 2007).

Los anteriores problemas se pueden resolver utilizando el modelo de regresión beta propuesto por (Ferrari & Cribari-Neto 2004, Simas et al. 2010), pero en éste no se considera la correlación espacial entre las observaciones. Por otro lado, se podría emplear los modelos lineales generalizados espaciales mixtos (spatial generalised linear mixed models, SGLMMs) propuestos por (Diggle et al. 1998, Zhang 2002, Christensen 2004), pero éstos modelos no consideran respuestas beta en sus ajustes. Por lo tanto, en este capítulo se propone una metodología que enlaza esas dos metodologías con la finalidad de dar una solución a los anteriores problemas.

En este capítulo se propone un modelo lineal beta espacial mixto (beta spatial linear mixed model, BSLMM) con dispersión variable utilizando MCML. El método propuesto se utiliza en situaciones donde la variable respuesta es una razón o proporción que esta relacionada con determinadas variables explicativas. Para este fin, se desarrolla una aproximación utilizando SGLMMs con la transformación Box-Cox en el modelo de precisión. Por lo tanto, el proceso de optimización de los parámetros se realiza tanto para modelo espacial de media (spatial mean model, SMM) como para el modelo espacial de dispersión variable (spatial variable dispersion model, SVDM). Todos los parámetros se estiman por máxima verosimilitud utilizando MCMC, la cual es una técnica muy factible y útil para este modelo. Además, se realiza la infe-

rencia estadística sobre los parámetros utilizando las aproximaciones obtenidas a partir de la normalidad asintótica del estimador de máxima verosimilitud, mediante una adaptación de la estadística tradicional para ajustar el modelo propuesto. También se desarrolla el diagnóstico del modelo y la predicción de una nueva observación. En este sentido, el método presentado permite dar una buena solución a los problemas de los contenidos de arcilla y magnesio, respectivamente.

Es de recalcar que para el problema de predicción espacial, las únicas variables explicativas que están en muchos casos disponibles son aquellas de las localizaciones muestreadas; en este sentido, no siempre se pueden obtener predicciones en localizaciones no muestreadas donde estas variables explicativas no se han observado. En solo uno de los dos estudios presentados en este capítulo dichas variables si fueron observadas y por ello, se pueden hacer predicciones en todas las localizaciones dentro de la región de estudio. En el otro caso, solo fue posible hacer las predicciones en las localizaciones observadas.

Este capítulo esta dividido de la siguiente manera; en la Sección 4.2 se desarrolla la metodología propuesta: modelo lineal beta espacial mixto, representación espectral, algoritmo de máxima verosimilitud MCMC para el BSLMM, medidas de bondad de ajuste y predicción espacial de nuevos sujetos. En la sección 4.3 se presenta un experimento simulado utilizando el BSLMM. En la sección 4.4 se desarrolla dos aplicaciones que ilustran la metodología propuesta. Entonces, el trabajo termina con algunas conclusiones.

## 4.2 Modelo lineal espacial beta mixto

Suponga que se que esta interesado en alguna variable respuesta beta que se puede relacionar con información geo-referenciada en cada localización muestreada, tal como la latitud y longitud y variables explicativas binarias, categóricas y continuas. Si  $M = \{M(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^d\}$ , denota un campo aleatorio gaussiano con  $\mathbf{s}$  una localización en el espacio euclidiano  $d$ -dimensional y con funciones de media  $E[M(\mathbf{s})] = \beta_0 + \mathbf{x}^t(\mathbf{s})\boldsymbol{\beta}$  y covarianza  $\text{Cov}(M(\mathbf{s}), M(\mathbf{s}')) = \sigma^2 \rho(\mathbf{s}, \mathbf{s}'; \varsigma) + \tau^2 1_{\{\mathbf{s}=\mathbf{s}'\}}$ , donde  $\beta_0 \in \mathbb{R}$  es un parámetro relacionado al intercepto desconocido,  $\boldsymbol{\beta} \in \mathbb{R}^p$  es un vector de parámetros de regresión desconocidos,  $\mathbf{x}(\mathbf{s})$ 's son covariables conocidas dependientes de la ubicación espacial,  $\sigma^2 > 0$  es un parámetro de dispersión,  $\rho(s, s', \varsigma)$  es una función de correlación en  $\mathbb{R}^2$ ,  $\varsigma$  es un parámetro de correlación y  $\tau^2 \geq 0$  es llamado el efecto pepita.

Condicionando sobre  $M$ , el proceso espacial estocástico,  $\{Y(\mathbf{s}), \mathbf{s} \in \mathbb{R}^d\}$ , esta conformado por variables aleatorias mutuamente independientes. Entonces, se dice que una variable aleatoria  $\{Y(\mathbf{s}) | M\}$  sigue una distribución beta con parámetros  $p(\mathbf{s}), q(\mathbf{s}) > 0$ , denotado por  $\text{Beta}(p(\mathbf{s}), q(\mathbf{s}))$ , si la distribución de

$\{Y(\mathbf{s}) \mid M\}$  admite la siguiente función de densidad con respecto a la medida de Lebesgue

$$f(y(\mathbf{s}) \mid p(\mathbf{s}), q(\mathbf{s})) = \frac{\Gamma[p(\mathbf{s}) + q(\mathbf{s})]}{\Gamma[p(\mathbf{s})]\Gamma[q(\mathbf{s})]} y(\mathbf{s})^{p(\mathbf{s})-1} [1 - y(\mathbf{s})]^{q(\mathbf{s})-1} I_{(0,1)}(y(\mathbf{s})) \quad (4.1)$$

donde  $\Gamma(\cdot)$  denota la función gamma. La media y varianza de  $\{Y(\mathbf{s}) \mid M\}$  están dadas, respectivamente, por

$$\begin{aligned} E(Y(\mathbf{s}) \mid M) = \mu(\mathbf{s}) &= \frac{p(\mathbf{s})}{p(\mathbf{s}) + q(\mathbf{s})} \\ \text{Var}(Y(\mathbf{s}) \mid M) &= \frac{p(\mathbf{s})q(\mathbf{s})}{(p(\mathbf{s}) + q(\mathbf{s}))^2(p(\mathbf{s}) + q(\mathbf{s}) + 1)} \end{aligned}$$

La función de densidad, en (4.1) presenta diferentes formas de dependencia sobre los valores de estos dos parámetros (ver detalles en Ferrari & Cribari-Neto (2004)). Por lo tanto, el modelo espacial beta puede ser apropiado para explicar el porcentaje de arcilla y de magnesio como una función de las variables explicativas estudiadas en cada caso y presentadas anteriormente.

De acuerdo a Ferrari & Cribari-Neto (2004), se puede reparametrizar la función de densidad dada en (4.1) como  $\phi(\mathbf{s}) = p(\mathbf{s}) + q(\mathbf{s})$ ; así, se encuentra que  $p(\mathbf{s}) = \mu(\mathbf{s})\phi(\mathbf{s})$  y  $q(\mathbf{s}) = \phi(\mathbf{s})(1 - \mu(\mathbf{s}))$ . Entonces utilizando esta reparametrización, la función de densidad de la distribución beta (4.1) se puede reescribir como

$$\begin{aligned} f(y(\mathbf{s}) \mid \mu(\mathbf{s}), \phi(\mathbf{s})) &= \frac{\Gamma(\phi(\mathbf{s}))}{\Gamma[\mu(\mathbf{s})\phi(\mathbf{s})]\Gamma[(1 - \mu(\mathbf{s}))\phi(\mathbf{s})]} y(\mathbf{s})^{[\mu(\mathbf{s})\phi(\mathbf{s})-1]} \\ &\quad \times [1 - y(\mathbf{s})]^{[(1 - \mu(\mathbf{s}))\phi(\mathbf{s})-1]} I_{(0,1)}(y(\mathbf{s})) \end{aligned} \quad (4.2)$$

donde  $0 < \mu(\mathbf{s}) < 1$ ,  $\phi(\mathbf{s}) > 0$  y  $\Gamma(\cdot)$  es la función gamma. Además,  $E(Y(\mathbf{s}) \mid M) = \mu(\mathbf{s})$  y  $\text{Var}(Y(\mathbf{s}) \mid M) = \text{Var}(\mu(\mathbf{s}))/[1 + \phi(\mathbf{s})]$ , donde  $\text{Var}(\mu(\mathbf{s})) = \mu(\mathbf{s})[1 - \mu(\mathbf{s})]$  es la función de varianza,  $\mu(\mathbf{s})$  es la media de la variable respuesta y  $\phi(\mathbf{s})$  se puede interpretar como un parámetro de precisión en el sentido que para un  $\mu(\mathbf{s})$  fijo, un valor grande de  $\phi(\mathbf{s})$  lleva a una menor varianza de  $(Y(\mathbf{s}) \mid M)$ .

Para formar un modelo de regresión espacial con la aproximación del GLMM extendido, se utiliza dos funciones de enlace; una para el parámetro de localización ( $\mu(\mathbf{s})$ ) y la otra para el parámetro de precisión ( $\phi(\mathbf{s})$ ). Este proceso es similar al empleado por Ferrari & Cribari-Neto (2004), Smithson & Verkuilen (2006) y Simas et al. (2010). La función de enlace es una función no lineal, suave y monótona que relaciona el espacio no limitado del predictor lineal en el espacio muestral apropiado de las observaciones, por lo que “enlaza” el predictor lineal con las observaciones.

Considere  $n$  distintas ubicaciones  $\mathbf{s}_1, \dots, \mathbf{s}_n$  y suponga que se observa una realización  $\mathbf{y}_s = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))^t$  de  $\mathbf{Y}_s = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))^t$ , donde la densidad condicional de  $Y(\mathbf{s}_i)$  dada  $M(\mathbf{s}_i)$  está dada por (4.2) con media  $\mu(\mathbf{s}_i)$  y dispersión desconocida  $\phi(\mathbf{s}_i)$ ,  $i = 1, \dots, n$ .

La interpolación propuesta se construye para un modelo de campo aleatorio no gaussiano considerando específicamente variables explicativas categóricas, continuas e indicadoras en el modelo de tendencia. El conjunto de datos generan un mecanismo condicional sobre la señal del modelo siguiendo un modelo lineal generalizado clásico como el descrito por McCullagh & Nelder (1989). Explícitamente, se asume que la variable respuesta  $Y(\mathbf{s}_i)$ ,  $i = 1, 2, \dots, n$ , tiene una distribución beta, donde el SMM y el SVDM son obtenidos, respectivamente, mediante

$$\begin{aligned} g_1(\mu(\mathbf{s}_i)) = M_1(\mathbf{s}_i) &= \sum_{j=0}^{p_1} v_j(\mathbf{s}_i) \zeta_j + z_1(\mathbf{s}_i) = \mathbf{v}^t(\mathbf{s}_i) \boldsymbol{\zeta} + z_1(\mathbf{s}_i) \\ g_2(\phi(\mathbf{s}_i)) = M_2(\mathbf{s}_i) &= \sum_{j=0}^{p_2} u_j(\mathbf{s}_i) \varphi_j + z_2(\mathbf{s}_i) = \mathbf{u}^t(\mathbf{s}_i) \boldsymbol{\varphi} + z_2(\mathbf{s}_i) \end{aligned} \quad (4.3)$$

donde  $v_0(\mathbf{s}_i) = 1$ ,  $u_0(\mathbf{s}_i) = 1$ ,  $\boldsymbol{\zeta}^t = (\zeta_0, \zeta_1, \dots, \zeta_{p_1})$  y  $\boldsymbol{\varphi}^t = (\varphi_0, \varphi_1, \dots, \varphi_{p_2})$  son vectores de parámetros de tendencia desconocidos, con  $\boldsymbol{\zeta} \in \mathbb{R}^{p_1+1}$  y  $\boldsymbol{\varphi} \in \mathbb{R}^{p_2+1}$ ,  $p_1 + p_2 < n$ . Los  $z_1(\mathbf{s}_i)$ 's y  $z_2(\mathbf{s}_i)$ 's se asumen que tienen estructura de campo aleatorio, cuyas funciones de covarianza están dadas por  $\text{Cov}(z_1(\mathbf{s}_i), z_1(\mathbf{s}_{i'})) = \sigma_1^2 \rho_1(\mathbf{s}_i, \mathbf{s}_{i'}; \varsigma_1) + \tau_1^2 1_{\{\mathbf{s}_i = \mathbf{s}_{i'}\}}$  y  $\text{Cov}(z_2(\mathbf{s}_i), z_2(\mathbf{s}_{i'})) = \sigma_2^2 \rho_2(\mathbf{s}_i, \mathbf{s}_{i'}; \varsigma_2) + \tau_2^2 1_{\{\mathbf{s}_i = \mathbf{s}_{i'}\}}$ , respectivamente; donde los  $\sigma_j^2 > 0$ 's son los parámetros de dispersión  $j = 1, 2$ , los  $\rho_j(\mathbf{s}_i, \mathbf{s}_{i'}, \varsigma_j)$ 's son las funciones de correlación, cada una en  $\mathbb{R}^2$ , los  $\varsigma_j$ 's son los parámetros de correlación y los  $\tau_j^2 \geq 0$ 's son los efectos pepita. Además,  $\mathbf{v}^t(\mathbf{s}_i) = (v_0(\mathbf{s}_i), v_1(\mathbf{s}_i), \dots, v_{p_1}(\mathbf{s}_i))$  y  $\mathbf{u}^t(\mathbf{s}_i) = (u_0(\mathbf{s}_i), u_1(\mathbf{s}_i), \dots, u_{p_2}(\mathbf{s}_i))$  son vectores de observaciones de  $p_1 + 1$  y  $p_2 + 1$  covariables conocidas, respectivamente. Los modelos (4.3) son llamados el SMM y el SVDM de  $Y(\mathbf{s}_i)$ , respectivamente.

Las funciones enlace  $g_1 : (0, 1) \rightarrow \mathbb{R}$  y  $g_2 : (0, \infty) \rightarrow \mathbb{R}$  son estrictamente monótonas y doble diferenciables. Algunas posibles elecciones para la función enlace  $g_1(\mu(\mathbf{s}_i))$  son: la logit,  $g_1(\mu(\mathbf{s}_i)) = \log\{\mu(\mathbf{s}_i)/[1 - \mu(\mathbf{s}_i)]\}$ ; la probit,  $g_1(\mu(\mathbf{s}_i)) = \Phi^{-1}(\mu(\mathbf{s}_i))$  donde  $\Phi(\cdot)$  es la función de distribución acumulada de una variable normal estándar; el complemento loglog (cloglog),  $g_1(\mu(\mathbf{s}_i)) = \log\{-\log[1 - \mu(\mathbf{s}_i)]\}$ ; y la loglog,  $g_1(\mu(\mathbf{s}_i)) = -\log\{-\log(\mu(\mathbf{s}_i))\}$ . Una rica discusión de estas funciones enlace se presentan en (Atkinson 1985, McCullagh & Nelder 1989). Tentativamente las funciones enlace para  $g_2$  son (Ferrari & Cribari-Neto 2004, Smithson & Verkuilen 2006, Simas et al. 2010): la logaritmo,  $g_2(\phi(\mathbf{s}_i)) = \log \phi(\mathbf{s}_i)$ ; la raíz cuadrada,  $g_2(\phi(\mathbf{s}_i)) = \sqrt{\phi(\mathbf{s}_i)}$ ; y la identidad,  $g_2(\phi(\mathbf{s}_i)) = \phi(\mathbf{s}_i)$ , entre otras.

En forma matricial, los modelos lineales beta espaciales mixtos presentados

en (4.3) se pueden expresar como

$$g_1(\boldsymbol{\mu}_s) = M_1 = \mathbf{V}_s \boldsymbol{\zeta} + \mathbf{E}_1 \mathbf{z}_1 \quad g_2(\boldsymbol{\phi}_s) = M_2 = \mathbf{U}_s \boldsymbol{\varphi} + \mathbf{E}_2 \mathbf{z}_2 \quad (4.4)$$

donde  $\mathbf{V}_s = (\mathbf{1}, \mathbf{V}_1, \dots, \mathbf{V}_{p_1})$ ,  $\mathbf{U}_s = (\mathbf{1}, \mathbf{U}_1, \dots, \mathbf{U}_{p_2})$ ,  $\boldsymbol{\mu}_s = E(\mathbf{y}_s | \mathbf{V}_s, \mathbf{z}_1) = (\mu(\mathbf{s}_1), \dots, \mu(\mathbf{s}_n))^t$ ,  $\boldsymbol{\phi}_s = (\phi(\mathbf{s}_1), \dots, \phi(\mathbf{s}_n))^t$ ,  $\mathbf{1}$  es un vector de unos de tamaño  $n \times 1$ ,  $\mathbf{V}_i$  ( $i = 1, \dots, p_1$ ) y  $\mathbf{U}_j$  ( $j = 1, \dots, p_2$ ) son vectores de variables explicativas espaciales asociadas con los SMM y SVDM, respectivamente. Por otra parte, tanto  $\mathbf{V}_s$  como  $\mathbf{U}_s$  pueden involucrar variables continuas, categóricas y binarias o incluso una mezcla de ellas. Además,  $\mathbf{z}_1 = (z_1(\mathbf{s}_1), \dots, z_1(\mathbf{s}_n))^t$ ,  $\mathbf{z}_2 = (z_2(\mathbf{s}_1), \dots, z_2(\mathbf{s}_n))^t$ , y  $\mathbf{E}_1$  y  $\mathbf{E}_2$  son matrices diseño no aleatorias compatibles con los efectos aleatorios  $\mathbf{z}_1$  y  $\mathbf{z}_2$ , respectivamente.

Para el SVDM, todos los anteriores casos se pueden considerar en una clase general de funciones enlace, que tiene la siguiente forma

$$g_{\nu_2}(\phi(\mathbf{s}_i)) = \begin{cases} (\phi^{\nu_2}(\mathbf{s}_i))/\nu_2 & \text{if } \nu_2 > 0 \\ \log(\phi(\mathbf{s}_i)) & \text{if } \nu_2 = 0 \end{cases}$$

Como la función de enlace  $g_{\nu_2}(\phi(\mathbf{s}_i))$  relaciona a  $\boldsymbol{\phi}_s$  con  $M_2$ , se tiene que

$$\boldsymbol{\phi}_s = m_2 g_{\nu_2}^{-1}(M_2) \quad (4.5)$$

donde  $m_2$  es una función determinística.

Si  $\nu_2 > 0$  entonces  $g_{\nu_2}(\mathbb{R}) = (-1/\nu_2, \infty)$  y es necesario definir  $g_{\nu_2}^{-1}(\phi(\mathbf{s}_i)) = 0$  cuando  $\phi(\mathbf{s}_i) \notin g_{\nu_2}(\mathbb{R})$ . Por lo tanto, se hace  $f(y(\mathbf{s}_i) | \mu(\mathbf{s}_i), \phi(\mathbf{s}_i)) = 1_{\{y(\mathbf{s}_i)=0\}}$  cuando  $\phi(\mathbf{s}_i) = 0$ .

Los modelos (4.4) no son un simple nuevo formato porque estos modelos se acomodan a una estructura de datos mas complejos, como lo son los datos espaciales. Por ejemplo, con  $\mathbf{E}_1$  y  $\mathbf{E}_2$  y los efectos aleatorios  $\mathbf{z}_1$  y  $\mathbf{z}_2$  definidos apropiadamente en ambos modelos, respectivamente, éstos abarcan datos agrupados no normales y datos de factores cruzados (ver (Breslow & Clayton 1993)). Cuando se definen  $\mathbf{E}_1$  y  $\mathbf{E}_2$  como matrices indicadoras pertenecientes a regiones espaciales (por ejemplo, condados o distritos censales), los modelos (4.4) funcionan también para datos de área.

La base de muchos procedimientos, incluyendo el descrito anteriormente, es una distribución normal multivariada (multivariate normal, MN). En este caso, considere los procesos  $\mathbf{z}_1$  y  $\mathbf{z}_2$ , y asuma que ellos tienen distribución MN, es decir,  $\mathbf{z}_1 \sim MV(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{z}_1})$  y  $\mathbf{z}_2 \sim MV(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{z}_2})$ , respectivamente. La clave para el modelamiento espacial en problemas de grandes muestras es la parametrización de estas normales multivariadas de una manera eficiente. Reescribiendo  $\mathbf{z}_j$ ,  $j = 1, 2$ , en términos de una descomposición espectral (ver detalles en el apéndice 4.1), se tiene

$$\mathbf{z}_j = \boldsymbol{\Psi}_j \boldsymbol{\delta}_j, \quad j = 1, 2 \quad (4.6)$$

donde  $\Psi_j = (\psi_{j1}, \dots, \psi_{jn})$  es una matriz de tamaño  $n \times n$  con  $\psi_{ji} = (\psi_{ji}(\mathbf{s}_1), \dots, \psi_{ji}(\mathbf{s}_n))^t$  ( $j = 1, 2$  y  $i = 1, \dots, n$ ) y  $\delta_j$  es un vector  $n \times 1$  de coeficientes espectrales (proyecciones del proceso  $\mathbf{z}_j$  sobre las funciones base) con función de distribución  $\delta_j \sim MN(\mathbf{0}, \Sigma_{\delta_j})$ . Observe que si las funciones base son ortogonales, entonces  $\delta_j = \Psi_j^t \mathbf{z}_j$  y  $\Sigma_{\delta_j} = \Psi_j^t \Sigma_{\mathbf{z}_j} \Psi_j$  porque  $\Psi_j^t \Psi_j = \Psi_j \Psi_j^t = \mathbf{I}$ .

Existen varias ventajas al escribir el proceso espacial  $\mathbf{z}_j$  en términos del proceso espectral  $\delta_j$ . La primera es que el operador espectral frecuentemente actúa como una forma de eliminar las correlaciones en  $\Sigma_{\mathbf{z}_j}$ ; de este modo, las correlaciones en  $\Sigma_{\delta_j}$  son relativamente bajas. Otros beneficios son: la eficiencia computacional de esta descomposición y que en algunos casos se logra una reducción de la dimensión; la formulación de la base se puede truncar (Royle & Wikle 2005).

Por último, utilizando la transformación Box-Cox, sustituyendo (4.6) en el modelo (4.4), se tienen los siguientes modelos

$$\begin{aligned} M_1 = g_1(\boldsymbol{\mu}_s) &= \mathbf{V}_s \boldsymbol{\zeta} + \mathbf{E}_1 \Psi_1 \boldsymbol{\delta}_1 & M_2 = g_{\nu_2}(\boldsymbol{\phi}_s) &= \mathbf{U}_s \boldsymbol{\varphi} + \mathbf{E}_2 \Psi_2 \boldsymbol{\delta}_2 \\ &= \mathbf{V}_s \boldsymbol{\zeta} + \mathbf{E}_1^* \boldsymbol{\delta}_1 & &= \mathbf{U}_s \boldsymbol{\varphi} + \mathbf{E}_2^* \boldsymbol{\delta}_2 \end{aligned} \quad (4.7)$$

donde  $\mathbf{E}_j^* = \mathbf{E}_j \Psi_j$ ,  $j = 1, 2$ .

### 4.2.1 Algoritmo de máxima verosimilitud con Monte Carlo para el BSLMM

La aplicación de métodos basados en verosimilitud para SGLMMs basado en respuesta beta es obstaculizado por las dificultades computacionales que surgen de la alta dimensionalidad de los vectores aleatorios no observados,  $\boldsymbol{\delta}_1$  y  $\boldsymbol{\delta}_2$ . En esta sección se considera la máxima verosimilitud a través de MCMC (Geyer & Thompson 1992, Geyer 1994, Højbjerg 2003, Christensen 2004) para BSLMM.

Así como se realizó anteriormente, considérese  $n$  distintas localizaciones  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$  y supóngase que se observa una realización  $\mathbf{y}_s = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))^t$  de  $\mathbf{Y}_s = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))^t$ . Así,  $M_1 = (M_1(\mathbf{s}_1), \dots, M_1(\mathbf{s}_n))^t$  sigue una distribución multinormal con vector de media  $\mathbf{V}_s \boldsymbol{\zeta}$  y matriz de covarianza  $\sigma_1^2 R(\vartheta_1) + \tau_1^2 \mathbf{I}_n$ , donde  $R(\vartheta_1)$  es la matriz de correlación con entradas  $R_{ij}(\vartheta_1) = \rho(\mathbf{s}_i, \mathbf{s}_j; \vartheta_1)$ . Similarmente,  $M_2 = (M_2(\mathbf{s}_1), \dots, M_2(\mathbf{s}_n))^t \sim MN(\mathbf{U}_s \boldsymbol{\varphi}; \sigma_2^2 R(\vartheta_2) + \tau_2^2 \mathbf{I}_n)$ .

Por independencia de  $Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n)$  dados  $M_1, M_2$  y (4.5), la densidad condicional de  $\mathbf{Y}_s$  dada  $M_1 = \boldsymbol{\eta}_1$  y  $M_2 = \boldsymbol{\eta}_2$  está dada por

$$f(\mathbf{y}_s \mid \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \nu_2) = \prod_{i=1}^n f\{y(\mathbf{s}_i) \mid g_1^{-1}[\boldsymbol{\eta}_1(\mathbf{s}_i)], m_{i2} g_{\nu_2}^{-1}[\boldsymbol{\eta}_2(\mathbf{s}_i)]\}$$

donde  $m_{i2} = m_2(\mathbf{s}_i)$  y  $\nu_2$  determina la función de enlace para el SVDM.

Ahora, desde una perspectiva clásica, la función de verosimilitud basada en las variables aleatorias observadas  $\mathbf{y}_s$  se obtiene marginalizando con respecto a las variables aleatorias no observadas  $\boldsymbol{\delta}_1$  y  $\boldsymbol{\delta}_2$ , llevando a la obtención de la verosimilitud del modelo mixto. Entonces, la función de verosimilitud para un BSLMM no se puede expresar en forma cerrada, sino sólo como una doble integral de alta dimensionalidad como se presenta a continuación

$$\begin{aligned} L(b, \nu_2) &= f(\mathbf{y}_s | b, \nu_2) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} f(\mathbf{y}_s | \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \nu_2) f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{V}_s, \mathbf{U}_s; b) d\boldsymbol{\eta}_1 d\boldsymbol{\eta}_2 \\ &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \prod_{i=1}^n f(y(\mathbf{s}_i) | g_1^{-1}[\eta_1(\mathbf{s}_i)], m_{i2} g_{\nu_2}^{-1}[\eta_2(\mathbf{s}_i)]) f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{V}_s, \mathbf{U}_s; b) \\ &\quad d\eta_1(\mathbf{s}_1) \cdots d\eta_1(\mathbf{s}_n) d\eta_2(\mathbf{s}_1) \cdots d\eta_2(\mathbf{s}_n) \end{aligned} \quad (4.8)$$

donde  $b = (\boldsymbol{\zeta}, \boldsymbol{\varphi}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ ,  $f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{V}_s, \mathbf{U}_s; b)$  denota la función de distribución conjunta de  $(M_1, M_2)$  dada las covariables observadas  $\mathbf{V}_s$  y  $\mathbf{U}_s$ , con  $\boldsymbol{\theta}_1$  y  $\boldsymbol{\theta}_2$  los vectores de parámetros de covarianza asociados a  $\boldsymbol{\delta}_1$  y  $\boldsymbol{\delta}_2$ , respectivamente. La doble integral anterior es también la constante de normalización en la función de densidad condicional  $(M_1, M_2)$  dado  $\mathbf{Y}_s = \mathbf{y}_s$ ,

$$\begin{aligned} f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{y}_s, b, \nu_2) &\propto \prod_{i=1}^n f\{y(\mathbf{s}_i) | g_1^{-1}[\eta_1(\mathbf{s}_i)], m_{i2} g_{\nu_2}^{-1}[\eta_2(\mathbf{s}_i)]\} \\ &\quad \times f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{V}_s, \mathbf{U}_s; b) \end{aligned} \quad (4.9)$$

MCMC proporciona un método para la simulación de (4.9) y aproximación de (4.8).

Las dos integrales tienen una alta dimensión, y en consecuencia, son intratables para encontrar los MLEs por maximización directa. Entonces, se considera el caso donde  $\nu_2$  es fijo, por lo cual se puede suprimir. La función de verosimilitud presentada en (4.8) se puede reescribir como

$$\begin{aligned} L(b) &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} f(\mathbf{y}_s | \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{V}_s, \mathbf{U}_s; b) d\boldsymbol{\eta}_1 d\boldsymbol{\eta}_2 \\ &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \frac{f(\mathbf{y}_s | \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{V}_s, \mathbf{U}_s; b)}{\tilde{f}(\mathbf{y}_s, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)} \tilde{f}(\mathbf{y}_s, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) d\boldsymbol{\eta}_1 d\boldsymbol{\eta}_2 \\ &\propto \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \frac{f(\mathbf{y}_s | \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{V}_s, \mathbf{U}_s; b)}{f(\mathbf{y}_s | \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \tilde{f}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)} \tilde{f}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{y}_s) d\boldsymbol{\eta}_1 d\boldsymbol{\eta}_2 \\ &= \tilde{\mathbb{E}} \left[ \frac{f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{V}_s, \mathbf{U}_s; b)}{\tilde{f}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)} \middle| \mathbf{y}_s \right] \end{aligned} \quad (4.10)$$

donde  $\tilde{f}(\mathbf{y}_s, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = f(\mathbf{y}_s | \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \tilde{f}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$  y  $\tilde{f}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$  es alguna función de densidad con soporte en  $\mathbb{R}^n \times \mathbb{R}^n$ , la función de densidad condicional  $\tilde{f}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \mathbf{y}_s) \propto f(\mathbf{y}_s | \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \tilde{f}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$  y  $\tilde{\mathbb{E}}(\cdot | \mathbf{y}_s)$  denota la esperanza con respecto a

$\tilde{f}(\cdot | \mathbf{y}_s)$  y depende de una estimación inicial de  $b$ . Los MLEs se pueden calcular mediante la maximización de la aproximación de Monte Carlo presentada en (4.10),

$$L_r(b) = \frac{1}{r} \sum_{k=1}^r \frac{f(\boldsymbol{\eta}_1(k), \boldsymbol{\eta}_2(k) | \mathbf{V}_s, \mathbf{U}_s; b)}{\tilde{f}(\boldsymbol{\eta}_1(k), \boldsymbol{\eta}_2(k))} \quad (4.11)$$

donde  $\boldsymbol{\eta}_1(k)$  y  $\boldsymbol{\eta}_2(k)$  ( $k = 1, \dots, r$ ) son muestreados por MCMC a partir de la función de distribución  $\tilde{f}(\cdot | \mathbf{y}_s)$ . Como se ha señalado en (4.10), se debe elegir  $\tilde{f}(\cdot)$  cercano a  $f(\cdot | \hat{b})$ , donde  $\hat{b}$  es el MLE de  $b$ , porque de lo contrario uno o muy pocos de los términos  $f(\boldsymbol{\eta}_1(k), \boldsymbol{\eta}_2(k) | \mathbf{V}_s, \mathbf{U}_s; b) / \tilde{f}(\boldsymbol{\eta}_1(k), \boldsymbol{\eta}_2(k))$ ,  $k = 1, \dots, r$  pueden dominar a los otros en  $L_r(b)$ , lo cual hace la aproximación menos útil.

Ahora, se presenta un procedimiento numérico por maximización de la aproximación de Monte Carlo (4.11). Por lo tanto, si  $\boldsymbol{\theta}_j = (\sigma_j^2, \Delta_j)$  para  $j = 1, 2$ , entonces  $b = (\boldsymbol{\zeta}, \boldsymbol{\varphi}, \sigma_1^2, \sigma_2^2, \boldsymbol{\Delta})$  donde  $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$  y cada  $\boldsymbol{\Delta}_j$  denota la función de correlación con valores de  $\vartheta_j$  y  $\tau_j^2$ ,  $j = 1, 2$ . La maximización de  $L_r$  con respecto a  $\boldsymbol{\zeta}$ ,  $\boldsymbol{\varphi}$ ,  $\sigma_1^2$  y  $\sigma_2^2$  dado  $\boldsymbol{\Delta}$  es bastante sencillo porque las derivadas de primer y segundo orden de la función de densidad normal  $f(\boldsymbol{\eta}_1(k), \boldsymbol{\eta}_2(k) | \mathbf{V}_s, \mathbf{U}_s, b)$ ,  $k = 1, \dots, r$ , con respecto a esos parámetros son simples, lo que hace un proceso iterativo como el de Newton-Raphson factible y computacionalmente rápido. Para este método iterativo se pueden tomar los siguientes valores iniciales

$$\begin{pmatrix} \boldsymbol{\zeta}(k) \\ \boldsymbol{\varphi}(k) \end{pmatrix} = \left[ \begin{pmatrix} \mathbf{V}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_s \end{pmatrix}^t \mathbf{C}_{M_1, M_2}^{-1} \begin{pmatrix} \mathbf{V}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_s \end{pmatrix} \right]^{-1} \begin{pmatrix} \mathbf{V}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_s \end{pmatrix}^t \\ \times \mathbf{C}_{M_1, M_2}^{-1} \begin{pmatrix} \boldsymbol{\eta}_1(k) \\ \boldsymbol{\eta}_2(k) \end{pmatrix}$$

$$\sigma_1^2(k) = \frac{1}{n} [\boldsymbol{\eta}_1(k) - \mathbf{V}_s \boldsymbol{\zeta}(k)]^t [\mathbf{R}(\vartheta_1) + \tau_{R_1}^2 \mathbf{I}_n]^{-1} [\boldsymbol{\eta}_1(k) - \mathbf{V}_s \boldsymbol{\zeta}(k)]$$

$$\sigma_2^2(k) = \frac{1}{n} [\boldsymbol{\eta}_2(k) - \mathbf{U}_s \boldsymbol{\varphi}(k)]^t [\mathbf{R}(\vartheta_2) + \tau_{R_2}^2 \mathbf{I}_n]^{-1} [\boldsymbol{\eta}_2(k) - \mathbf{U}_s \boldsymbol{\varphi}(k)]$$

$k = 1, \dots, r$ , los cuales corresponde a los estimadores de máxima verosimilitud para las funciones de densidad normal  $f(\boldsymbol{\eta}_1(k), \boldsymbol{\eta}_2(k) | \mathbf{V}_s, \mathbf{U}_s, b)$  y donde

$$\mathbf{C}_{M_1, M_2} = \begin{pmatrix} \sigma_1^2 (\mathbf{R}(\vartheta_1) + \tau_{R_1}^2 \mathbf{I}_n) & \text{Cov}(M_1, M_2) \\ \text{Cov}^t(M_1, M_2) & \sigma_2^2 (\mathbf{R}(\vartheta_2) + \tau_{R_2}^2 \mathbf{I}_n) \end{pmatrix}, \quad (4.12)$$

donde  $\sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)$  es la matriz de covarianza de  $\boldsymbol{\eta}_j$  ( $j = 1, 2$ ) y  $\tau_{R_j}^2 = \tau_j^2 / \sigma_j^2$  es una pepita relativa en lugar de  $\tau_j^2$  para  $j = 1, 2$ . Además,  $\text{Cov}(M_1, M_2)$  es la covarianza cruzada entre  $M_1$  y  $M_2$ .

Los valores de  $\boldsymbol{\zeta}$ ,  $\boldsymbol{\varphi}$ ,  $\sigma_1^2$  y  $\sigma_2^2$  que maximizan  $L_r(b)$  para un valor fijo de  $\boldsymbol{\Delta}$ ,  $\hat{\boldsymbol{\zeta}}(\boldsymbol{\Delta})$ ,  $\hat{\boldsymbol{\varphi}}(\boldsymbol{\Delta})$ ,  $\hat{\sigma}_1^2(\boldsymbol{\Delta})$  y  $\hat{\sigma}_2^2(\boldsymbol{\Delta})$  están conectados en  $L_r$ , y se obtiene



$\tilde{L}_r(\Delta) = L_r(\hat{\zeta}(\Delta), \hat{\varphi}(\Delta), \hat{\sigma}_1^2(\Delta), \hat{\sigma}_2^2(\Delta))$ . Esta función es maximizada con respecto a  $\vartheta_1, \vartheta_2, \tau_{R_1}^2$  y  $\tau_{R_2}^2$  para una función de correlación dada utilizando optimización numérica. Los parámetros  $\vartheta_1, \vartheta_2, \tau_{R_1}^2$  y  $\tau_{R_2}^2$  entran en  $\tilde{L}_r$  via la matriz  $\mathbf{C}_{M_1, M_2}$ , y ya que la inversa de la matriz es computacionalmente exigente, la maximización podría ser relativamente lenta. La maximización también puede ser sensible a valores iniciales en este proceso porque la aproximación  $L_r$  puede ser multimodal. En este sentido, el resultado debería ser investigado cuidadosamente considerando una variedad de valores iniciales.

Por otro lado, cuando el interés es investigar cuál función de enlace es apropiada en el SVDM, es útil integrar con respecto a  $\phi_s = (\phi(\mathbf{s}_1), \dots, \phi(\mathbf{s}_n))^t$ , donde  $\phi(\mathbf{s}_i) = m_{i2}g_{\nu_2}^{-1}(\eta_2(\mathbf{s}_i)), i = 1, \dots, n$ . El determinante del Jacobiano para esta transformación es

$$J_{\nu_2}(\phi_s) = \prod_{i=1}^n \frac{g'_{\nu_2}(\phi(\mathbf{s}_i)/m_{i2})}{m_{i2}}$$

Asumiendo que la función de enlace satisface  $g_{\nu_2}(\mathbb{R}) = \mathbb{R}$  para cada  $\nu_2$  de interés y definiendo

$$f_{\nu_2}(\mu_s, \phi_s \mid \mathbf{V}_s, \mathbf{U}_s, \delta_1, \delta_2, b) = J_{\nu_2}(\phi_s) f(g_1^{-1}(\eta_1), g_{\nu_2}^{-1}(\phi_s) \mid \mathbf{V}_s, \mathbf{U}_s, \delta_1, \delta_2, b)$$

se obtiene

$$\begin{aligned} L(\mu_s, \phi_s, \nu_2) &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} f(\mathbf{y}_s \mid \mu_s, \phi_s) f_{\nu_2}(\mu_s, \phi_s \mid \mathbf{V}_s, \mathbf{U}_s, \delta_1, \delta_2, b) d\mu_s d\phi_s \\ &\propto \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \frac{f(\mathbf{y}_s \mid \mu_s, \phi_s) f_{\nu_2}(\mu_s, \phi_s \mid \mathbf{V}_s, \mathbf{U}_s, \delta_1, \delta_2, b)}{f(\mathbf{y}_s \mid \mu_s, \phi_s) \tilde{f}_{(\nu_2)_0}(\mu_s, \phi_s)} \\ &\quad \times \tilde{f}_{(\nu_2)_0}(\mu_s, \phi_s \mid \mathbf{y}_s) d\mu_s d\phi_s \\ &= \tilde{\mathbb{E}} \left[ \frac{f_{\nu_2}(\mu_s, \phi_s \mid \mathbf{V}_s, \mathbf{U}_s, \delta_1, \delta_2, b)}{\tilde{f}_{(\nu_2)_0}(\mu_s, \phi_s)} \mid \mathbf{y}_s \right] \end{aligned} \quad (4.13)$$

donde  $\tilde{f}_{(\nu_2)_0}(\mu_s, \phi_s) = J_{(\nu_2)_0}(\phi_s) \tilde{f}(g_1^{-1}(\eta_1), g_{\nu_2}^{-1}(\phi_s))$ , la función de densidad condicional es  $\tilde{f}_{(\nu_2)_0}(\mu_s, \phi_s \mid \mathbf{y}_s) \propto f(\mathbf{y}_s \mid \mu_s, \phi_s) \tilde{f}_{(\nu_2)_0}(\mu_s, \phi_s)$  y  $\tilde{\mathbb{E}}(\cdot \mid \mathbf{y}_s)$  denota la esperanza con respecto a  $\tilde{f}_{(\nu_2)_0}(\cdot \mid \mathbf{y}_s)$ . El MLE se puede calcular maximizando la aproximación de Monte Carlo en (4.13) como

$$L(\mu_s, \phi_s, \nu_2) = \frac{1}{r} \sum_{k=1}^r \frac{f_{\nu_2}(\mu_s(k), \phi_s(k) \mid \mathbf{V}_s, \mathbf{U}_s, \delta_1, \delta_2, b)}{\tilde{f}_{(\nu_2)_0}(\mu_s(k), \phi_s(k))} \quad (4.14)$$

donde el  $\mu_s(k)$  y  $\phi_s(k)$  ( $k = 1, \dots, r$ ) son muestreados por MCMC a partir de la función de distribución  $\tilde{f}(\cdot \mid \mathbf{y}_s)$ . Si  $g_{\nu_2}(\mathbb{R}) \neq \mathbb{R}$  entonces se mantiene (4.13).

Una vez estimados los parámetros de tendencia y de correlación espacial,  $b$ , se esta preparado para discutir las medidas espaciales de bondad de ajuste.

### 4.2.2 Medidas de bondad de ajuste

Después de ajustar el BSLMM es importante llevar a cabo un análisis de diagnóstico para verificar la bondad de ajuste del modelo estimado. Una medida global de la variación explicada se obtiene calculando el pseudo  $R_s^2$  definida como

$$R_s^2 = \frac{l(\tilde{b}) - l(\hat{b})}{l(\tilde{b})}, \quad 0 \leq R_s^2 \leq 1$$

donde  $l(\tilde{b})$  es la función de log-verosimilitud para el modelo saturado evaluado en  $\tilde{b}$  y  $l(\hat{b})$  denota el valor máximo de la función de log-verosimilitud para el modelo de interés utilizando MCMC. Observe que  $l(\tilde{b})$  será más grande que cualquier otra función de verosimilitud para las observaciones medidas ya que proporciona la más completa descripción de los datos, asumiendo la misma función de distribución y función enlace.

La discrepancia del modelo ajustado se puede determinar a través de lo bien que el modelo ajustado es significativamente diferente del modelo saturado, que contiene tantos parámetros como observaciones hay en el modelo. Para este fin, sea

$$D(\mathbf{y}_s, b) = 2 [l(\hat{b}) - l(\tilde{b})]$$

Una aproximación a esta cantidad se puede expresar como

$$D(\mathbf{y}_s, b) = \sum_{i=1}^n (r(\mathbf{s}_i))^2 \quad (4.15)$$

el cual es conocida como la *pseudo-deviance* y

$$r_i = r(\mathbf{s}_i) = \text{sign}[y(\mathbf{s}_i) - \hat{y}(\mathbf{s}_i)] \{2 [l(y(\mathbf{s}_i), \hat{b}) - l(y(\mathbf{s}_i), \tilde{b})]\}^{1/2}$$

donde  $l(y(\mathbf{s}_i), \tilde{b})$  es la máxima log-verosimilitud para el modelo saturado asociado a la  $i$ -ésima observación y  $l(y(\mathbf{s}_i), \hat{b})$  es el máximo valor de la función de la log-verosimilitud para el modelo de interés asociado a la  $i$ -ésima observación.  $r(\mathbf{s}_i)$  es la  $i$ -ésima deviance residual porque una observación con un valor absoluto grande de  $r(\mathbf{s}_i)$  se puede ver como discrepancia. Como se esperaba, la log-verosimilitud asociada con el modelo saturado debe ser más grande que la del modelo con  $p_1 < n$  parámetros.

A continuación se discuten las técnicas espaciales para predecir el valor de un campo aleatorio en una localización determinada a partir de las observaciones cercanas.

### 4.2.3 Predicción espacial de nuevos individuos

En este caso, al igual que en la Sección 3.3.1, se utiliza la técnica de kriging, la cual es una técnica mediante la cual se puede interpolar el vector de valores  $\mathbf{y}_0 = (y(\mathbf{s}_{n+1}), \dots, y(\mathbf{s}_{n+n'}))^t$  de un vector en el campo aleatorio  $\mathbf{Y}_0$  en  $n'$  pre-especificadas localizaciones a partir de las observaciones  $y(\mathbf{s}_i)$ ,  $i = 1, \dots, n$ .

El enfoque se centra en la interpolación de efectos aleatorios sobre un área espacial continua cuando las observaciones son una proporción o tasa. Por consiguiente, se puede considerar la predicción funcional para el SMM de  $\boldsymbol{\eta}_1^0 = (\eta_1(\mathbf{s}_{n+1}), \dots, \eta_1(\mathbf{s}_{n+n'}))^t$  y el SVDM de  $\boldsymbol{\eta}_2^0 = (\eta_2(\mathbf{s}_{n+1}), \dots, \eta_2(\mathbf{s}_{n+n'}))^t$ . En los dos modelos, sea  $f(\boldsymbol{\eta}_j^0, \boldsymbol{\eta}_j)$  la función de densidad conjunta de  $\boldsymbol{\eta}_j$  y un vector  $\boldsymbol{\eta}_j^0$ ,  $j = 1, 2$ . Limitando el interés a los predictores lineales pseudo insesgados de la forma

$$\tilde{\boldsymbol{\eta}}_j = \mathbf{p}_j + \mathbf{Q}_j \boldsymbol{\eta}_j, \quad j = 1, 2 \quad (4.16)$$

para algún vector conformable  $\mathbf{p}_j$  y la matriz  $\mathbf{Q}_j$  (McCulloch et al. 2008), y donde  $\tilde{\boldsymbol{\eta}}_j$  es el predictor de  $\boldsymbol{\eta}_j^0$ ; así, minimizando el error cuadrático medio de la predicción, se encuentran los mejores pseudo predictores lineales insesgados, los cuales están dados por (ver detalles en el Apéndice 4.2)

$$\tilde{\boldsymbol{\eta}}_j = \mathbf{E}(\boldsymbol{\eta}_j^0) + \text{Cov}^t(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) [\text{Var}(\boldsymbol{\eta}_j)]^{-1} [\boldsymbol{\eta}_j - \mathbf{E}(\boldsymbol{\eta}_j)], \quad j = 1, 2$$

Además, la matriz de covarianzas para la predicción tiene la siguiente forma general (ver detalles en el Apéndice 4.2)

$$\begin{aligned} \text{Var}(\tilde{\boldsymbol{\eta}}_j | \mathbf{y}_s) &\approx \boldsymbol{\Sigma}_j + \text{Cov}^t(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) [\sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)]^{-1} [\widetilde{\text{Var}}_r(\boldsymbol{\eta}_j | \mathbf{y}_s)] \\ &\quad \times [\sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)]^{-1} \text{Cov}(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) \end{aligned}$$

para  $j = 1, 2$ .

## 4.3 Experimento simulado

En esta simulación, se presenta un estudio de simulación para analizar la elección de valores iniciales y la determinación del tamaño de muestra de Monte Carlo. Se simularon 100 puntos en una rejilla espacial regular de  $(0, 1) \times (0, 1)$  utilizando dos variables aleatorias explicativas: la primera es una variable continua que se muestreó de una distribución normal con media 1 y desviación estándar 4 y la segunda es una variable categórica muestreada de una función de distribución multinomial asumiendo tres valores con probabilidades 0.45, 0.25 y 0.3. Como existen tres niveles en la variable categórica, únicamente se consideran dos variables dummy ( $V_2$  y  $V_3$ ) para evitar el problema de singularidad. Para cada una de las variables explicativas se generaron 100

observaciones independientes. Condicionando sobre los procesos estacionarios gaussianos  $z_1(\mathbf{s}_i)$  y  $z_2(\mathbf{s}_i)$ , la variable respuesta  $Y$  se simuló de una función de distribución beta con  $p(\mathbf{s}_i) = \mu(\mathbf{s}_i)\phi(\mathbf{s}_i)$  y  $q(\mathbf{s}_i) = \phi(\mathbf{s}_i)(1 - \mu(\mathbf{s}_i))$ , donde el SMM y el SVDM están dados por

$$\mu(\mathbf{s}_i) = \{1 + \exp\{-[1.3 + 0.01v_1(\mathbf{s}_i) + 0.4v_2(\mathbf{s}_i) - 0.7v_3(\mathbf{s}_i) + 1.6w_1(\mathbf{s}_i) - 1.7w_2(\mathbf{s}_i) + z_1(\mathbf{s}_i)]\}\}^{-1} \quad (4.17)$$

$$\phi(\mathbf{s}_i) = \exp[2.1 - 0.23u_1(\mathbf{s}_i) - 0.8u_2(\mathbf{s}_i) + 0.5u_3(\mathbf{s}_i) + 2.3w_1(\mathbf{s}_i) - 1.1w_2(\mathbf{s}_i) + z_2(\mathbf{s}_i)] \quad (4.18)$$

$i = 1, \dots, 100$ , y donde  $w_1(\mathbf{s}_i)$  y  $w_2(\mathbf{s}_i)$  son las  $i$ -ésimas coordenadas espaciales,  $v_3(\mathbf{s}_i) = u_3(\mathbf{s}_i) \sim N(1, 4)$ ,  $v_2(\mathbf{s}_i) = u_2(\mathbf{s}_i)$  y  $v_3(\mathbf{s}_i) = u_3(\mathbf{s}_i)$  son dos variables dummies asociadas a la variable categórica, y  $z_1(\mathbf{s}_i)$  y  $z_2(\mathbf{s}_i)$  siguen dos procesos estacionarios gaussiano isotrópicos con variogramas esféricos dados por  $\gamma_1(h_1) = 1 + 4 \left(1.5 \left(\frac{h_1}{0.4}\right) - \frac{1}{2} \left(\frac{h_1}{0.4}\right)^3\right)$  y  $\gamma_2(h_2) = 5 + 10 \left(1.5 \left(\frac{h_2}{0.3}\right) - \frac{1}{2} \left(\frac{h_2}{0.3}\right)^3\right)$ , respectivamente, para  $h_1 > 0$ ,  $h_2 > 0$  y  $\gamma_1(0) = \gamma_2(0) = 0$ .

En primer lugar se ajustan los datos a un BSLMM utilizando la función de enlace logit con efectos aleatorios independientes e igualmente distribuidos en el SMM y función de enlace log con efectos aleatorios independientes e igualmente distribuidos en el SVDM. Para la estimación de máxima verosimilitud, se utiliza la aproximación presentada en la Sección 4.2 con  $\tilde{f}(\cdot)$  igual a  $f(\cdot | b_0)$ , donde  $b_0 = (\boldsymbol{\zeta}_0, \boldsymbol{\varphi}_0, \boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02})^t$  con  $\boldsymbol{\zeta}_0 = (0.6, -0.04, 0.07, -1, 3, -2)^t$ ,  $\boldsymbol{\varphi}_0 = (0, -0.08, -1.5, -0.3, 4, 0.2)^t$ ,  $\boldsymbol{\theta}_{01} = (10, 0.5, 3)^t$  y  $\boldsymbol{\theta}_{02} = (10, 0.3, 5)^t$ . Luego de haber hecho una ejecución inicial del algoritmo MCMC para determinar los valores apropiados de los parámetros  $b_0$ , se presentan los resultados utilizando una función de correlación esférica tanto para el SMM como para el SVDM. El tamaño de muestra por Monte Carlo fue de 1000, y las primeras 10000 iteraciones fueron descartados ya que fue tomado como período de entrenamiento de la cadena de Markov. Un total de 1000000 de simulaciones se ejecutaron, tomando muestras cada 1000 iteraciones. Se estudiaron los siguientes modelos:

- $H_1$ : No correlación espacial ni en el modelo de media ni en el modelo de dispersión variable.
- $H_2$ : No correlación espacial en el modelo de media pero con función de correlación espacial esférica en el modelo de dispersión variable.
- $H_3$ : Función de correlación esférica en el modelo de media pero no hay correlación espacial en el modelo de dispersión variable.
- $H_4$ : Función de correlación esférica en el modelo de media con  $\vartheta_1 > 0$  y  $\tau_1^2 = 0$ , y función de correlación esférica en el modelo de dispersión variable con  $\vartheta_2 > 0$  y  $\tau_2^2 = 0$ .

- $H_5$ : Función de correlación esférica en el modelo de media con  $\vartheta_1 > 0$  y  $\tau_1^2 \geq 0$ , y función de correlación esférica en el modelo de dispersión variable con  $\vartheta_2 > 0$  y  $\tau_2^2 \geq 0$ .
- $H_6$ : Función de correlación esférica en el modelo de media con  $\vartheta_1 > 0$  y  $\tau_1^2 \geq 0$ , y función de correlación gaussiana en el modelo de dispersión variable con  $\vartheta_2 > 0$  y  $\tau_2^2 \geq 0$ .
- $H_7$ : Función de correlación gaussiana en el modelo de media con  $\vartheta_1 > 0$  y  $\tau_1^2 \geq 0$ , y función de correlación exponencial en el modelo de dispersión con  $\vartheta_2 > 0$  y  $\tau_2^2 \geq 0$ .
- $H_8$ : Función de correlación en el modelo de media con  $\vartheta_1 > 0$  y  $\tau_1^2 \geq 0$ , y función de correlación esférica en el modelo de dispersión variable con  $\vartheta_2 > 0$  y  $\tau_2^2 \geq 0$ .

Los resultados obtenidos luego de ejecutar el algoritmo MCMC para el BSLMM se presentan en la Tabla 4.1. Al comparar la diferencia entre los  $\log \hat{L}$ 's para  $H_5$  y  $H_1$  con la distribución  $0.5\chi_{(6)}^2$  (cuyo cuantil 95 % es 6.296), se encuentra que hay una clara correlación espacial en los dos modelos: de media y dispersión variable. Además, cuando se compara la diferencia de los  $\log \hat{L}$ 's en  $H_3$  y  $H_2$  con respecto a  $H_1$ , se encuentra una evidente correlación espacial en el SVDM y el SMM, respectivamente ( $0.5\chi_{(3)}^2 = 3.907$ ). Sin embargo, si se compara  $H_5$  con respecto a  $H_3$  y  $H_2$ , se observa que es necesario la correlación espacial en el SMM y el SVDM para construir el BSLMM. En esta simulación, hay presencia de efecto pepita, el cual se observa mediante la comparación de las diferencias de  $\log \hat{L}$ 's de  $H_5$  y  $H_4$  con respecto a la distribución  $0.5\chi_{(1)}^2$  (cuyo cuantil 95 % es 2.996). Observe que las estimaciones de  $\hat{\vartheta}_1$  y  $\hat{\vartheta}_2$  para  $H_4$  son más pequeñas que las estimaciones de  $\hat{\vartheta}_1$  y  $\hat{\vartheta}_2$  para  $H_5$ , lo cual es consistente con los modelos en  $H_4$  ya que no tienen efecto pepita mientras los modelos en  $H_5$  si lo tienen.

Los resultados en la Tabla 4.1 también muestran que la selección de la función de correlación no es importante, aunque las funciones de correlación esférica ( $H_5$ ) tanto en el SMM como en el SVDM son ligeramente mejores que otros modelos ajustados. De esta manera, este resultado es concordante con la simulación desarrollada. Varias ejecuciones del algoritmo MCMC por el enfoque BSLMM se realizaron con diferentes valores iniciales ( $b_0$ ) y diferentes funciones de correlación. Los resultados no difieren mucho y el patrón fue el mismo al de la Tabla 4.1. La comparación entre los modelos se hace de manera informal ya que algunos de los modelos que se compararon no son anidados. El análisis también asume que la razón de log-verosimilitud puede aproximarse razonablemente a una distribución  $\chi^2$ , pero esto es difícil de demostrar rigurosamente.

En la Tabla 4.2 se presentan los resultados de las log-verosimilitudes ensayando diferentes funciones de enlace para el SMM y el SVDM, y utilizando

TABLA 4.1: Estimaciones beta espaciales por máxima verosimilitud para la simulación utilizando los modelos  $H_1, H_2, H_3, H_4, H_5, H_6, H_7$  y  $H_8$ .

Modelo espacial para la media										
Modelo	$\hat{\zeta}_0$	$\hat{\zeta}_1$	$\hat{\zeta}_2$	$\hat{\zeta}_3$	$\hat{\zeta}_4$	$\hat{\zeta}_5$	$\hat{\sigma}_1^2$	$\hat{\vartheta}_1$	$\hat{\tau}_1^2$	
$H_1$	-0.851	0.056	0.157	-0.209	1.471	0.047	*	*	*	*
$H_2$	-0.125	0.060	0.074	-0.302	1.563	-0.142	*	*	*	*
$H_3$	-0.028	0.014	0.183	0.152	1.381	-0.955	1.792	0.415	0.183	
$H_4$	1.822	-0.016	0.450	-0.215	1.541	-2.134	5.117	0.406	0.000	
$H_5$	1.652	0.005	0.377	-0.578	1.513	-2.083	3.688	0.429	0.960	
	(0.002)	(0.000)	(0.001)	(0.002)	(0.002)	(0.002)	(0.005)	(0.000)	(0.002)	
$H_6$	1.659	0.005	0.376	-0.580	1.504	-2.089	3.695	0.429	0.966	
$H_7$	1.776	0.006	0.299	-0.705	1.423	-2.159	2.929	0.218	1.773	
$H_8$	1.705	0.003	0.402	-0.591	1.448	-2.323	3.419	0.218	1.900	

Modelo espacial con dispersión variable										
Modelo	$\hat{\varphi}_0$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{\varphi}_4$	$\hat{\varphi}_5$	$\hat{\sigma}_2^2$	$\hat{\vartheta}_2$	$\hat{\tau}_2^2$	$\log \hat{L}$
$H_1$	-1.391	0.021	-0.403	-0.102	1.059	0.846	*	*	*	199.23
$H_2$	-0.620	0.000	-0.350	0.204	1.034	0.206	0.385	0.696	0.344	216.24
$H_3$	-0.235	-0.018	-0.656	-4.161	0.693	0.551	*	*	*	225.56
$H_4$	2.034	-0.052	-0.632	-0.283	1.850	-1.336	7.558	0.193	0.000	234.37
$H_5$	2.784	-0.157	-0.765	0.199	1.891	-1.384	6.798	0.329	5.404	323.12
	(0.008)	(0.001)	(0.005)	(0.006)	(0.007)	(0.008)	(0.035)	(0.000)	(0.023)	
$H_6$	2.831	-0.154	-0.791	0.188	1.882	-1.437	5.199	0.174	6.847	322.71
$H_7$	2.731	-0.128	-0.756	0.280	1.853	-1.499	10.017	0.099	1.241	317.30
$H_8$	2.414	-0.071	-0.911	-0.097	2.116	-0.546	2.705	1.174	6.682	306.66

en todos los casos, las funciones de correlación esférica en ambos modelos. De acuerdo a esta tabla se seleccionó los modelos con enlace logit en el SMM y con enlace log en el SVDM para ajustar el BSLMM, ya que éste tiene la log-verosimilitud más alta (323.12); este resultado esta de acuerdo a la simulación

realizada.

TABLA 4.2: Log-verosimilitudes para la simulación utilizando diferentes funciones de enlace en el SMM y el SVDM, con funciones de correlación esférica en los dos modelos.

Modelo espacial con dispersión variable ( $\nu_2$ )	Modelo espacial para la media			
	loglog	logit	cloglog	probit
0.0 (log)	300.71	323.12	270.00	288.17
0.5 (raíz cuadrada)	254.30	270.78	252.87	237.83
0.8	249.27	245.54	235.80	243.58
1.0 (identidad)	245.54	241.91	233.03	240.22

La Tabla 4.1 presenta los parámetros estimados y las desviaciones estándar del modelo seleccionado ( $H_5$ ). Es posible ver que todas las estimaciones están de acuerdo con los respectivos valores verdaderos de los parámetros y todos ellos tienen desviaciones estándar pequeñas. Las Figuras 4.1 y 4.2 para los parámetros de media espacial y dispersión variable espacial, respectivamente, presentan las muestras de la cadena y su distribución, lo cual confirma los resultados de que todos los parámetros son significativos. Al juzgar el desempeño del MCMC, se debe tener en cuenta que un conjunto particular de los valores obtenidos de efectos aleatorios fueron utilizados en MCMC para ajustar el BSLMM en ambos modelos (SMM y SVDM). Considerando esto, las estimaciones MCMC son satisfactorias.

## 4.4 Aplicaciones

En esta sección, se vuelve a retomar las dos aplicaciones presentadas en la introducción: la primera es el contenido de arcilla en muestras de suelo tomadas de 0-10 cm de profundidad de la capa en cada una de las 147 observaciones del perfil de suelo, y la segunda, es el contenido de magnesio en muestras de suelo tomadas de 0-20 cm de profundidad de la capa en 178 localizaciones.

### 4.4.1 Contenido de arcilla

El conjunto de datos contiene 147 observaciones del perfil de suelo del área de investigación de la TCP (Yemefack et al. 2005). Específicamente, las medidas de estudio se hicieron en muestras de capas de suelo de 0-10 cm. Se desea investigar la relación entre el contenido de arcilla (peso en % de mineral de tierra fina, medida en  $mmol_c/dm^3$ ) y las variables explicativas: altura en metros

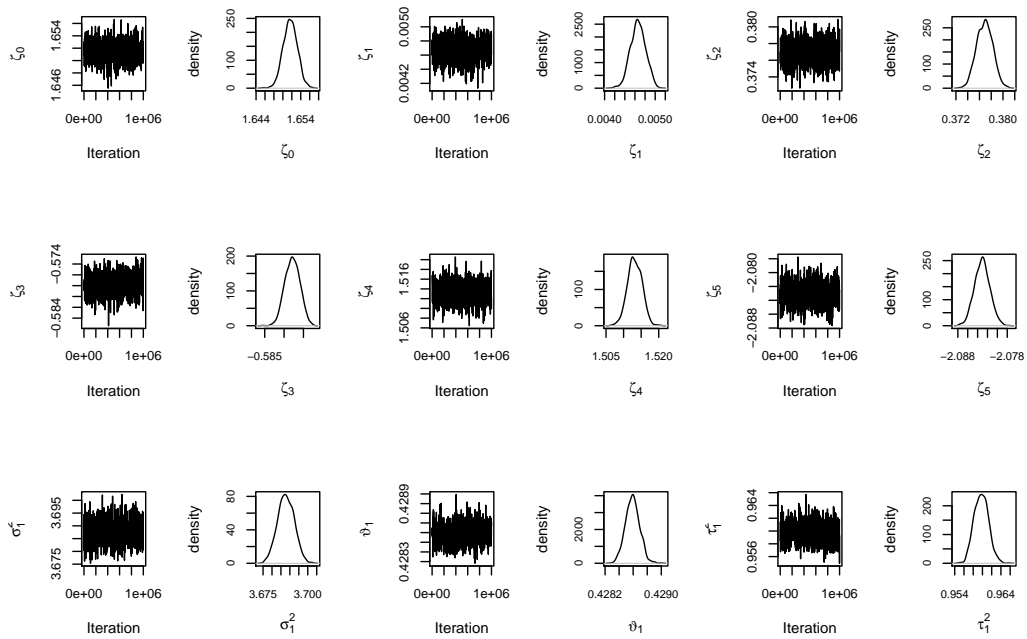


FIGURA 4.1: Comportamiento de la cadena muestreada para los parámetros del SMM, modelo (4.17).

sobre el nivel del mar (ELE), cuatro zonas agro-ecológicas (ZONE) y dos referencias del suelo (WRB) que son Acrisols y Ferralsols. Para cada observación, se registraron las coordenadas del este y norte (en metros).

En este sentido,  $y(\mathbf{s}_i)$  denota el contenido de arcilla en las localizaciones  $\mathbf{s}_i$ ,  $i = 1, \dots, n$ . En el BSLMM se asume que los  $y(\mathbf{s}_i)$ 's son variables beta condicionalmente independientes dados los procesos estocásticos espaciales no observados  $z_1(\mathbf{s}_i)$  y  $z_2(\mathbf{s}_i)$  para el SMM y el SVDM, respectivamente. Además, se asume que: el SMM en  $\mathbf{s}_i$  depende de las variables explicativas altura y zona agroecológica observadas en la localización  $\mathbf{s}_i$  sobre  $z_1(\mathbf{s}_i)$  y, el SVDM en  $\mathbf{s}_i$  depende de las variables explicativas zona agro-ecológica y grupo de referencia del suelo observado en la localización  $\mathbf{s}_i$  sobre  $z_2(\mathbf{s}_i)$ .

La Figura 4.3 muestra la relación entre las covariables potenciales y el contenido de arcilla. De acuerdo a esta figura, parece que hay una relación positiva entre el contenido de arcilla y la altura. El diagrama de cajas en la parte inferior izquierda (panel(c) de la Figura 4.3) sugiere que las medias de las distribuciones de contenido de arcilla son diferentes en las distintas zonas agro-ecológicas; lo mismo sucede con el diagrama de cajas de las dos referencias de grupos de suelo (parte inferior derecha, panel (d) de la Figura 4.3). No se incluyó la relación entre las localizaciones (E-W y N-S) y los contenidos de



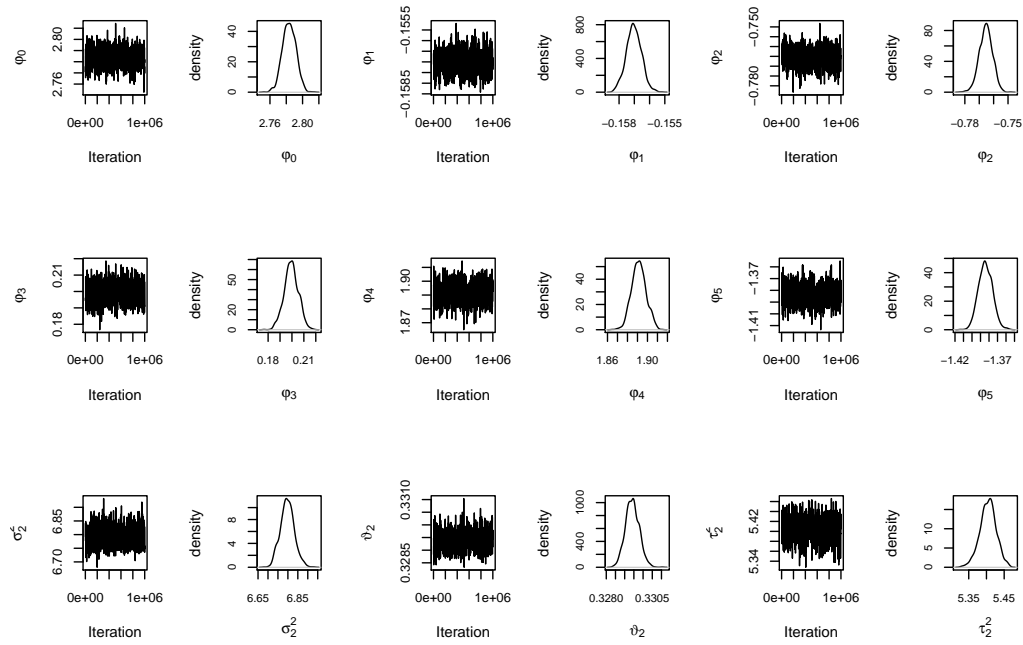


FIGURA 4.2: Comportamiento de la cadena muestreada para los parámetros del SVDM, modelo (4.18).

arcilla porque estás no fueron muy claras.

Adicionalmente, no se encontró alguna evidencia de anisotropía en el SMM y el SVDM; por lo cual se omitieron los detalles correspondientes. Por lo tanto, se restringe el estudio a funciones de correlación isotrópicas, donde la función de correlación  $\rho(h_j; \vartheta_j)$  depende únicamente de la distancia euclidiana  $h_j = \|\mathbf{s}_i - \mathbf{s}_{i'}\|_j$  entre ubicaciones,  $j = 1, 2$ ,  $i, i' = 1, \dots, n$ . Los  $z_1(\mathbf{s}_i)$  y  $z_2(\mathbf{s}_i)$  en los dos modelos tienen como finalidad capturar los residuales de la variación espacial en el SMM y el SVDM, respectivamente, después de ajustar por las variables explicativas.

Similar al caso simulado, primero se ajustan los datos mediante un BSLMM utilizando la función de enlace logit con efectos aleatorios independientes e igualmente distribuidos en el SMM y la función enlace log con efectos aleatorios independientes e igualmente distribuidos en el SVDM. Para la estimación de máxima verosimilitud, se utilizan  $b_0 = (\zeta_0, \varphi_0, \theta_{01}, \theta_{02})^t$  con  $\zeta_0 = (-0.9, 0.002, -0.4, -0.6, -0.9)^t$ ,  $\varphi_0 = (6, -1.7, -0.8, -2.2, -1.8)^t$ ,  $\theta_{01} = (0.3, 0.81, 0.1)^t$  y  $\theta_{02} = (0.2, 0.6, 0.1)^t$  para aproximar  $\tilde{f}(\cdot)$  a  $f(\cdot | b_0)$ . Una ejecución inicial del algoritmo MCMC sirve para determinar los valores apropiados de los parámetros  $b_0$ . El tamaño de muestra por Monte Carlo fue de 1000 y las primeras 10000 iteraciones fueron descartadas en el periodo de en-

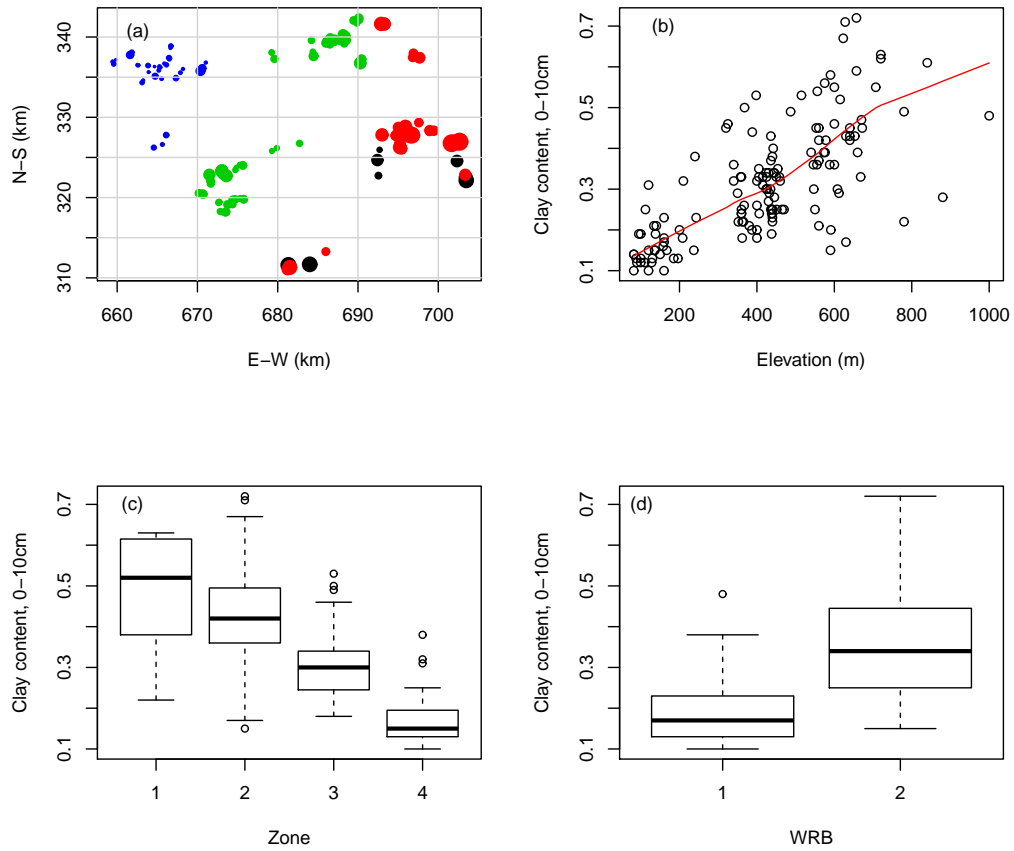


FIGURA 4.3: (a) Ubicaciones del contenido de arcilla, (b) Diagrama de dispersión del contenido de arcilla contra la altura, donde la recta representa la curva lowess, (c) Diagrama de cajas del contenido de arcilla en cada una de las cuatro zonas y (d) Diagrama de cajas del contenido de arcilla en cada uno de los dos grupos de referencia de suelo.

trenamiento, un total de 1000000 de simulaciones fueron ejecutadas usando una función de correlación esférica tanto para el SMM como para el SVDM, tomando muestras cada 1000 iteraciones. Se estudiaron los mismos modelos ( $H_1 - H_8$ ) que en el caso simulado.

En la Tabla 4.3, se presentan los resultados obtenidos después de ejecutar el algoritmo MCMC utilizando el BSLMM. Al comparar la diferencia entre los  $\log \hat{L}$ 's para  $H_5$  (o equivalentemente  $H_8$ ) y  $H_1$  (o equivalentemente  $H_3$ ) con la distribución  $0.5\chi_{(3)}^2$  (cuyo cuartil 95 % es 3.907), se observa que hay una ligera correlación espacial. Por lo tanto, se elige el modelo espacial  $H_5$  (o  $H_8$ ) porque hay una ligera correlación espacial en el modelo de dispersión variable, aunque

no está presente en el modelo de media. El efecto pepita en ambos modelos no fue significativo en ninguno de los modelos propuestos ( $H_1 - H_8$ ). La Tabla 4.3 también muestra que la elección de la función de correlación no es relevante, aunque el modelo con función de correlación esférica ( $H_5$  or  $H_8$ ) para la parte de dispersión variable es ligeramente mejor que los otros modelos. Además, los resultados no difieren mucho y el patrón es el mismo al de la Tabla 4.3 cuando se realizaron varias ejecuciones del algoritmo MCMC por el enfoque del BSLMM con diferentes valores iniciales ( $b_0$ ) y funciones de correlación.

TABLA 4.3: Estimaciones beta espaciales por máxima verosimilitud para el contenido de arcilla utilizando los modelos  $H_1, H_2, H_3, H_4, H_5, H_6, H_7$  y  $H_8$ .

Modelo espacial para la media								
Modelo	$\hat{\zeta}_0$	$\hat{\zeta}_1$	$\hat{\zeta}_2$	$\hat{\zeta}_3$	$\hat{\zeta}_4$	$\hat{\sigma}_1^2$	$\hat{\vartheta}_1$	$\hat{\tau}_1^2$
$H_1, H_3$	-1.878	0.002	0.541	0.343	0.045	0.000	0.000	0.000
$H_2, H_4$	-1.850	0.002	0.619	0.359	0.017	*	*	*
$H_5, H_8$	-2.085	0.002	0.630	0.457	0.212	0.000	0.000	0.000
$H_6$	-1.965	0.002	0.569	0.380	0.110	0.000	0.000	0.000
$H_7$	-2.140	0.002	0.655	0.490	0.266	0.000	0.000	0.000

Modelo espacial con dispersión variable									
Modelo	$\hat{\varphi}_0$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{\varphi}_4$	$\hat{\sigma}_2^2$	$\hat{\vartheta}_2$	$\hat{\tau}_2^2$	$\log \hat{L}$
$H_1, H_3$	3.100	1.123	2.000	0.793	-1.620	*	*	*	157.27
$H_2, H_4$	3.141	1.226	2.164	0.759	-1.743	0.048	1.334	0.000	158.86
$H_5, H_8$	2.932	1.170	2.434	1.265	-1.655	0.147	5.415	0.000	161.41
$H_6$	2.952	1.082	2.271	1.188	-1.569	0.097	1.209	0.000	160.72
$H_7$	2.958	1.173	2.402	1.292	-1.659	0.132	2.412	0.000	160.62

En la Tabla 4.4 se presentan los resultados de las log-verosimilitudes ensayando diferentes funciones de enlace para el SMM y el SVDM, y utilizando la función de correlación esférica en ambas partes del modelo. De acuerdo a los resultados de esta tabla, se elige el modelo con función de enlace logit para el SMM y el modelo con función de enlace log para SVDM para ajustar el BSLMM ya que éste tiene la log-verosimilitud más alta (161.41).

Luego, se asume que el BSLMM ajustado esta basado en los modelos espaciales mixtos dados por

$$\begin{aligned} \text{logit}(\hat{\mu}(\mathbf{s}_i)) &= \log\left(\frac{\hat{\mu}(\mathbf{s}_i)}{1 - \hat{\mu}(\mathbf{s}_i)}\right) = -2.085 + 0.002 \times ELEV(\mathbf{s}_i) + 0.630 \\ &\quad \times ZONE_2(\mathbf{s}_i) + 0.457 \times ZONE_3(\mathbf{s}_i) + 0.212 \times ZONE_4(\mathbf{s}_i) \\ \log(\hat{\phi}(\mathbf{s}_i)) &= 2.932 + 1.170 \times ZONE_2(\mathbf{s}_i) + 2.434 \times ZONE_3(\mathbf{s}_i) + 1.265 \\ &\quad \times ZONE_4(\mathbf{s}_i) - 1.655 \times WRB_2(\mathbf{s}_i) + \hat{z}_2(\mathbf{s}_i) \end{aligned}$$

TABLA 4.4: Log-verosimilitudes para el contenido de arcilla utilizando diferentes funciones de enlace en el SMM y el SVDM con función de correlación esférica en ambos modelos.

Modelo espacial con dispersión variable ( $\nu_2$ )	Modelo espacial para la media			
	loglog	logit	cloglog	probit
0.0 (log)	161.36	161.41	160.87	158.23
0.5 (raíz cuadrada)	157.13	159.23	158.60	159.39
0.7	157.77	156.81	158.49	159.17
1.0 (identidad)	155.18	157.10	158.49	157.43

donde  $\hat{y}(\mathbf{s}_i)$  es el contenido de arcilla en la ubicación  $i$ -ésima ( $i = 1, \dots, 147$ ),  $ELEV(\mathbf{s}_i)$  es la altura sobre el nivel del mar en la ubicación  $i$ -ésima,  $ZONE_j(\mathbf{s}_i)$  es la  $j$ -ésima zona agro-ecológica en la ubicación  $i$ -ésima ( $j = 2, 3, 4$ ),  $WRB_2(\mathbf{s}_i)$  es el grupo de suelo Ferralsol en la ubicación  $i$ -ésima y  $\hat{z}_2(\mathbf{s}_i)$  es un proceso estacionario gaussiano con variograma isotrópico esférico ajustado  $\hat{\gamma}_2(h_2) = 0.147 \left( \frac{3}{2} \left( \frac{h_2}{5.415} \right) - \frac{1}{2} \left( \frac{h_2}{5.415} \right)^3 \right)$  para  $h_2 > 0$  y  $\hat{\gamma}_2(0) = 0$ .

La Tabla 4.5 presenta los parámetros estimados, desviación estándar (S.D.) y los intervalos de confianza al 95 % para cada uno de los parámetros en el BSLMM ( $H_5$  or  $H_8$ ). En el SMM, la variable altura tiene un efecto significativo positivo sobre el contenido de arcilla, mientras la zona 1 agro-ecológica reduce el contenido de arcilla. Por otro lado, en el SVDM, el suelo Ferralsol no es significativo pero difiere del suelo Acrisol, mientras que las zonas agro-ecológicas tienen efectos significativos. Además, los parámetros de correlación espacial sin efecto pepita son relevantes para ajustar el modelo de dispersión (precisión) variable. En particular, para el SVDM utilizado en el ajuste del BSLMM, el intervalo al 95 % para  $\sigma_2^2$  es (0.0001; 0.2471) en la escala log y el intervalo para  $\vartheta_2$  es de (0.0025; 9.6836) km. Además utilizando la función de enlace logit en el SMM y la función de enlace log en el SVDM para ajustar el BSLMM, se encuentra que el pseudo  $R_s^2 = 0.51$ , lo cual sugiere una bondad de ajuste del modelo propuesto un poco baja.

Además, la Figura 4.4 (panel derecho) de la deviance residual ( $r_i$ ) contra valores pronosticados del contenido de arcilla muestra la ausencia de cualquier relación obvia, indicando un adecuado ajuste de primer orden en el modelo propuesto. El panel izquierdo de la Figura 4.3 muestra la ausencia de cualquier relación entre la deviance residual ( $r_i$ 's) y el ordenamiento de la capa de suelo.

Una vez ajustado el modelo, la idea en muchos estudios es hacer un mapa de predicción del contenido de arcilla; sin embargo, un análisis de esta clase requiere que las variables explicativas sean medidas en las ubicaciones o localizaciones a predecir. En este estudio, no se tiene esta información, por ello este

TABLA 4.5: Estimaciones e intervalos de confianza de 95 % para los parámetros involucrados en el BSLMM para ajustar el contenido de arcilla.

Modelo espacial para la media				
Parámetro	Estimación	S.D.	Cuantil	
			2.5 %	97.5 %
$\zeta_0$	-2.085	0.646	-2.9783	-0.4990
$\zeta_1$	0.002	0.000	0.0001	0.0028
$\zeta_2$	0.630	0.349	-0.2751	1.0547
$\zeta_3$	0.457	0.430	-0.6372	1.0343
$\zeta_4$	0.212	0.569	-1.1810	0.9723

Modelo espacial con dispersión variable				
Parámetro	Estimación	S.D.	Cuantil	
			2.5 %	97.5 %
$\varphi_0$	2.932	0.818	1.5422	4.6789
$\varphi_1$	1.170	0.628	-0.9638	1.5706
$\varphi_2$	2.434	0.633	0.0367	2.5415
$\varphi_3$	1.265	0.896	-1.3104	2.3072
$\varphi_4$	-1.655	1.102	-3.1615	1.0566
$\sigma_2^2$	0.147	0.069	0.0001	0.2471
$\vartheta_2$	5.415	2.557	0.0025	9.6836

mapa no se obtiene en esta aplicación. Sin embargo, en la siguiente aplicación, se obtienen el mapa de predicción de la variable respuesta y su correspondiente desviación estándar para el error de predicción.

#### 4.4.2 Contenido de Magnesio

El conjunto de datos que se consideran en esta sección corresponden a muestras de suelo recogidas con un taladro de tipo holandés en una rejilla regular incompleta a una distancia de aproximadamente 50 metros, con coordenadas geográficas: norte y este a 900 metros de distancia en ambas direcciones. Las muestras de suelo fueron tomadas de 0 a 20 cm de profundidad de la capa en cada una de las 178 ubicaciones (ver Figura 4.6). El contenido de magnesio fue medido en  $mmol_c/dm^3$ . La región de estudio se dividió en tres subregiones que han experimentado diferentes sistemas de manejo de suelos. La primera en la esquina superior izquierda (ver Figura 4.6), normalmente se inunda durante cada temporada de lluvias y ya no se utiliza como un área experimental debido a su altitud variable. La segunda, corresponde a la mitad de la región de estudio (parte inferior de la Figura 4.6), y la tercera, en la esquina superior derecha

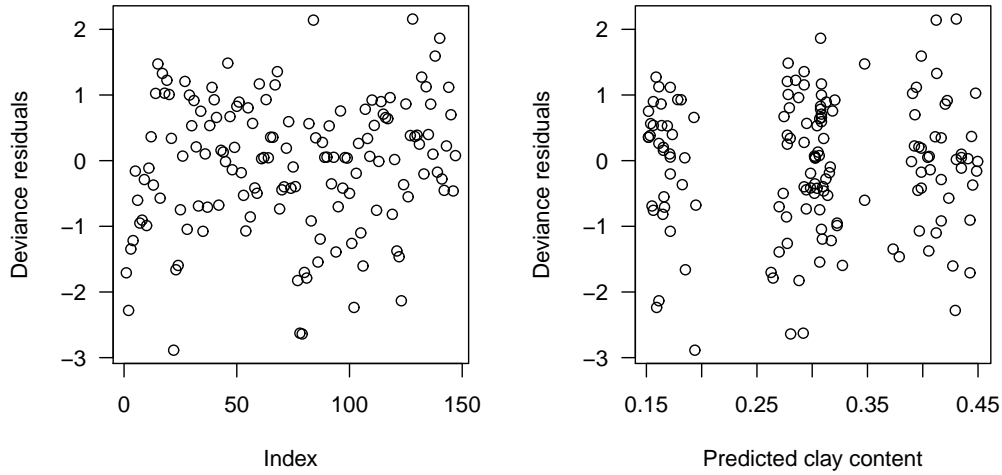


FIGURA 4.4: Deviance residual contra: el ordenamiento de la profundidad de la capa de suelo (*panel izquierdo*) y los valores pronosticados (*panel derecho*) utilizando el BSLMM para el contenido de arcilla.

de la Figura 4.6, han recibido fertilizantes en el pasado: la segunda se utiliza con frecuencia como un área experimental, mientras que la tercera se utiliza como un área experimental (Capeche et al. 1997, Diggle & Ribeiro 2007).

Por lo tanto, este conjunto de datos tiene información acerca del contenido de magnesio ( $y(\mathbf{s}_i)$ ), las coordenadas espaciales ( $w_1, w_2$ ), altitud (ALT) y la sub-región (SR), la cual está asociada con los tres períodos de fertilización en las diferentes áreas. Los datos se tomaron de (Capeche et al. 1997) y el objetivo principal del estudio es hacer una planificación adecuada del uso de la tierra, lo que permite la gestión racional y sostenible, evitando el proceso de erosión. Esto es importante con el fin de asignar los subsidios en los campos experimentales para realizar búsquedas que se pueden extrapolar a suelos y zonas climáticas similares. En este estudio, se asume que los  $y(\mathbf{s}_i)$ 's son variables beta condicionalmente independientes dados los procesos espaciales estocásticos no observados  $z_1(\mathbf{s}_i)$  y  $z_2(\mathbf{s}_i)$  para el SMM y el SVDM, respectivamente.

Similar a la aplicación del contenido de arcilla y a la simulación, no se encontró ninguna evidencia de anisotropía tanto en el SMM y el SVDM. Además, se ajustan los datos a un BSLMM utilizando la función de enlace cloglog con efectos aleatorios independientes e igualmente distribuidos en el SMM y la transformación 0.3 como función de enlace con los efectos aleatorios independientes e igualmente distribuidos en el SVDM. Para la estimación de máxima verosimilitud, se utiliza  $b_0 = (\zeta_0, \varphi_0, \theta_{01}, \theta_{02})^t$

con  $\zeta_0 = (-5.9, -0.0001, 0.0008, 0.2, 0.5, 0.4)^t$ ,  $\varphi_0 = (29.5, -6.6, -3.5, -1.2)^t$ ,  $\theta_{01} = (4, 0.6, 1)^t$  y  $\theta_{02} = (5, 0.5, 1)^t$  para aproximar  $\tilde{f}(\cdot)$  a  $f(\cdot | b_0)$ . El tamaño de muestra usando Monte Carlo se desarrolló similarmente como en el caso de la simulación y la aplicación del contenido de arcilla.

En la Tabla 4.6 se muestran los resultados obtenidos luego de ejecutar el algoritmo MCMC utilizando el BSLMM. Al comparar la diferencia entre los  $\log \hat{L}$ 's para  $H_8$  y  $H_1$  con la distribución  $0.5\chi_{(6)}^2$  (cuyo cuantil 95 % es 6.296), se puede decir que existe una clara correlación espacial tanto en el SMM como en el SVDM. Sin embargo, cuando se comparan las diferencia de los  $\log \hat{L}$ 's en  $H_3$  y  $H_2$  con respecto a  $H_1$ , no se encuentra por separado evidencia de correlación espacial en el SVDM y en el SMM ( $0.5\chi_{(3)}^2 = 3.907$ ). Si se compara  $H_8$  con respecto a  $H_3$  y  $H_2$ , se observa que es necesaria la correlación espacial en el SMM y en el SVDM para ajustar el BSLMM.

La Tabla 4.6 también muestra que la selección de la función de correlación es relevante; la función de correlación gaussiana para el SMM y la función de correlación esférica para el SVDM ( $H_8$ ) es ligeramente mejor que el modelo  $H_7$ . En  $H_8$ , los efectos pepita en los dos modelos no son importantes. Además, similarmente a los estudios anteriores, cuando se utilizaron diferentes valores iniciales ( $b_0$ ) y funciones de correlación, los resultados no cambiaron mucho y el patrón fue el mismo al de la Tabla 4.6 al realzar varias ejecuciones del algoritmo MCMC empleando el BSLMM.

En la Tabla 4.7 se presentan los resultados de las log-verosimilitudes ensayando diferentes funciones de enlace para el SMM y el SVDM, y utilizando las funciones de correlación gaussianas y esféricas en el SMM y el SVDM, respectivamente. De acuerdo a estos resultados, se elige el modelo con función de enlace cloglog para el SMM y la transformación 0.3 como enlace para el SVDM para ajustar el BSLMM ya que éste tiene la más alta log-verosimilitud (338.71).

Por lo tanto, asumimos el ajuste BSML con modelo espacial mixto dado por

$$\begin{aligned} \text{cloglog}(\hat{\mu}(\mathbf{s}_i)) &= \log\{-\log[1 - \hat{\mu}(\mathbf{s}_i)]\} = -7.154 - 0.023 * w_1(\mathbf{s}_i) + 0.964 * w_2(\mathbf{s}_i) \\ &\quad + 0.137 * ALT(\mathbf{s}_i) + 0.294 * SR_2(\mathbf{s}_i) + 0.215 * SR_3(\mathbf{s}_i) + \hat{z}_1(\mathbf{s}_i) \\ (\hat{\phi}(\mathbf{s}_i))^{0.3} &= 88.263 + 2.571 * w_1(\mathbf{s}_i) - 13.167 * w_2(\mathbf{s}_i) - 3.926 * ALT(\mathbf{s}_i) + \hat{z}_2(\mathbf{s}_i) \end{aligned}$$

donde  $\hat{y}(\mathbf{s}_i)$  es el contenido de magnesio en la ubicación  $i$ -ésima ( $i = 1, \dots, 178$ ),  $w_1(\mathbf{s}_i)$  y  $w_2(\mathbf{s}_i)$  son las  $i$ -ésimas coordenadas espaciales,  $ALT(\mathbf{s}_i)$  es la altitud por encima del nivel del mar en la ubicación  $i$ -ésima,  $SR_j(\mathbf{s}_i)$  es la  $j$ -ésima sub-región en la ubicación  $i$ -ésima ( $j = 2, 3$ ),  $\hat{z}_1(\mathbf{s}_i)$  es un proceso estacionario gaussiano con variograma ajustado isotrópico gaussiano  $\hat{\gamma}_2(h_2) = 0.047(1 - \exp(-3h_1^2/0.497^2))$  para  $h_1 > 0$  y  $\hat{\gamma}_1(0) = 0$ , y  $\hat{z}_2(\mathbf{s}_i)$  es

TABLA 4.6: Estimaciones beta espaciales por máxima verosimilitud para el contenido de magnesio utilizando los modelos  $H_1$ ,  $H_2$ ,  $H_3$ ,  $H_4$ ,  $H_5$ ,  $H_6$ ,  $H_7$  and  $H_8$ .

Modelo espacial para la media									
Modelo	$\hat{\zeta}_0$	$\hat{\zeta}_1$	$\hat{\zeta}_2$	$\hat{\zeta}_3$	$\hat{\zeta}_4$	$\hat{\zeta}_5$	$\hat{\sigma}_1^2$	$\hat{\vartheta}_1$	$\hat{\tau}_1^2$
$H_1$	-5.829	-0.089	0.721	0.194	0.385	0.361	*	*	*
$H_2$	-5.929	-0.085	0.732	0.195	0.399	0.381	*	*	*
$H_3$	-6.035	-0.095	0.770	0.190	0.396	0.366	0.000	0.000	0.000
$H_4$	-5.994	-0.093	0.760	0.189	0.399	0.370	0.000	0.000	0.000
$H_5$	-4.102	-0.061	0.545	0.136	0.272	0.259	0.000	0.000	0.000
$H_6$	-5.934	-0.085	0.733	0.195	0.400	0.382	0.000	0.000	0.000
$H_7$	-6.917	-0.083	0.969	0.125	0.471	0.387	0.032	0.373	0.000
$H_8$	-7.154	-0.023	0.964	0.137	0.294	0.215	0.047	0.497	0.000

Modelo espacial con dispersión variable								
Modelo	$\hat{\varphi}_0$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{\sigma}_2^2$	$\hat{\vartheta}_2$	$\hat{\tau}_2^2$	$\log \hat{L}$
$H_1$	112.560	0.335	-13.965	-4.856	*	*	*	327.78
$H_2$	105.463	0.432	-13.473	-4.716	0.000	0.000	0.338	328.44
$H_3$	108.183	0.574	-14.088	-4.787	*	*	*	327.99
$H_4$	101.101	0.705	-12.890	-4.764	0.143	0.600	0.000	328.06
$H_5$	105.879	0.654	-13.807	-4.695	0.000	0.000	0.280	328.58
$H_6$	105.348	0.453	-13.477	-4.710	0.141	0.037	0.203	328.45
$H_7$	88.450	1.844	-12.564	-3.821	0.302	0.024	0.000	337.56
$H_8$	88.263	2.571	-13.167	-3.926	0.287	0.075	0.000	338.71

TABLA 4.7: Log-verosimilitudes para el contenido de magnesio utilizando diferentes funciones de enlace en el SMM y el SVDM con funciones de correlación gaussiana y esférica, respectivamente.

Modelo espacial de dispersión variable ( $\nu_2$ )	Modelo espacial para la media			
	loglog	logit	cloglog	probit
0.0 (log)	327.98	337.43	337.16	336.48
0.3	328.55	338.42	338.71	338.24
0.5 (raíz cuadrada)	328.39	338.37	338.60	338.15
1.0 (identidad)	337.69	337.41	327.91	337.16

un proceso estacionario gaussiano con variograma ajustado isotrópico esférico  $\hat{\gamma}_2(h_2) = 0.287 \left( \frac{3}{2} \left( \frac{h_2}{0.075} \right) - \frac{1}{2} \left( \frac{h_2}{0.075} \right)^3 \right)$  para  $h_2 > 0$  y  $\hat{\gamma}_2(0) = 0$ .

La Tabla 4.8 muestra los parámetros estimados, las desviaciones estándar (S.D.) y los intervalos de confianza de 95 % para cada uno de los parámetros en el BSLMM ( $H_8$ ). En el SMM, la variable Norte-Sur ( $w_2$ ), altitud, y subregión tienen un efecto significativo positivo sobre el contenido de magnesio, mientras que la coordenada Este-Oeste ( $w_1$ ) no tiene un efecto significativo. Por otro lado, en el SVDM, la variable Norte-Sur y la altitud tienen un efecto signi-



ficativo negativo sobre la respuesta de dispersión, y similarmente al modelo de media, la coordenada Este-Oeste no es significativa. Además, los parámetros de correlación espacial sin el efecto pepita son relevantes para ajustar la variación en ambos modelos (SMM y SVDM).

En particular, para el SMM usado en el BSLMM, el intervalo del 95 % para  $\sigma_1^2$  es (0.001;0.084) con el enlace cloglog y el intervalo para  $\vartheta_1$  de (0.111;0.949)km. Para el SVDM utilizado el BSLMM, el intervalo del 95 % para  $\sigma_2^2$  es (0.001;0.821) con la transformación 0.3 y el intervalo para  $\vartheta_2$  de (0.051; 1.000)km. Además de usar la función de enlace cloglog en el SMM y la transformación 0.3 como función de enlace en el SVDM para ajustar el BSLMM, se encuentra que el pseudo  $R_s^2 = 0.42$ , lo cual sugiere una bondad de ajuste para el modelo propuesto un poco baja.

TABLA 4.8: Estimaciones e intervalos de confianza de 95 % para los parámetros involucrados en el ajuste de contenido de magnesio en el BSLMM.

Modelo espacial para la media				
Parámetro	Estimación	S.D.	Cuantil	
			2.5 %	97.5 %
$\zeta_0$	-7.154	1.758	-8.841	-2.110
$\zeta_1$	-0.023	0.196	-0.234	0.483
$\zeta_2$	0.964	0.317	0.105	1.311
$\zeta_3$	0.137	0.056	0.047	0.255
$\zeta_4$	0.294	0.145	0.012	0.526
$\zeta_5$	0.215	0.130	0.026	0.508
$\sigma_2^2$	0.047	0.024	0.001	0.084
$\vartheta_2$	0.497	0.162	0.111	0.949

Modelo espacial con dispersión variable				
Parámetro	Estimación	S.D.	Cuantil	
			2.5 %	97.5 %
$\varphi_0$	88.263	18.662	21.939	94.241
$\varphi_1$	2.571	1.492	-2.929	2.961
$\varphi_2$	-13.167	2.476	-16.650	-6.793
$\varphi_3$	-3.926	0.760	-4.741	-2.228
$\sigma_2^2$	0.287	0.281	0.001	0.821
$\vartheta_2$	0.075	0.238	0.051	1.000

Además, en la Figura 4.5 (panel derecho) se gráfica la deviance residual ( $r_i$ ) contra los valores de la predicción del contenido de magnesio; esta gráfica muestra la ausencia deseada de cualquier relación obvia, indicando un adecuado ajuste de primer orden del modelo propuesto. El panel izquierdo de la Figura 4.5 muestra la ausencia de cualquier relación entre la deviance residual ( $r_i$ 's) y el ordenamiento de la capa de suelo.

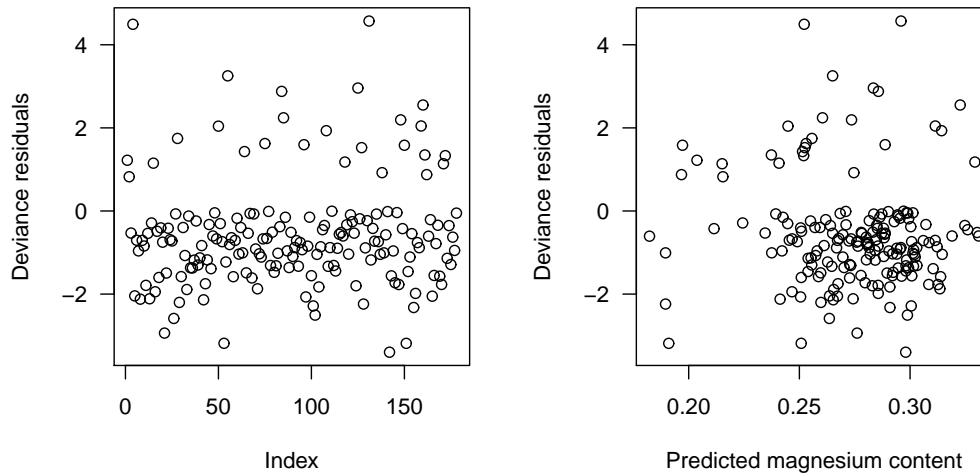


FIGURA 4.5: Deviance residual contra: ordenamiento de la capa de suelo (*panel izquierdo*) y los valores pronosticados (*panel derecho*) utilizando el BSLMM para el contenido de magnesio.

Una vez ajustado el modelo, se obtuvieron el mapa de predicción del contenido de magnesio para el SMM y el SVDM y el mapa de sus correspondientes desviaciones estándar para los errores de predicción utilizando el BSLMM. Los resultados se presentan en la Figura 4.6. Observe que en las ubicaciones observadas o cercanos a éstas, la función de distribución predictiva muestra un reducido error estándar alrededor de estos puntos observados en ambos modelos (ver paneles (b) y (d)). En el panel (a) (modelo de media) de la Figura 4.6, se puede ver que las áreas dentro de las gamas de colores de color naranja pálido y amarillo están los altos contenidos de magnesio porque las  $\hat{\phi}(\mathbf{s}_i)$ 's son altas, mientras que las áreas en el rango de color rojo-naranja se encuentran los niveles bajos de magnesio, los cuales no exceden el umbral de 25%. Por otra parte en el panel (b) (mapa de dispersión variable) de la Figura 4.6, se puede observar la precisión sobre los puntos observados en el mapa de media. En este panel, las áreas dentro de los rangos de color naranja pálido y amarillo son áreas con una alta precisión de predicción del contenido de magnesio; además, áreas en el intervalo de color rojo-naranja son áreas con una baja precisión de la predicción del contenido de magnesio, las cuales no exceden el umbral 100. Observe que los mapas de predicción muestran discontinuidades en los límites entre las subregiones como consecuencia del tratamiento de la sub-región en un factor de tres niveles.

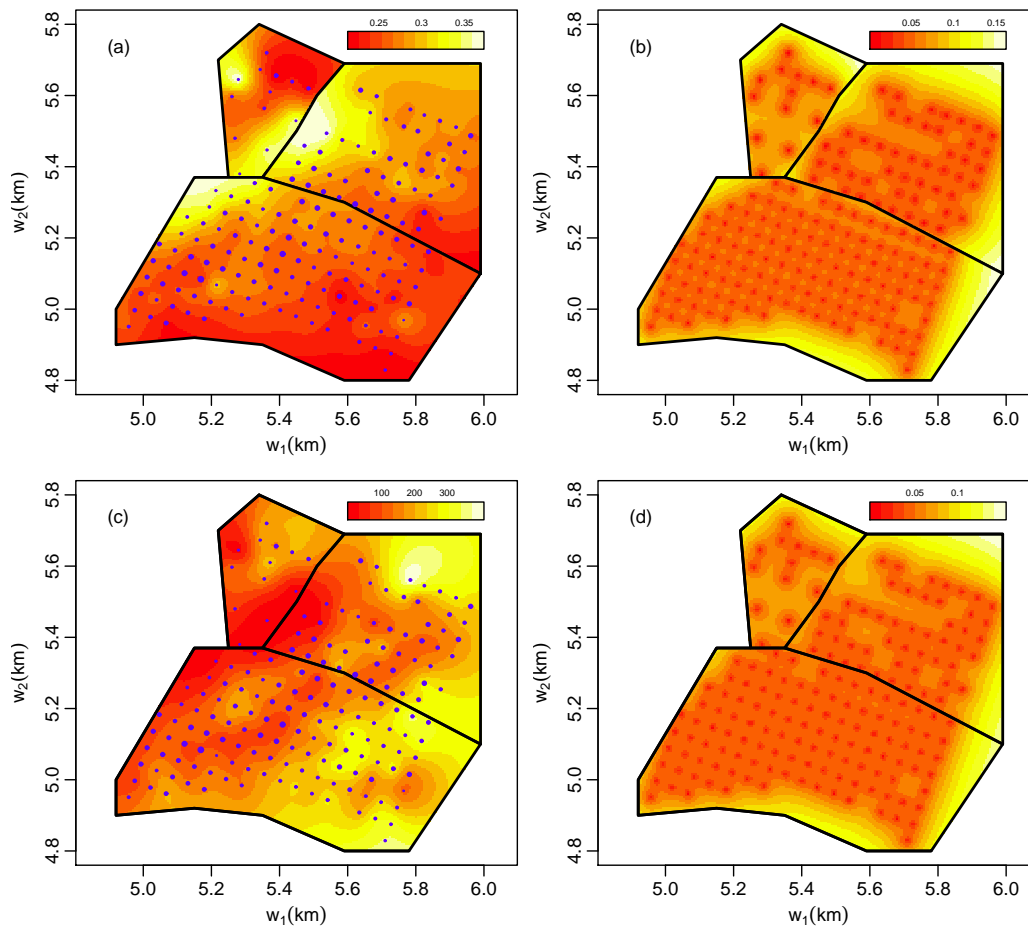


FIGURA 4.6: (a): Estimaciones puntuales utilizando el SMM, superpuesto con los contenidos de magnesio observados en los campos estudiados, donde cada punto es proporcional a la correspondiente medida del contenido de magnesio, y las líneas delimitan las sub-regiones con diferentes prácticas en el manejo del suelo. (b): Errores de predicción estándar para el SMM. (c): Estimaciones puntuales usando el SVDM (modelo de precisión), superpuesto con el contenido de magnesio observado en los campos estudiados. (d): Errores de predicción estándar para el SVDM.

## Apéndice 4.1. Representación espectral

En esta sección se hace la representación espectral para los procesos  $z_1$  y  $z_2$ . De acuerdo a Royle & Wikle (2005), los procesos  $z_j$ 's ( $j = 1, 2$ ) se pueden expandir en términos de funciones bases de Fourier (es decir senos y cosenos). Además asumiendo que los procesos espaciales están definidos sobre una rejilla regular, para las localizaciones  $\mathbf{s}_i$ ,  $i = 1, \dots, n$ , y frecuencias espaciales  $w_q = q/n$  para

$q = 0, \dots, n/2$  ( $n$  incluso), se tiene

$$\begin{aligned} z_j(\mathbf{s}_i) &= \sum_{q=0}^{n/2} \delta_j^{(1)}(q) \cos(2\pi \mathbf{s}_i w_q) + \sum_{q=1}^{n/2-1} \delta_j^{(2)}(q) \sin(2\pi \mathbf{s}_i w_q) \\ &= (\boldsymbol{\psi}_i^{(1)})^t \boldsymbol{\delta}_j^{(1)} + (\boldsymbol{\psi}_i^{(2)})^t \boldsymbol{\delta}_j^{(2)} = \boldsymbol{\psi}_i^t \boldsymbol{\delta}_j, \quad j = 1, 2 \end{aligned}$$

donde  $\boldsymbol{\psi}_i^{(1)} = (\psi_i^{(1)}(w_0), \dots, \psi_i^{(1)}(w_{n/2}))^t$ ,  $\boldsymbol{\psi}_i^{(2)} = (\psi_i^{(2)}(w_1), \dots, \psi_i^{(2)}(w_{n/2-1}))^t$ ,  $\psi_i^{(1)}(w_q) = \cos(2\pi \mathbf{s}_i w_q)$ ,  $\psi_i^{(2)}(w_q) = \sin(2\pi \mathbf{s}_i w_q)$ ,  $\boldsymbol{\psi}_i = ((\boldsymbol{\psi}_i^{(1)})^t, (\boldsymbol{\psi}_i^{(2)})^t)^t$ ,  $\boldsymbol{\delta}_j^{(1)} = (\delta_j^{(1)}(0), \dots, \delta_j^{(1)}(n/2))^t$ ,  $\boldsymbol{\delta}_j^{(2)} = (\delta_j^{(2)}(1), \dots, \delta_j^{(2)}(n/2 - 1))^t$  y  $\boldsymbol{\delta}_j = ((\boldsymbol{\delta}_j^{(1)})^t, (\boldsymbol{\delta}_j^{(2)})^t)^t$ .

Es bien conocido que para procesos aleatorios estacionarios de segundo orden, los coeficientes  $\delta_{(q)}$ 's son casi incorrelacionados y sus varianzas a una frecuencia dada son aproximadamente iguales a la mitad de la función de densidad espectral de potencia a esa frecuencia, excepto para las frecuencias  $w_0$  y  $w_{n/2}$ , en donde la varianza es igual a la función de densidad espectral de potencia asociada (Shumway & Stoffer 2000). Así, para  $j = 1, 2$ , asumiendo que  $\mathbf{z}_j$  es un proceso estacionario de segundo orden con matriz de covarianza  $\boldsymbol{\Sigma}_{z_j} = \sigma_1^2 \mathbf{R}_j$ , donde  $\mathbf{R}_j$  es la matriz de correlación, entonces  $\text{Cov}(\boldsymbol{\delta}_j) \approx \sigma_j^2 \mathbf{C}_j$ , donde  $\mathbf{C}_j$  es una matriz diagonal con los elementos en la diagonal dados por  $[f_j(w_0), \frac{1}{2}f_j(w_1), \dots, \frac{1}{2}f_j(w_{n/2}), \frac{1}{2}f_j(w_1), \dots, \frac{1}{2}f_j(w_{n/2-1})]$ , donde  $f_j(w_q)$  es la densidad espectral a la frecuencia  $w_q$  correspondiente a la función de correlación utilizada para construir  $\mathbf{R}_j$ .

En este caso, se parametriza la matriz de correlación  $\mathbf{R}_j = \mathbf{R}(\boldsymbol{\theta}_j)$ ,  $j = 1, 2$ , en términos de un vector de parámetros de dependencia espacial  $\boldsymbol{\theta}_j$ . Así, la matriz diagonal  $\mathbf{D}(\boldsymbol{\theta}_j)$  es también una función de  $\boldsymbol{\theta}_j$ . En el análisis presentado aquí la matriz de covarianza Matérn con función de densidad espectral asociada a la frecuencia  $w$  está dada por Royle & Wikle (2005) como

$$f_j(w) = \frac{2^{\kappa_j-1} \sigma_j \Gamma(\kappa_j + d_w/2) \vartheta_j^{2\kappa_j}}{\pi^{d_w/2} (\vartheta_j^2 + w^2)^{\kappa_j + d_w/2}}, \quad \sigma_j > 0, \vartheta_j > 0, \kappa_j > 0, \quad j = 1, 2 \quad (4.19)$$

donde  $d_w$  es la dimensionalidad del proceso espacial (Stein 1999, p. 49),  $\kappa_j$  está relacionado con el grado de suavidad del  $j$ -ésimo proceso espacial,  $\vartheta_j$  está relacionado con el  $j$ -ésimo rango de correlación y  $\sigma_j$  es proporcional a la  $j$ -ésima varianza del proceso (Stein 1999, p. 48). Así, si se elige  $\boldsymbol{\Psi}_j$  como las funciones base de Fourier, entonces (4.19) sugiere la forma de  $\boldsymbol{\Sigma}_{\delta_j}(\boldsymbol{\theta}_j)$  (matriz diagonal, con los elementos de la diagonal correspondientes a la frecuencia  $w$  dada por (4.19)).

## Apéndice 4.2. Interpolación de efectos aleatorios utilizando los predictores lineales insesgados

La minimización del error cuadrático medio de la predicción se realiza a través de

$$\begin{aligned} \mathbb{E} \left[ (\tilde{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j^0)^t \mathbf{A}_j (\tilde{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j^0) \right] &= \int \int (\tilde{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j^0)^t \mathbf{A}_j (\tilde{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j^0) \\ &\quad \times f(\boldsymbol{\eta}_j^0, \boldsymbol{\eta}_j) d\boldsymbol{\eta}_j^0 d\boldsymbol{\eta}_j, \quad j = 1, 2 \end{aligned} \quad (4.20)$$

donde  $f(\boldsymbol{\eta}_j^0, \boldsymbol{\eta}_j)$  es la función de densidad conjunta de  $\boldsymbol{\eta}_j^0$  y  $\boldsymbol{\eta}_j$ , y  $\mathbf{A}_j$  es una matriz simétrica definida positiva.

Utilizando (4.16), el lado izquierdo de (4.20) se puede escribir como

$$\begin{aligned} \mathbf{q}_j &= \mathbb{E} \left[ (\mathbf{p}_j + \mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0)^t \mathbf{A}_j (\mathbf{p}_j + \mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0) \right] \\ &= \mathbf{p}_j^t \mathbf{A}_j \mathbf{p}_j + 2\mathbf{p}_j^t \mathbf{A}_j \mathbb{E} (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0) \\ &\quad + \mathbb{E} \left[ (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0)^t \mathbf{A}_j (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0) \right], \quad j = 1, 2 \end{aligned} \quad (4.21)$$

Derivando parcialmente la ecuación (4.21) con respecto a  $\mathbf{p}_j$  e igualando a  $\mathbf{0}$ , se tiene que

$$\begin{aligned} \frac{\partial \mathbf{q}_j}{\partial \mathbf{p}_j} &= 2\mathbf{A}_j [\mathbf{p}_j + \mathbb{E} (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0)] = \mathbf{0} \\ \mathbf{p}_j &= -\mathbb{E} (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0), \quad j = 1, 2 \end{aligned} \quad (4.22)$$

Sustituyendo (4.22) en (4.21), se obtiene

$$\begin{aligned} \mathbf{q}_j &= -\left[ \mathbb{E} (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0) \right]^t \mathbf{A}_j \mathbb{E} (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0) \\ &\quad + \mathbb{E} \left[ (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0)^t \mathbf{A}_j (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0) \right] = \text{tr} \{ \mathbf{A}_j \text{Var} (\mathbf{Q}_j \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^0) \} \\ &= \text{tr} \{ \mathbf{A}_j [\mathbf{Q}_j \text{Var} (\boldsymbol{\eta}_j) \mathbf{Q}_j^t + \text{Var} (\boldsymbol{\eta}_j^0) - \mathbf{Q}_j \text{Cov} (\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) \\ &\quad - \text{Cov}^t (\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) \mathbf{Q}_j^t] \}, \quad j = 1, 2 \end{aligned} \quad (4.23)$$

donde  $\text{Var} (\boldsymbol{\eta}_j) = \sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)$ ,  $\text{Var} (\boldsymbol{\eta}_j^0)$  es la matriz de covarianza de  $\boldsymbol{\eta}_j^0$  y  $\text{Cov} (\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0)$  es la matriz de covarianzas entre  $\boldsymbol{\eta}_j$  y  $\boldsymbol{\eta}_j^0$ .

Ahora se desea minimizar (4.23) con respecto a  $\mathbf{Q}_j$ , para lo cual se ignora  $\mathbf{A}_j$  y  $\text{Var} (\boldsymbol{\eta}_j^0)$  porque no involucran a  $\mathbf{Q}_j$ . Entonces, derivando parcialmente (4.23) con respecto a  $\mathbf{Q}_j$  e igualando a  $\mathbf{0}$ , se tiene que

$$\begin{aligned} \frac{\partial \mathbf{q}_j}{\partial \mathbf{Q}_j} &= 2\mathbf{Q}_j \text{Var} (\boldsymbol{\eta}_j) - 2\text{Cov}^t (\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) = \mathbf{0} \\ \mathbf{Q}_j &= \text{Cov}^t (\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) [\text{Var} (\boldsymbol{\eta}_j)]^{-1}, \quad j = 1, 2 \end{aligned} \quad (4.24)$$

Así, sustituyendo (4.22) y (4.24) en (4.16), el mejor pseudo predictor lineal insesgado esta dado por

$$\tilde{\boldsymbol{\eta}}_j = \mathbf{E}(\boldsymbol{\eta}_j^0) + \text{Cov}^t(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) [\text{Var}(\boldsymbol{\eta}_j)]^{-1} [\boldsymbol{\eta}_j - \mathbf{E}(\boldsymbol{\eta}_j)], \quad j = 1, 2 \quad (4.25)$$

Específicamente, la predicción para el SMM y SVDM están dados, respectivamente, por

$$\begin{aligned} \tilde{\boldsymbol{\eta}}_1 &= \mathbf{V}_s^0 \boldsymbol{\zeta} + \text{Cov}^t(\boldsymbol{\eta}_1, \boldsymbol{\eta}_1^0) [\sigma_1^2 (\mathbf{R}(\vartheta_1) + \tau_{R_1}^2 \mathbf{I}_n)]^{-1} [\boldsymbol{\eta}_1 - \mathbf{V}_s \boldsymbol{\zeta}] \\ \tilde{\boldsymbol{\eta}}_2 &= \mathbf{U}_s^0 \boldsymbol{\varphi} + \text{Cov}^t(\boldsymbol{\eta}_2, \boldsymbol{\eta}_2^0) [\sigma_2^2 (\mathbf{R}(\vartheta_2) + \tau_{R_2}^2 \mathbf{I}_n)]^{-1} [\boldsymbol{\eta}_2 - \mathbf{U}_s \boldsymbol{\varphi}] \end{aligned} \quad (4.26)$$

donde  $\mathbf{V}_s^0$  y  $\mathbf{U}_s^0$  son matrices de variables explicativas para  $n'$  nuevos individuos en el espacio para el SMM y el SVDM, respectivamente.

Tomando esperanza en (4.26), se obtiene

$$\begin{aligned} \mathbf{E}(\tilde{\boldsymbol{\eta}}_1 | \mathbf{y}_s) &= \mathbf{V}_s^0 \boldsymbol{\zeta} + \text{Cov}^t(\boldsymbol{\eta}_1, \boldsymbol{\eta}_1^0) [\sigma_1^2 (\mathbf{R}(\vartheta_1) + \tau_{R_1}^2 \mathbf{I}_n)]^{-1} [\mathbf{E}(\boldsymbol{\eta}_1 | \mathbf{y}_s) - \mathbf{V}_s \boldsymbol{\zeta}] \\ &\approx \mathbf{V}_s^0 \boldsymbol{\zeta} + \text{Cov}^t(\boldsymbol{\eta}_1, \boldsymbol{\eta}_1^0) [\sigma_1^2 (\mathbf{R}(\vartheta_1) + \tau_{R_1}^2 \mathbf{I}_n)]^{-1} [\tilde{\mathbf{E}}_r(\boldsymbol{\eta}_1 | \mathbf{y}_s) - \mathbf{V}_s \boldsymbol{\zeta}] \\ \mathbf{E}(\tilde{\boldsymbol{\eta}}_2 | \mathbf{y}_s) &= \mathbf{U}_s^0 \boldsymbol{\varphi} + \text{Cov}^t(\boldsymbol{\eta}_2, \boldsymbol{\eta}_2^0) [\sigma_2^2 (\mathbf{R}(\vartheta_2) + \tau_{R_2}^2 \mathbf{I}_n)]^{-1} [\mathbf{E}(\boldsymbol{\eta}_2 | \mathbf{y}_s) - \mathbf{U}_s \boldsymbol{\varphi}] \\ &\approx \mathbf{U}_s^0 \boldsymbol{\varphi} + \text{Cov}^t(\boldsymbol{\eta}_2, \boldsymbol{\eta}_2^0) [\sigma_2^2 (\mathbf{R}(\vartheta_2) + \tau_{R_2}^2 \mathbf{I}_n)]^{-1} [\tilde{\mathbf{E}}_r(\boldsymbol{\eta}_2 | \mathbf{y}_s) - \mathbf{U}_s \boldsymbol{\varphi}] \end{aligned}$$

donde  $\tilde{\mathbf{E}}_r$  es el vector de medias empíricas basado en las muestras  $\boldsymbol{\eta}_j(1), \dots, \boldsymbol{\eta}_j(r)$ ,  $j = 1, 2$ .

La matriz de covarianza para la predicción presentada en (4.25) está dada por

$$\begin{aligned} \text{Var}(\tilde{\boldsymbol{\eta}}_j | \mathbf{y}_s) &= \boldsymbol{\Sigma}_j + \text{Cov}^t(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) [\sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)]^{-1} [\text{Var}(\boldsymbol{\eta}_j | \mathbf{y}_s)] \\ &\quad [\sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)]^{-1} \text{Cov}(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) \\ &\approx \boldsymbol{\Sigma}_j + \text{Cov}^t(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) [\sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)]^{-1} [\widetilde{\text{Var}}_r(\boldsymbol{\eta}_j | \mathbf{y}_s)] \\ &\quad [\sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)]^{-1} \text{Cov}(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) \end{aligned}$$

donde  $\boldsymbol{\Sigma}_j = \text{Var}(\boldsymbol{\eta}_j^0) - \text{Cov}^t(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0) [\sigma_j^2 (\mathbf{R}(\vartheta_j) + \tau_{R_j}^2 \mathbf{I}_n)]^{-1} \text{Cov}(\boldsymbol{\eta}_j, \boldsymbol{\eta}_j^0)$  y  $\widetilde{\text{Var}}_r$  es la matriz de varianzas covarianzas basada en las muestras  $\boldsymbol{\eta}_j(1), \dots, \boldsymbol{\eta}_j(r)$ ,  $j = 1, 2$ .



## Capítulo 5

# Modelo autoregresivo espacio-tiempo lineal generalizado basado en distancias con perturbaciones autorregresivas espacio-tiempo

### 5.1 Introducción

La idea de combinar datos de corte transversales y series de tiempo es posterior a las sugerencias de Marschak (1939) y ha recibido considerable atención en la literatura econométrica, bioestadística, ciencias sociales y ciencias geográficas. Otras descripciones de las problemáticas más destacadas y soluciones sugeridas se pueden encontrar en Dielman (1983), Chamberlain (1984), Hsiao (1985) y Anselin (1988). En los últimos años, la literatura espacio-tiempo ha mostrado un creciente interés en la especificación y estimación de las relaciones econométricas y de las ciencias sociales basada en paneles espaciales. Los paneles espaciales suelen referirse a los datos que contienen las observaciones de series de tiempo de un número de unidades espaciales (códigos postales, municipios, regiones, estados, jurisdicciones, países, etc.). Este interés se puede explicar por el hecho de que los datos panel ofrecen a los investigadores extensas posibilidades de modelamiento en comparación al ajuste de una sola ecuación de corte transversal, lo cual fue el objetivo principal de la literatura econométrica espacial por un largo tiempo (Hsiao 2003). Un grupo considerablemente más pequeño (Elhorst 2003, Lee 2004, Elhorst 2005, Elhorst 2008) se ha ocupado del análisis social y econométrico de la dinámica espacial en los modelos de datos panel temporalmente estáticos.



Los datos panel son generalmente más informativos ya que contienen más variación y menos colinealidad entre las variables. El uso de los resultados de datos panel ayuda a contar en apariencia con una mayor disponibilidad de grados de libertad, y por lo tanto, se aumenta aparentemente la eficiencia en la estimación de los parámetros involucrados en el modelo a ajustar. Los datos panel también permiten la especificación de hipótesis de comportamiento más complejas, incluyendo efectos que no pueden ser abordados mediante datos de solo corte transversal (ver Hsiao (2003) para mayores detalles).

Los datos espacio-tiempo tratan con la interacción espacio-tiempo (autocorrelación espacio-tiempo) y la estructura espacio-tiempo (heterogeneidad espacio-tiempo) en modelos de regresión de corte transversal y datos panel (Paelinck & Klaassen 1979, Anselin 1988). Tal enfoque en la localización e interacción espacio-tiempo ha ganado recientemente un lugar más central no sólo en lo aplicado, sino también en la econometría teórica y ciencias sociales. En el pasado, los modelos que incorporaban explícitamente espacio o localización geográfica se encontraban principalmente en campos especializados tales como ciencia regional, urbana, economía de bienes inmuebles y geografía económica (Anselin & Florax 1995, Anselin & Kelejian 1997, Pace et al. 1998, Anselin et al. 2004).

Recientemente los métodos espaciales y espacio-tiempo cada vez se han aplicado más en una amplia gama de investigaciones empíricas en los campos más tradicionales de la economía y las ciencias sociales, incluyendo entre otros, estudios en el análisis de la demanda, el crecimiento económico, economía internacional, mercado laboral, índices de empleo, el desplazamiento por violencia armada, economía pública, finanzas públicas locales, producción agrícola y contaminación ambiental. Muchos de estos estudios tienen una variable de respuesta continua, sin embargo, cuando la variable respuesta es un conteo, una tasa o una respuesta binaria, no hay demasiada literatura que resuelva el problema mediante el uso de un modelo autorregresivo espacio-tiempo incluyendo perturbaciones espacio-tiempo autorregresivas de variables de estado estacionarias. Por lo tanto, estas aplicaciones no solo han dado lugar a nuevas ideas, desarrollos y ampliaciones, sino también a la nuevas preguntas. Algunos científicos han demostrado que los datos con dependencia espacial pueden alterar, e incluso revertir, los resultados de los modelos estándar de series de tiempo.

En la clásica variable respuesta continua, una sencilla versión de estos modelos, típicamente es un modelo espacial autorregresivo (SAR) que aumenta el modelo de regresión lineal incluyendo una variable adicional conocida como un rezago espacial. Cada observación de la variable rezagada en espacio-tiempo es un promedio ponderado de los valores de la variable dependiente observada para las otras unidades de corte transversal. Algunas de las versiones generalizadas del modelo SAR también permiten que las perturbaciones sean generadas

mediante un proceso autorregresivo espacio-tiempo y que las variables explicativas sean rezagadas espacialmente. Elhorst (2003, 2005, 2008) y Fingleton (2008) han proporcionado una revisión de muchas inquietudes que se plantean en la estimación de los cuatro modelos (modelos de efectos fijos, efectos aleatorios, coeficientes fijos y coeficientes aleatorios) de datos panel de uso común en la investigación aplicada, ampliando éstos al incluir la autocorrelación en el error espacial o una variable dependiente rezagada espacialmente.

Muchos de los trabajos y estudios presentados anteriormente establecen la importancia de integrar rezagos espaciales y temporales en el análisis de datos panel cuando la variable respuesta no es normal. Sin embargo, la literatura en modelos con dinámica espacial y temporal solo han presentado algunos progresos al tratar con esta clase de variable respuesta, pero en muchos casos por separado.

En este capítulo, se presenta una solución a problemas donde la variable respuesta es un conteo, una razón o una respuesta binaria (dicotómica) utilizando modelos lineales generalizados autorregresivos espacio-tiempo basados en distancias con perturbaciones autorregresivas espacio-tiempo (distance-based generalised linear space-time-autoregressive models with space-time-autoregressive disturbances, DBGLSTARAR). Este modelo puede también utilizar variables explicativas espaciales adicionales, así como variables explicativas asociadas al tiempo. Se incorporan medidas generales de distancia/disimilitud que se pueden aplicar a variables explicativas espaciales o espacio temporales: numéricas, categóricas o una mezcla de ellas. Se desarrolla la estimación de los efectos fijos y se determinan sus niveles de significancia.

Se amplía la posibilidad de contrastar o juzgar la especificación de efectos fijos contra la especificación de efectos aleatorios en modelos de datos panel, para incluir la autocorrelación del error espacio-tiempo o una variable dependiente rezagada espacio-tiempo utilizando pruebas de especificación. Se estima la matriz de varianzas y covarianzas de los parámetros en estos modelos extendidos. El proceso de estimación de los diferentes parámetros se hace mediante el método de ecuaciones de estimación generalizada (generalised estimating equations, GEEs) para espacio-tiempo, aunque en la sección de estimación se presentan detalladamente dos opciones que se pueden emplear: máxima verosimilitud y el método MCMC obtenido mediante máxima verosimilitud. Además, se presenta una medida de bondad de ajuste y el mejor predictor lineal insesgado cuando se utilizan estos modelos para propósitos de predicción.

En la aplicación presentada en este capítulo, se evalúan de forma empírica los valores de la I de Moran (Anselin et al. 2004) mediante una prueba estadística (la desviación estándar, I de Moran), que indica la significancia estadística de autocorrelación espacio-temporal, por ejemplo en los modelos residuales. Además, los residuos de Pearson y de deviance del modelo propues-

to se pueden graficar en un mapa que revela de manera más explícita patrones particulares de autocorrelación espacio-tiempo.

Este capítulo está dividido de la siguiente manera: en la Sección 5.2 se desarrolla la metodología propuesta: se presenta el GLM dinámico espacio-tiempo utilizando el método basado en distancias. En la Sección 5.3 se presentan tres métodos para la estimación de los parámetros: máxima verosimilitud, MCMC vía máxima verosimilitud y el método GEE para espacio-tiempo, este último es utilizado en el presente capítulo. En la Sección 5.4 se presenta la selección, validación y predicción del modelo ajustado utilizando el método GEE para espacio-tiempo; se expone una medida de bondad de ajuste, se dan algunas medidas para realizar el análisis de residuos, se hace el proceso de selección de las coordenadas principales y se realiza la predicción espacio-tiempo de un nuevo sujeto. Finalmente en la Sección 5.5 se presenta una aplicación en el que se modela el número de acciones armadas de los grupos guerrilleros de las FARC-EP y el ELN en Colombia, con el cual se ilustra la metodología propuesta.

## 5.2 Modelo lineal generalizado dinámico espacio-tiempo utilizando el método basado en distancias

Sea  $\{y(\mathbf{s}, t), \mathbf{s} \in D, t \in T\}$  un proceso estocástico espacio-temporal, el conjunto de índices  $D$  están en una superficie continua o un conjunto finito de ubicaciones discretas y  $T \subseteq \mathbb{Z}$ . De este modo, el modelo desarrollado es adecuado para tiempo discreto. Una distribución pertenece a la familia exponencial si su función de densidad está dada por (McCullagh & Nelder 1989)

$$f(y(\mathbf{s}, t); \alpha_{st}) = h_1(y(\mathbf{s}, t)) \exp\{\eta(\alpha_{st})h_2(y(\mathbf{s}, t)) - b(\alpha_{st})\}$$

donde  $\eta(\alpha_{st})$ ,  $b(\alpha_{st})$ ,  $h_1(y(\mathbf{s}, t))$  y  $h_2(y(\mathbf{s}, t))$  son funciones que toman valores en la recta real.

La propuesta de interpolación está construida para un modelo no gaussiano aleatorio de dinámica espacio-tiempo, considerando específicamente variables continuas e indicadoras en el modelo de tendencia. Los datos generados por el mecanismo condicional sobre la señal del modelo siguen un GLM como el descrito por McCullagh & Nelder (1989). Específicamente, nos enfocamos en modelos de rezago espacial y de error rezagado espacialmente; en el que se utiliza el rezago espacial tanto en el tiempo como en la variable dependiente. Esta dependencia se refiere a las ubicaciones vecinas en un período diferente.

Por lo tanto, el punto de partida es el siguiente modelo

$$\begin{aligned}\eta_{it} = \eta(\mathbf{s}_i, t) &= g(\mu_{it}) = \mathbf{v}_{1i}^t \boldsymbol{\gamma}_0 + \mathbf{v}_{2it}^t \boldsymbol{\gamma}_t + \pi_t \sum_{i'=1}^n w_{ii'}^{(1)} \eta_{i't} + \varepsilon_{it} \\ \varepsilon_{it} = \varepsilon(\mathbf{s}_i, t) &= \psi_t \sum_{i'=1}^n w_{ii'}^{(2)} \varepsilon_{i't} + e_{it}\end{aligned}\quad (5.1)$$

con  $i = 1, \dots, n$ ,  $t = 1, \dots, T$ ,  $|\pi_t| < 1$  y  $|\psi_t| < 1$ , donde  $\mu_{it} = \mu(\mathbf{s}_i, t) = E[y(\mathbf{s}_i, t) | \mathbf{v}_{1i}, \mathbf{v}_{2it}, \varepsilon_{it}]$ ,  $g(\cdot)$  es una función de enlace que es invertible y continua,  $\mathbf{v}_{1i}^t \boldsymbol{\gamma}_0 + \mathbf{v}_{2it}^t \boldsymbol{\gamma}_t$  es la tendencia,  $\mathbf{v}_{1i}^t = \mathbf{v}_1^t(\mathbf{s}_i) = (1, v_{i1}, \dots, v_{ip_1})$  es un vector que contiene variables explicativas asociadas a la  $\mathbf{s}_i$ -ésima ubicación espacial,  $\boldsymbol{\gamma}_0 = (\gamma_0, \gamma_1, \dots, \gamma_{p_1})^t$  es un vector de parámetros de regresión desconocidos,  $\mathbf{v}_{2it}^t = \mathbf{v}_2^t(\mathbf{s}_i, t) = (v_{it1}, \dots, v_{itp_{2t}})$  es un vector que contiene variables explicativas asociadas al espacio-tiempo en la  $\mathbf{s}_i$ -ésima ubicación y el  $t$ -ésimo tiempo, y  $\boldsymbol{\gamma}_t = (\gamma_{t1}, \dots, \gamma_{tp_{2t}})^t$  es un vector de parámetros de regresión desconocidos espacio-tiempo. Además,  $\pi_t$  es el coeficiente autorregresivo espacial en el  $t$ -ésimo período de tiempo,  $\varepsilon_{it}$  refleja el término de error autocorrelacionado espacial para la  $\mathbf{s}_i$ -ésima ubicación en el  $t$ -ésimo tiempo,  $\psi_t$  es llamado el coeficiente de autocorrelación espacial en el  $t$ -ésimo período de tiempo y  $e_{it} = e(\mathbf{s}_i, t)$  es un error normal idénticamente distribuido asociado a la  $\mathbf{s}_i$ -ésima ubicación en el  $t$ -ésimo tiempo con media cero y covarianza  $E(e_{it}, e_{it'}) = \sigma_{tt'}$  para  $t, t' = 1, \dots, T$ , con  $E(e_{it}, e_{i't}) = 0$  para  $i, i' = 1, \dots, n$ .

Por simplicidad, se asume inicialmente que  $w_{ii'}^{(1)} = w_{ii'}^{(2)} = w_{ii'}$ , con  $w_{ii'}$  un elemento de la matriz espacial de pesos o ponderaciones  $\mathbf{W}$ , la cual describe el arreglo espacial de las unidades en la muestra. Un paso inicial hacia la ponderación basada en la ubicación o localización podría ser excluir del modelo de calibración observaciones que están más allá de cierta distancia  $d_1$  del punto de regresión (Fotheringham & Zhan 1996). Esto es equivalente a dejar sus pesos en cero, produciendo la siguiente función de pesos

$$w_{ii'} = \begin{cases} 1 & \text{si } i' \in N(\mathbf{s}_i) \\ 0 & \text{en otro caso} \end{cases}$$

donde  $N(\mathbf{s}_i)$  es el conjunto de todos los vecinos cerca a la ubicación  $\mathbf{s}_i$ , la cual se construye por medio de la distancia ( $d_{ii'}$ ) entre dos ubicaciones  $\mathbf{s}_i$  y  $\mathbf{s}_{i'}$ . Esas dos ubicaciones se dicen que son vecinas si sus distancias son menores que un umbral ( $d_1$ ). Un camino para combatir el problema de continuidad en los pesos es definir  $w_{ii'}$  como una función continua de  $d_{ii'}$ . Otra elección entonces puede ser mediante la función bi-cuadrado dada por

$$w_{ii'} = \begin{cases} [1 - (d_{ii'}/d_2)^2]^2 & \text{si } d_{ii'} < d_2 \\ 0 & \text{en otro caso} \end{cases}$$

donde  $d_2$  se conoce como el ancho de banda. Este peso o ponderación es particularmente útil porque de esta manera hay continuidad (Brunsdon et al. 1996, Brunsdon et al. 1998, Fotheringham et al. 1998).

En este caso, la función de distancia lee los puntos de las coordenadas espaciales,  $(w_x, w_y)$ , y genera una matriz de pesos  $\mathbf{W}$ . Algunas medidas de distancias que se pueden utilizar son: Euclidiana ( $d_{ii'} = \sqrt{(w_{x_i} - w_{x_{i'}})^2 + (w_{y_i} - w_{y_{i'}})^2}$ ), Chebyshev ( $d_{ii'} = \max\{|w_{x_i} - w_{x_{i'}}|, |w_{y_i} - w_{y_{i'}}|\}$ ), Bray-Curtis ( $d_{ii'} = (|w_{x_i} - w_{x_{i'}}| + |w_{y_i} - w_{y_{i'}}|) / (|w_{x_i} + w_{x_{i'}}| + |w_{y_i} + w_{y_{i'}}|)$ ) y Canberra ( $d_{ii'} = (|w_{x_i} - w_{x_{i'}}| + |w_{y_i} - w_{y_{i'}}|) / (|w_{x_i}| + |w_{x_{i'}}| + |w_{y_i}| + |w_{y_{i'}}|)$ ), entre otras.

La función enlace  $g(\cdot)$  dada en (5.1) es estrictamente monótona y doblemente diferenciable. Algunas posibles elecciones para la función enlace,  $g(\mu_{it})$ , son: logit,  $g(\mu_{it}) = \log\{\mu_{it}/(1 - \mu_{it})\}$ ; probit,  $g(\mu_{it}) = \Phi^{-1}(\mu_{it})$  donde  $\Phi(\cdot)$  es la función de distribución acumulativa de una variable normal estándar; complemento loglog (cloglog),  $g(\mu_{it}) = \log\{-\log(1 - \mu_{it})\}$ ; y log-log,  $g(\mu_{it}) = -\log\{-\log(\mu_{it})\}$ . Una rica discusión de las funciones enlace se presentan en Atkinson (1985) y McCullagh & Nelder (1989).

Varios de los anteriores casos se pueden considerar en una clase general de funciones de enlace, Aranda-Ordaz (1981) propuso una familia de funciones de enlace para analizar datos de proporciones, que viene dada por

$$\eta_{it} = g_\nu(\mu(\mathbf{s}_i, t)) = \log \left[ \frac{(1 - \mu(\mathbf{s}_i, t))^{-\nu} - 1}{\nu} \right]$$

donde  $\nu$  es una constante desconocida, que tiene como casos particulares: el modelo logístico cuando  $\nu = 1$  y el complemento loglog cuando  $\nu \rightarrow 0$ .

Otra forma general de funciones enlace, propuestas por Box (Box & Cox 1964) y utilizada principalmente para datos con media positiva, es la transformación de Box-Cox dada por

$$\eta_{it} = g_\nu(\mu(\mathbf{s}_i, t)) = \begin{cases} (\mu^\nu(\mathbf{s}_i, t))/\nu & \text{si } \nu > 0 \\ \log(\mu(\mathbf{s}_i, t)) & \text{si } \nu = 0 \end{cases}$$

Dentro del campo de modelos lineales, es usual trabajar con el modelo en su forma canónica  $\eta_{it}(\alpha(\mathbf{s}_i, t)) = \alpha(\mathbf{s}_i, t) = \alpha_{it}$ ,  $h_2(y(\mathbf{s}_i, t)) = y(\mathbf{s}_i, t)$ , que incluye un parámetro de dispersión  $\phi > 0$ . Específicamente, condicionando sobre las variables explicativas  $(\mathbf{v}_{it})$  y el error espacial no observado  $\varepsilon_{it}$ ,  $y(\mathbf{s}_i, t)$  sigue una distribución de la familia exponencial, es decir,

$$\begin{aligned} y(\mathbf{s}_i, t) | \mathbf{v}_1(\mathbf{s}_i), \mathbf{v}_2(\mathbf{s}_i, t), \varepsilon_{it} &\stackrel{ind}{\sim} f(y(\mathbf{s}_i, t) | \mathbf{v}_1(\mathbf{s}_i), \mathbf{v}_2(\mathbf{s}_i, t), \varepsilon_{it}) \\ f(y(\mathbf{s}_i, t) | \mathbf{v}_1(\mathbf{s}_i), \mathbf{v}_2(\mathbf{s}_i, t), \varepsilon_{it}) &= \exp \left\{ \frac{1}{\phi} [y(\mathbf{s}_i, t)\alpha_{it} - b(\alpha_{it})] + c(y(\mathbf{s}_i, t), \phi) \right\} \end{aligned} \quad (5.2)$$

donde  $\phi$  es un parámetro de extra-variación o dispersión y  $c(\cdot)$  es una función específica. La media condicional,  $\mu_{it}$ , se relaciona con  $\alpha_{it}$  a través de la identidad  $\mu_{it} = \frac{\partial b(\alpha_{it})}{\partial \alpha_{it}}$ . Esto es modelado, después de hacer una transformación adecuada, utilizando el GLM presentado en (5.1) considerando los efectos fijos espacio-tiempo ( $\mu_{it}$ ) y los efectos aleatorios espacio-tiempo ( $\varepsilon_{it}$ ). La Tabla 5.1 muestra cómo se pueden reescribir en la forma (5.2) algunas de las distribuciones exponenciales más populares.

En forma matricial, la ecuación (5.1) se puede escribir para cada período de tiempo  $t$  ( $t = 1, \dots, T$ ) como

$$\begin{aligned}\boldsymbol{\eta}_t &= \mathbf{V}_1 \boldsymbol{\gamma}_0 + \mathbf{V}_{2t} \boldsymbol{\gamma}_t + \pi_t \mathbf{W}_1 \boldsymbol{\eta}_t + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\varepsilon}_t &= \psi_t \mathbf{W}_2 \boldsymbol{\varepsilon}_t + \mathbf{e}_t\end{aligned}\quad (5.3)$$

donde  $\boldsymbol{\eta}_t = (\eta_{1t}, \dots, \eta_{nt})^t$  es un vector  $n \times 1$ ,  $\mathbf{V}_1 = (\mathbf{v}_{11}, \dots, \mathbf{v}_{1n})^t$  es una matriz  $n \times (p_1 + 1)$  de variables explicativas espaciales,  $\mathbf{V}_{2t} = (\mathbf{v}_{21t}, \dots, \mathbf{v}_{2nt})^t$  es una matriz  $n \times p_{2t}$  de variables explicativas espacio-tiempo,  $\mathbf{e}_t = (e_{1t}, \dots, e_{nt})^t$  es un vector  $n \times 1$  y  $\mathbf{W}_1$  and  $\mathbf{W}_2$  son matrices  $n \times n$  que describen el rezago espacial y el error espacial de las unidades observadas en la muestra, respectivamente.

Observe en la ecuación (5.3) que el número de variables explicativas para  $\mathbf{V}_{2t}$ ,  $p_{2t}$ , pueden ser diferentes para período de tiempo. Además, el GLM espacio-tiempo sólo debe utilizarse puede ser puesto en funcionamiento sólo cuando hay más observaciones en la dimensión espacial que en la dimensión del tiempo ( $n > T$ ). En el caso más típico donde  $T > n$ , se aplica la habitual regresión no correlacionada utilizando el GLM.

Por lo tanto, se puede reformular el modelo (5.1) en forma vectorial como

$$\begin{aligned}\boldsymbol{\eta}_{st} &= g(\boldsymbol{\mu}_{st}) = g\{\mathbf{E}(\mathbf{y}_{st} \mid \mathbf{V}_{st}, \boldsymbol{\varepsilon}_{st})\} = \mathbf{V}_{st} \boldsymbol{\gamma} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1) \boldsymbol{\eta}_{st} + \boldsymbol{\varepsilon}_{st} \\ \boldsymbol{\varepsilon}_{st} &= (\boldsymbol{\Psi} \otimes \mathbf{W}_2) \boldsymbol{\varepsilon}_{st} + \mathbf{e}_{st}\end{aligned}\quad (5.4)$$

donde  $\otimes$  denota el producto Kronecker,

$$\begin{aligned}\mathbf{V}_{st} &= \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_{21} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{V}_1 & \mathbf{0} & \mathbf{V}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V}_{2T} \end{bmatrix}, & \boldsymbol{\gamma} &= \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_T \end{bmatrix}_{nT \times 1}, \\ \boldsymbol{\Pi} &= \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_T \end{bmatrix}_{T \times T}, & \boldsymbol{\Psi} &= \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_T \end{bmatrix}_{T \times T}, \\ \boldsymbol{\varepsilon}_{st} &= \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}_{nT \times 1}, & \boldsymbol{\eta}_{st} &= \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_T \end{bmatrix}_{nT \times 1}\end{aligned}$$

TABLA 5.1: Características de algunas distribuciones univariadas comunes en la familia exponencial como las dadas en (5.2) o (6.5)

Distribución	Notación	$\phi$	$\alpha_{it}(\mu_{it})$	$b(\alpha_{it})$	$c(y(\mathbf{s}_i, t), \phi)$	$\mu_{st}(\alpha_{it})$	$\text{Var}(\mu_{it})$
Normal	$N(\mu_{it}, \sigma^2)$	$\sigma^2$	$\mu_{it}$	$\frac{\alpha_{it}^2}{2}$	$-\frac{1}{2} \left[ \frac{y(\mathbf{s}_i, t)^2}{\sigma^2} + \log(2\pi\sigma^2) \right]$	$\alpha_{it}$	1
Poisson	$P(\mu_{it})$	1	$\log \mu_{it}$	$e^{\alpha_{it}}$	$-\log y(\mathbf{s}_i, t)!$	$\exp(\alpha_{it})$	$\mu_{it}$
Binomial	$B(n, \mu_{it})$	$1/n$	$\log \left( \frac{\mu_{it}}{1 - \mu_{it}} \right)$	$\log(1 + e^{\alpha_{it}})$	$\log \binom{n}{y(\mathbf{s}_i, t)}$	$\frac{1}{(1 + e^{-\alpha_{it}})}$	$\mu_{it}(1 - \mu_{it})$
Bin. Neg.	$\text{BN}(\mu_{it}, \kappa)$	1	$\log \left( \frac{\mu_{it}}{\mu_{it} + \kappa} \right)$	$-\kappa \log(1 - e^{\alpha_{it}})$	$\log \left[ \frac{\Gamma(\kappa + y(\mathbf{s}_i, t))}{\Gamma(\kappa) y(\mathbf{s}_i, t)!} \right]$	$\kappa \frac{e^{\alpha_{it}}}{1 - e^{\alpha_{it}}}$	$\mu_{it} \left( \frac{\mu_{it}}{\kappa} + 1 \right)$
Gamma	$G(\mu_{it}, v)$	$v^{-1}$	$-\frac{1}{\mu_{it}}$	$-\log(-\alpha_{it})$	$v \log(y(\mathbf{s}_i, t)) - \log \Gamma(v)$	$-\frac{1}{\alpha_{it}}$	$\mu_{it}^2$
Inv. Gau.	$\text{IG}(\mu_{it}, \sigma^2)$	$\sigma^2$	$-\frac{1}{2\mu_{it}^2}$	$-( -2\alpha_{it} )^{1/2}$	$-\log y(\mathbf{s}_i, t)$	$(-2\alpha_{it})^{-1/2}$	$\mu_{it}^3$
					$-\frac{1}{2} \log(2\pi\sigma^2 y^3(\mathbf{s}_1, t))$		
					$\frac{1}{1 - \frac{1}{2\sigma^2 y(\mathbf{s}_i, t)}}$		

$\phi$  es el parámetro de dispersión,  $\alpha_{it}(\mu_{it})$  es el enlace canónico,  $b(\alpha_{it})$  es la función cumulante,  $\text{Var}(\mu_{it})$  es la función varianza

con  $\boldsymbol{\eta}_t = (\eta_{1t}, \dots, \eta_{nt})^t$  un vector  $n \times 1$  ( $t = 1, \dots, T$ ),  $\boldsymbol{\mu}_{st} = \mathbb{E}(\mathbf{y}_{st} \mid \mathbf{V}_{st}, \boldsymbol{\varepsilon}_{st})$ ,  $\boldsymbol{\mu}_{st} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_T)^t$  es un vector  $nT \times 1$  con  $\boldsymbol{\mu}_t = (\mu_{1t}, \dots, \mu_{nt})^t$  un vector  $n \times 1$  y  $\mathbf{y}_{st} = (\mathbf{y}_1, \dots, \mathbf{y}_T)^t$  es un vector con  $nT \times 1$  con  $\mathbf{y}_t = (y(\mathbf{s}_1, t), \dots, y(\mathbf{s}_n, t))^t$  un vector  $n \times 1$ .  $\mathbf{V}_{st} = ((\mathbf{1}_n \vdash \mathbf{I}_T) \otimes \mathbf{H}_n) \mathbf{V}_{st}^*$  es una matriz  $nT \times p^*$  con  $p^* = p_1 + \sum_{t=1}^T p_{2t} + 1$ ,  $\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t$  una matriz centrada,  $\mathbf{I}_n$  la matriz identidad de tamaño  $n \times n$ ,  $\mathbf{1}_n$  un vector de unos de tamaño  $n \times 1$ ,  $\mathbf{I}_T$  la matriz identidad de tamaño  $T \times T$  y  $\mathbf{V}_{st}^*$  una matriz de variables explicativas originales; observe que  $\mathbf{V}_{st}^*$  puede estar conformada por variables continuas, categóricas y binarias, o incluso una mezcla de ellas. Además  $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \dots, \varepsilon_{nt})^t$  es un vector  $n \times 1$  y  $\mathbf{e}_{st} = (\mathbf{e}_1, \dots, \mathbf{e}_T)^t \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_T \otimes \mathbf{I}_n)$ , y éste es un vector  $nT \times 1$  con  $\mathbf{e}_t = (e_{1t}, \dots, e_{nt})^t$  un vector de  $n \times 1$ .

El modelo presentado en la ecuación (5.4) es un modelo lineal generalizado autorregresivo espacio-tiempo con perturbación autorregresiva espacio-tiempo (generalised linear space-time-autoregressive models with space-time-autoregressive disturbances, GLSTARAR), que incluye regresores exógenos. Las interacciones espacio-tiempo son modeladas a través de rezagos espacio-tiempo y errores espacio-tiempo. Además, el modelo permite interacciones espacio-tiempo en la variable dependiente, las variables exógenas y las perturbaciones.

El modelo (5.4) se puede expresar como

$$\begin{aligned} \boldsymbol{\eta}_{st} &= \mathbf{V}_{st} \boldsymbol{\gamma} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1) \boldsymbol{\eta}_{st} + [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1} \mathbf{e}_{st} \\ [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)] \boldsymbol{\eta}_{st} &= [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)] \mathbf{V}_{st} \boldsymbol{\gamma} + [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)] \\ &\quad \times (\boldsymbol{\Pi} \otimes \mathbf{W}_1) \boldsymbol{\eta}_{st} + \mathbf{e}_{st} \\ \boldsymbol{\eta}_{st} &= \mathbf{V}_{st} \boldsymbol{\gamma} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2) \mathbf{V}_{st} \boldsymbol{\gamma} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1) \boldsymbol{\eta}_{st} \\ &\quad + (\boldsymbol{\Psi} \otimes \mathbf{W}_2) \boldsymbol{\eta}_{st} - (\boldsymbol{\Psi} \boldsymbol{\Pi} \otimes \mathbf{W}_2 \mathbf{W}_1) \boldsymbol{\eta}_{st} + \mathbf{e}_{st} \end{aligned} \quad (5.5)$$

donde  $\mathbf{I}_{nT} = \mathbf{I}_T \otimes \mathbf{I}_n$ . Esta forma del modelo es válida siempre que  $[\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]$  sea una matriz no singular, lo cual es válido porque  $|\psi_t| < 1$  para  $t = 1, \dots, T$ .

Entonces realizando algunos procedimientos algebraicos en (5.5), se obtiene

$$\begin{aligned} \boldsymbol{\eta}_{st} &= \mathbf{V}_{st} \boldsymbol{\gamma} - (\mathbf{I}_T \otimes \mathbf{W}_2 \mathbf{V}_1) (\boldsymbol{\psi} \otimes \boldsymbol{\gamma}_0) - \left( \bigoplus_{t=1}^T \mathbf{W}_2 \mathbf{V}_{2t} \right) (\boldsymbol{\psi} \otimes \boldsymbol{\gamma}^*) \\ &\quad + \left( \bigoplus_{t=1}^T \mathbf{W}_1 \boldsymbol{\eta}_t \right) \boldsymbol{\pi} + \left( \bigoplus_{t=1}^T \mathbf{W}_2 \boldsymbol{\eta}_t \right) \boldsymbol{\psi} - \left( \bigoplus_{t=1}^T \mathbf{W}_2 \mathbf{W}_1 \boldsymbol{\eta}_t \right) (\boldsymbol{\psi} \odot \boldsymbol{\pi}) + \mathbf{e}_{st} \\ &= \mathbf{M}_1 \boldsymbol{\delta}_1 + \mathbf{M}_2 \boldsymbol{\delta}_2 + \mathbf{M}_3 \boldsymbol{\delta}_3 + \mathbf{M}_4 \boldsymbol{\delta}_4 + \mathbf{M}_5 \boldsymbol{\delta}_5 + \mathbf{M}_6 \boldsymbol{\delta}_6 + \mathbf{e}_{st} \\ &= \mathbf{M} \boldsymbol{\delta} + \mathbf{e}_{st} \end{aligned} \quad (5.6)$$

donde  $\odot$  denota el producto Hadamard,  $\bigoplus$  denota la suma directa,  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_T)^t$ ,  $\boldsymbol{\gamma}^* = (\gamma_1^t, \dots, \gamma_T^t)$ ,  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_T)^t$ ,  $\mathbf{M}_1 = \mathbf{V}_{st}$ ,  $\mathbf{M}_2 =$



$\mathbf{I}_T \otimes \mathbf{W}_2 \mathbf{V}_1$ ,  $\mathbf{M}_3 = \bigoplus_{t=1}^T \mathbf{W}_2 \mathbf{V}_{2t}$ ,  $\mathbf{M}_4 = \bigoplus_{t=1}^T \mathbf{W}_1 \boldsymbol{\eta}_t$ ,  $\mathbf{M}_5 = \bigoplus_{t=1}^T \mathbf{W}_2 \mathbf{W}_1 \boldsymbol{\eta}_t$ ,  $\mathbf{M}_6 = \bigoplus_{t=1}^T \mathbf{W}_2 \mathbf{W}_1 \boldsymbol{\eta}_t$ ,  $\boldsymbol{\delta}_1 = \boldsymbol{\gamma}$ ,  $\boldsymbol{\delta}_2 = -\boldsymbol{\psi} \otimes \boldsymbol{\gamma}_0$ ,  $\boldsymbol{\delta}_3 = -\boldsymbol{\psi} \otimes \boldsymbol{\gamma}^*$ ,  $\boldsymbol{\delta}_4 = \boldsymbol{\pi}$ ,  $\boldsymbol{\delta}_5 = \boldsymbol{\psi}$ ,  $\boldsymbol{\delta}_6 = -\boldsymbol{\psi} \odot \boldsymbol{\pi}$ ,  $\mathbf{M} = (\mathbf{M}_1, \dots, \mathbf{M}_6)$  y  $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^t, \dots, \boldsymbol{\delta}_6^t)$ . Las restricciones sobre los parámetros en el modelo (5.6) son:  $\boldsymbol{\delta}_2 = -\boldsymbol{\delta}_5 \otimes \boldsymbol{\delta}_{1(0)}$ ,  $\boldsymbol{\delta}_3 = -\boldsymbol{\delta}_4 \otimes \boldsymbol{\delta}_{1(1)}$  y  $\boldsymbol{\delta}_6 = -\boldsymbol{\delta}_5 \odot \boldsymbol{\delta}_4$ , donde  $\boldsymbol{\delta}_1 = (\boldsymbol{\delta}_{1(0)}^t, \boldsymbol{\delta}_{1(1)}^t)$ .

### 5.2.1 Modelo lineal generalizado espacio-tiempo basado en distancias

La matriz de variables explicativas en el modelo (5.4),  $\mathbf{V}_{st}$ , puede involucrar una elevada dimensión de parámetros que se pueden estimar. Entonces, se necesita reducir el espacio abarcado por las variables explicativas. Por lo tanto, como solución a dicho problema para construir el modelo se deben incorporar varias medidas generales de distancia/disimilitud que se pueden aplicar a las variables explicativas: continuas, categórica, o una mezcla de ellas. Para ello, se necesitan definir algunas medidas de similitud (o de distancia Euclídea), que dependen de las características de variables explicativas.

De acuerdo a Cuadras (1989) y Cuadras & Arenas (1990), sea  $\Omega = \{\omega_1, \dots, \omega_{nT}\}$  un conjunto que consiste de  $nT$  elementos. Sea  $d_{ii'} = d(\omega_i, \omega_{i'}) = d(\omega_{i'}, \omega_i) \geq d(\omega_i, \omega_i) = 0$  una función de distancia (o disimilaridad) definida sobre  $\Omega$ . Suponga que la matriz de inter-distancias (o inter-disimilaridades) con dimensión  $nT \times nT$ ,  $\mathbf{D} = (d_{ii'})$  es Euclidiana. Entonces, existe una configuración de puntos  $\mathbf{v}_1, \dots, \mathbf{v}_{nT} \in \mathbb{R}^{p^*}$ , (donde  $\mathbf{v}_i = (\mathbf{v}_1^t(\mathbf{s}_i), \mathbf{v}_2^t(\mathbf{s}_i, 1), \dots, \mathbf{v}_2^t(\mathbf{s}_i, T)) = (1, v_{i1}, \dots, v_{ip_1}, v_{i11}, \dots, v_{i1p_{21}}, \dots, v_{iT1}, \dots, v_{iT p_{2T}})^t$ ,  $i = 1, \dots, nT$ , el cual esta formado por variables binarias, categóricas y continuas) tal que la similitud de acuerdo a Gower (1968) se puede definir para variables mixtas como

$$m_{ii'} = \frac{\sum_{j=1}^{p_c} \left(1 - \frac{|v_{ij} - v_{i'j}|}{G_j}\right) + c_{1ii'} + v_{ii'}}{p_c + (p_b - c_{4ii'}) + p_q} \quad (5.7)$$

donde, para los efectos fijos,  $p_c$  es el número de variables continuas,  $c_{1ii'}$  y  $c_{4ii'}$  son el número de coincidencias positivas y negativas, respectivamente, para las  $p_b$  variables binarias, y  $v_{ii'}$  es el número de coincidencias para las  $p_q$  variables multiestado.  $G_j$  es el rango (o distancia) de la  $j$ -ésima variable continua.

En el caso en que las variables explicativas en (5.4) sean binarias o categóricas, las similitudes se pueden definir como

$$m_{ii'} = \frac{c_{1ii'} + c_{4ii'}}{c_{1ii'} + c_{2ii'} + c_{3ii'} + c_{4ii'}} \quad (\text{Sokal-Michener})$$

$$m_{ii'} = \frac{c_{1ii'}}{c_{1ii'} + c_{2ii'} + c_{3ii'}} \quad (\text{Jaccard})$$

donde  $c_{1ii'}$ ,  $c_{2ii'}$ ,  $c_{3ii'}$ ,  $c_{4ii'}$  son las frecuencias de (1,1), (1,0), (0,1) y (0,0), respectivamente. A través de la transformación

$$d_{ii'} = \sqrt{1 - m_{ii'}}$$

es posible obtener distancias Euclidianas. Si todas las variables explicativas en (5.4) son continuas, se puede definir la distancia como

$$d_{ii'} = \sqrt{(\mathbf{v}_i - \mathbf{v}_{i'})^t (\mathbf{v}_i - \mathbf{v}_{i'})} \quad (5.8)$$

o alternativamente por la distancia valor absoluto  $d_{ii'} = \sqrt{\sum_{j=1}^{p^*} |v_{ij} - v_{i'j}|}$ . Expresiones para la similaridad de Gower como la presentada en la ecuación (5.7) es útil en la medida que se tenga información asociada con variables explicativas mixtas, no sólo para los sitios de muestreo sino también para los lugares no muestreados, lo que limita su uso en áreas no muestreadas.

Estas distancias satisfacen que la distancia está cerca a 0 si las mediciones de  $\mathbf{v}$  en  $i$  e  $i'$  son muy similares, es decir  $d_{ii'} \cong 0$  si  $\mathbf{v}_i \cong \mathbf{v}_{i'}$ . Después de seleccionar una de las anteriores distancias, se define  $\mathbf{A}_{nT \times nT} = (-d_{ii'}^2/2)$  y  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$ . Entonces,  $\mathbf{B}$  es una matriz semidefinida positiva (Mardia et al. 2002) de rango  $nT - 1$  y la matriz de coordenadas principales,  $\mathbf{X}$ , se obtiene a partir de la siguiente descomposición espectral

$$\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t = \mathbf{X}\mathbf{X}^t$$

donde  $\mathbf{\Lambda}$  es una matriz diagonal que contiene los valores propios de  $\mathbf{B}$  y  $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}$  es una matriz  $nT \times nT$  de rango  $nT - 1$  ya que ésta tiene un vector propio igual a  $\mathbf{1}_{nT}$  y  $\mathbf{U}$  contiene las coordenadas estandarizadas.

Además, las filas  $\mathbf{x}_1^t, \dots, \mathbf{x}_{nT}^t$  de  $\mathbf{X}$  son las coordenadas principales de  $\mathbf{B}$  con respecto a la matriz de distancia  $\mathbf{D}$ . De la misma manera que  $\mathbf{v}_i \cong \mathbf{v}_{i'}$  cuando un individuo  $i$  es similar a otro individuo  $i'$  en (5.4), es claro que también  $\mathbf{x}_i \cong \mathbf{x}_{i'}$ .

Uno de los peligros potenciales en la predicción basada en distancia es la enorme sobreparametrización ya que el rango de  $\mathbf{B}$  puede ser tan grande como  $nT - 1$ . Entonces, el número de coordenadas principales (columnas de  $\mathbf{X}$ ) puede ser excesivo, lo que permite un modelo arbitrariamente sobre-ajustado.

Con el fin de evitar este tipo de problemas, se seleccionan únicamente las coordenadas principales más significativas utilizando cualquier método de la Sección 5.4.3. Por lo tanto, en forma matricial, los modelos lineales generalizados autorregresivos espacio-tiempo basados en distancias con perturbaciones autorregresivas espacio-tiempo (DBGLSTARAR) en forma reducida se pueden expresar como

$$\begin{aligned} \boldsymbol{\eta}_{st} &= g(\boldsymbol{\mu}_{st}) = g\{\mathbf{E}(\mathbf{y}_{st} \mid \mathbf{X}_{st}, \boldsymbol{\varepsilon}_{st})\} = \mathbf{X}_{st}\boldsymbol{\beta} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1)\boldsymbol{\eta}_{st} + \boldsymbol{\varepsilon}_{st} \\ \boldsymbol{\varepsilon}_{st} &= (\boldsymbol{\Psi} \otimes \mathbf{W}_2)\boldsymbol{\varepsilon}_{st} + \mathbf{e}_{st} \end{aligned} \quad (5.9)$$

donde  $\boldsymbol{\mu}_{st} = E(\mathbf{y}_{st} \mid \mathbf{X}_{st}, \boldsymbol{\varepsilon}_{st})$ ,  $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \dots, \beta_k)$  es un vector de parámetros desconocido  $(k+1) \times 1$ ,  $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ ,  $k \leq nT - 1$ ,  $\mathbf{X}_{st}$  contiene un subconjunto de  $k+1$  columnas relevantes de  $\mathbf{X}$  las cuales son las coordenadas principales significativamente correlacionadas con  $\mathbf{y}_{st}$ , incluyendo un vector de  $\mathbf{1}_{nT}$ . Por otra parte, cada  $\mathbf{X}_j$  ( $j = 1, \dots, k$ ) es una coordenada principal, es decir, un vector columna de  $\mathbf{X}_{st}$ . Observe que  $\mathbf{X}_j^t \mathbf{1}_{nT} = 0$ ,  $\mathbf{X}_j^t \mathbf{X}_j = \lambda_j$  y  $\mathbf{X}_j^t \mathbf{X}_{j'} = 0$  para  $j \neq j'$  con  $j, j' = 1, \dots, k$ .

El modelo presentado en la ecuación (5.9) es un DBGLSTARAR e incluye los regresores de coordenadas principales. Las interacciones espacio-tiempo son modeladas a través de rezagos espacio-tiempo y errores espacio-tiempo. Además, el modelo permite interacciones espacio-tiempo en la variable dependiente, las coordenadas principales y las perturbaciones.

Alternativamente, el modelo (5.9) se puede expresar como

$$\begin{aligned}\eta_{it} &= \mathbf{x}_{it}^t \boldsymbol{\beta} + \pi_t \sum_{i'=1}^n w_{ii'}^{(1)} \eta_{i't} + \varepsilon_{it} \\ \varepsilon_{it} &= \psi_t \sum_{i'=1}^n w_{ii'}^{(2)} \varepsilon_{i't} + e_{it}\end{aligned}\tag{5.10}$$

con  $i = 1, \dots, n$  y  $t = 1, \dots, T$ , y donde  $\mathbf{x}_{it}^t = (1, x_{i0}, \dots, x_{ik})$ ,  $|\pi_t| < 1$  y  $|\psi_t| < 1$ .

Los modelos en (5.9) se pueden reescribir como

$$\begin{aligned}\boldsymbol{\eta}_{st} &= \mathbf{X}_{st} \boldsymbol{\beta} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1) \boldsymbol{\eta}_{st} + \boldsymbol{\varepsilon}_{st} = \mathbf{X}_{st} \boldsymbol{\beta} + \left( \bigoplus_{t=1}^T \mathbf{W}_1 \boldsymbol{\eta}_t \right) \boldsymbol{\pi} + \boldsymbol{\varepsilon}_{st} \\ &= \mathbf{X} \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_{st}\end{aligned}\tag{5.11}$$

$$\begin{aligned}\boldsymbol{\varepsilon}_{st} &= (\boldsymbol{\Psi} \otimes \mathbf{W}_2) \boldsymbol{\eta}_{st} + \mathbf{e}_{st} = \left( \bigoplus_{t=1}^T \mathbf{W}_2 \boldsymbol{\varepsilon}_t \right) \boldsymbol{\psi} + \mathbf{e}_{st} \\ &= \boldsymbol{\varepsilon}_{st}^* \boldsymbol{\psi} + \mathbf{e}_{st}\end{aligned}\tag{5.12}$$

donde  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_T)^t$ ,  $\mathbf{X} = (\mathbf{X}_{st}, \bigoplus_{t=1}^T \mathbf{W}_1 \boldsymbol{\eta}_t)$ ,  $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^t, \boldsymbol{\pi}^t)$ ,  $\boldsymbol{\varepsilon}_{st}^* = (\bigoplus_{t=1}^T \mathbf{W}_2 \boldsymbol{\varepsilon}_t)$  y  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_T)^t$ .

Desarrollando procedimientos similares a los presentados en las ecuaciones (5.5) y (5.6), los modelos (5.11) y (5.12) pueden ser expresados mediante

$$\begin{aligned}\boldsymbol{\eta}_{st} &= \mathbf{X}_{st} \boldsymbol{\beta} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1) \boldsymbol{\eta}_{st} + \boldsymbol{\varepsilon}_{st} \\ &= \mathbf{X}_{st} \boldsymbol{\beta} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1) \boldsymbol{\eta}_{st} + [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1} \mathbf{e}_{st}\end{aligned}\tag{5.13}$$

donde  $\boldsymbol{\varepsilon}_{st} \sim MN(\mathbf{0}, \text{Var}(\boldsymbol{\varepsilon}_{st}))$  con  $\text{Var}(\boldsymbol{\varepsilon}_{st}) = [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1} (\boldsymbol{\Sigma}_T \otimes \mathbf{I}_n) \{[\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1}\}^t$ .

Así, el modelo presentado en (5.13) se puede escribir equivalentemente como

$$\begin{aligned}
\boldsymbol{\eta}_{st} &= \mathbf{X}_{st}\boldsymbol{\beta} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)\mathbf{X}_{st}\boldsymbol{\beta} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1)\boldsymbol{\eta}_{st} + (\boldsymbol{\Psi} \otimes \mathbf{W}_2)\boldsymbol{\eta}_{st} \\
&\quad - (\boldsymbol{\Psi}\boldsymbol{\Pi} \otimes \mathbf{W}_2\mathbf{W}_1)\boldsymbol{\eta}_{st} + \mathbf{e}_{st} \\
&= \mathbf{X}_{st}\boldsymbol{\beta} - \left( \bigoplus_{t=1}^T \mathbf{W}_2\mathbf{X}_{2t} \right) (\boldsymbol{\psi} \otimes \boldsymbol{\beta}) + \left( \bigoplus_{t=1}^T \mathbf{W}_1\boldsymbol{\eta}_t \right) \boldsymbol{\pi} + \left( \bigoplus_{t=1}^T \mathbf{W}_2\boldsymbol{\eta}_t \right) \boldsymbol{\psi} \\
&\quad - \left( \bigoplus_{t=1}^T \mathbf{W}_2\mathbf{W}_1\boldsymbol{\eta}_t \right) (\boldsymbol{\psi} \odot \boldsymbol{\pi}) + \mathbf{e}_{st} \\
&= \mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{Z}_2\boldsymbol{\alpha}_2 + \mathbf{Z}_3\boldsymbol{\alpha}_3 + \mathbf{Z}_4\boldsymbol{\alpha}_4 + \mathbf{Z}_5\boldsymbol{\alpha}_5 + \mathbf{e}_{st} = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e}_{st} \quad (5.14)
\end{aligned}$$

donde  $\mathbf{Z}_1 = \mathbf{X}_{st}$ ,  $\mathbf{Z}_2 = \bigoplus_{t=1}^T \mathbf{W}_2\mathbf{X}_{2t}$ ,  $\mathbf{Z}_3 = \bigoplus_{t=1}^T \mathbf{W}_1\boldsymbol{\eta}_t$ ,  $\mathbf{Z}_4 = \bigoplus_{t=1}^T \mathbf{W}_2\mathbf{W}_1\boldsymbol{\eta}_t$ ,  $\mathbf{Z}_5 = \bigoplus_{t=1}^T \mathbf{W}_2\mathbf{W}_1\boldsymbol{\eta}_t$ ,  $\boldsymbol{\alpha}_1 = \boldsymbol{\beta}$ ,  $\boldsymbol{\alpha}_2 = -\boldsymbol{\psi} \otimes \boldsymbol{\beta}$ ,  $\boldsymbol{\alpha}_3 = \boldsymbol{\pi}$ ,  $\boldsymbol{\alpha}_4 = \boldsymbol{\psi}$ ,  $\boldsymbol{\alpha}_5 = -\boldsymbol{\psi} \odot \boldsymbol{\pi}$ ,  $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_5)$ , y  $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^t, \dots, \boldsymbol{\alpha}_5^t)$ . Las restricciones sobre los parámetros en el modelo (5.14) son:  $\boldsymbol{\alpha}_2 = -\boldsymbol{\alpha}_4 \otimes \boldsymbol{\alpha}_1$  y  $\boldsymbol{\alpha}_5 = -\boldsymbol{\alpha}_4 \odot \boldsymbol{\alpha}_3$ .

**Ejemplo 5.1.** El GLM autorregresivo espacio-tiempo basado en distancias sin perturbaciones autorregresivas espacio-tiempo está dado por

$$\begin{aligned}
\boldsymbol{\eta}_{st} &= \mathbf{X}_{st}\boldsymbol{\beta} + (\boldsymbol{\Pi} \otimes \mathbf{W}_1)\boldsymbol{\eta}_{st} + \mathbf{e}_{st} \\
&= \mathbf{X}_{st}\boldsymbol{\beta} + \left( \bigoplus_{t=1}^T \mathbf{W}_1\boldsymbol{\eta}_t \right) \boldsymbol{\pi} + \mathbf{e}_{st}
\end{aligned}$$

donde  $\mathbf{X}_{st}$ ,  $\boldsymbol{\beta}$ ,  $\boldsymbol{\Pi} \otimes \mathbf{W}_1$ ,  $\bigoplus_{t=1}^T \mathbf{W}_1\boldsymbol{\eta}_t$ , y  $\boldsymbol{\pi}$  son definidas como en la ecuación (5.14), y  $\mathbf{e}_{st} \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_T \otimes \mathbf{I}_n)$ .

**Ejemplo 5.2.** El GLM autorregresivo espacio-tiempo basado en distancias con perturbaciones está dado por

$$\begin{aligned}
\boldsymbol{\eta}_{st} &= \mathbf{X}_{st}\boldsymbol{\beta} + [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1}\mathbf{e}_{st} \\
[\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]\boldsymbol{\eta}_{st} &= [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]\mathbf{X}_{st}\boldsymbol{\beta} + \mathbf{e}_{st} \\
\boldsymbol{\eta}_{st} &= \mathbf{X}_{st}\boldsymbol{\beta} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)\mathbf{X}_{st}\boldsymbol{\beta} + (\boldsymbol{\Psi} \otimes \mathbf{W}_2)\boldsymbol{\eta}_{st} + \mathbf{e}_{st} \\
&= \mathbf{X}_{st}\boldsymbol{\beta} - \left( \bigoplus_{t=1}^T \mathbf{W}_2\mathbf{X}_{2t} \right) (\boldsymbol{\psi} \otimes \boldsymbol{\beta}) + \left( \bigoplus_{t=1}^T \mathbf{W}_2\boldsymbol{\eta}_t \right) \boldsymbol{\psi} + \mathbf{e}_{st}
\end{aligned}$$

donde  $\mathbf{X}_{st}$ ,  $\boldsymbol{\beta}$ ,  $\boldsymbol{\Psi} \otimes \mathbf{W}_2$ ,  $\bigoplus_{t=1}^T \mathbf{W}_2\boldsymbol{\eta}_t$  y  $\boldsymbol{\psi}$  son definidas como en la ecuación (5.14).

## 5.3 Métodos de estimación de los parámetros

### 5.3.1 Estimación por máxima verosimilitud

Asumiendo que cada  $Y(\mathbf{s}_i, t)$  en el modelo (5.11) tiene una función de distribución de la familia exponencial, y por independencia de  $Y(\mathbf{s}_1, t), \dots, Y(\mathbf{s}_n, t)$

dados  $\mathbf{X}$ ,  $\boldsymbol{\varepsilon}_{st}$  y  $g^{-1}(\cdot)$ , la función de densidad condicional de  $\mathbf{Y}_{st} = \mathbf{y}_{st}$  dadas las covariables observadas  $\mathbf{X}$  y  $\boldsymbol{\varepsilon}_{st}$  es

$$f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) = \prod_{i=1}^n \prod_{t=1}^T f\{y(\mathbf{s}_i, t); g^{-1}[\mathbf{x}(\mathbf{s}_i, t); \boldsymbol{\beta}^*]\}$$

Ahora, desde una perspectiva clásica, la función de verosimilitud basada en las variables aleatorias observadas  $\mathbf{y}_{st}$  se obtiene marginalizando con respecto a las variables aleatorias no observadas  $\boldsymbol{\varepsilon}_{st}$ , dando lugar a la verosimilitud del modelo mixto. Entonces la función de verosimilitud para un modelo de vector autorregresivo generalizado espacio-tiempo no se puede escribir en forma cerrada, sino sólo como una integral de alta dimensión

$$\begin{aligned} L(\boldsymbol{\beta}^*, \boldsymbol{\theta}) &= f(\mathbf{y}_{st} | \boldsymbol{\beta}^*, \boldsymbol{\theta}) = \int_{\mathbb{R}^{nT}} f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta}) d\boldsymbol{\varepsilon}_{st} \\ &= \int_{\mathbb{R}^{nT}} \prod_{i=1}^n \prod_{t=1}^T f\{y(\mathbf{s}_i, t); g^{-1}[\mathbf{x}(\mathbf{s}_i, t), \boldsymbol{\varepsilon}(\mathbf{s}_i, t); \boldsymbol{\beta}^*]\} \\ &\quad \times f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta}) d\varepsilon_{11} \cdots d\varepsilon_{nT} \end{aligned} \quad (5.15)$$

donde  $f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta})$  denota la función de distribución normal multivariada de  $\boldsymbol{\varepsilon}_{st}$  dadas las covariables observadas  $\mathbf{X}$ , con  $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\Sigma}_T)$  el conjunto de los parámetros asociados a  $\boldsymbol{\varepsilon}_{st}$ .

### Parámetros para los efectos fijos

A pesar de que las ecuaciones de verosimilitud son numéricamente un poco complejas, éstas se pueden escribir en una forma más simple. A partir de (5.15), el log de la verosimilitud esta dado por

$$\ell = \log \int_{\mathbb{R}^{nT}} f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta}) d\boldsymbol{\varepsilon}_{st} = f(\mathbf{y}_{st} | \boldsymbol{\beta}^*, \boldsymbol{\theta}) \quad (5.16)$$

de modo que

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\beta}^*} &= \frac{\partial}{\partial \boldsymbol{\beta}^*} \int_{\mathbb{R}^{nT}} \frac{f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta})}{f(\mathbf{y}_{st} | \boldsymbol{\beta}^*, \boldsymbol{\theta})} d\boldsymbol{\varepsilon}_{st} \\ &= \int_{\mathbb{R}^{nT}} \left[ \frac{\partial}{\partial \boldsymbol{\beta}^*} f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) \right] \frac{f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta})}{f(\mathbf{y}_{st} | \boldsymbol{\beta}^*, \boldsymbol{\theta})} d\boldsymbol{\varepsilon}_{st} \end{aligned} \quad (5.17)$$

ya que  $f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta})$  no involucra  $\boldsymbol{\beta}^*$ . Observe que

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}^*} f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) &= \left( \frac{1}{f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)} \frac{\partial f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} \right) \\ &\quad \times f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) \\ &= \frac{\partial \log f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) \end{aligned} \quad (5.18)$$

y así la ecuación (5.17) se puede expresar como

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\beta}^*} &= \int_{\mathbb{R}^{nT}} \frac{\partial \log f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) \frac{f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta})}{f(\mathbf{y}_{st} | \boldsymbol{\beta}^*, \boldsymbol{\theta})} d\boldsymbol{\varepsilon}_{st} \\ &= \int_{\mathbb{R}^{nT}} \frac{\partial \log f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} f(\boldsymbol{\varepsilon}_{st} | \mathbf{y}_{st}, \boldsymbol{\theta}) d\boldsymbol{\varepsilon}_{st}\end{aligned}\quad (5.19)$$

donde

$$\frac{\partial \log f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} = \frac{\partial}{\partial \boldsymbol{\beta}^*} \sum_{i=1}^n \sum_{t=1}^T \left\{ \frac{1}{\phi} [y_{i,t} \alpha_{it} - b(\alpha_{it})] + c(y_{it}, \phi) \right\}$$

Entonces, las ecuaciones score obtenidas a partir del análisis de la verosimilitud tienen la forma

$$\begin{aligned}\frac{\partial \log f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} &= \frac{1}{\phi} \sum_{i=1}^n \sum_{t=1}^T \left( y_{it} \frac{\partial \alpha_{it}}{\partial \boldsymbol{\beta}^*} - \frac{\partial b(\alpha_{it})}{\partial \alpha_{it}} \frac{\partial \alpha_{it}}{\partial \boldsymbol{\beta}^*} \right) \\ &= \frac{1}{\phi} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \mu_{it}) \frac{\partial \alpha_{it}}{\partial \boldsymbol{\beta}^*} \\ &= \frac{1}{\phi} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \mu_{it}) \frac{\partial \alpha_{it}}{\partial \mu_{it}} \frac{\partial \mu_{it}}{\partial \boldsymbol{\beta}^*} \\ &= \frac{1}{\phi} \sum_{i=1}^n \sum_{t=1}^T \frac{(y_{it} - \mu_{it})}{v(\mu_{it})} \frac{\partial \mu_{it}}{\partial \eta_{it}} \frac{\partial \eta_{it}}{\partial \boldsymbol{\beta}^*} \\ &= \frac{1}{\phi} \sum_{i=1}^n \sum_{t=1}^T \frac{(y_{it} - \mu_{it})}{v(\mu_{it}) g'(\mu_{it})} \mathbf{x}_{it}^t \\ &= \frac{1}{\phi} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \mu_{it}) \xi_{it} g'(\mu_{it}) \mathbf{x}_{it}^t\end{aligned}\quad (5.20)$$

ya que  $E(y_{it}) = \mu_{it} = \partial b(\alpha_{it}) / \partial \alpha_{it}$ ,  $\text{Var}(y_{it}) = \phi \partial^2 b(\alpha_{it}) / \partial \alpha_{it}^2 = \phi v(\mu_{it})$  con  $v(\mu_{it})$  una función de varianza,  $\partial \alpha_{it} / \partial \mu_{it} = (\partial \mu_{it} / \partial \alpha_{it})^{-1} = (\partial^2 b(\alpha_{it}) / \partial \alpha_{it}^2)^{-1} = 1/v(\mu_{it})$ ,  $\partial \mu_{it} / \partial \eta_{it} = \partial \mu_{it} / \partial g(\mu_{it}) = (\partial g(\mu_{it}) / \partial \mu_{it})^{-1} = (g'(\mu_{it}))^{-1}$  y  $\partial \eta_{it} / \partial \boldsymbol{\beta}^* = \partial g(\mu_{it}) / \partial \mu_{it} = \mathbf{x}_{it}^t$  donde  $\mathbf{x}_{it}^t$  es la  $it$ -ésima fila de  $\mathbf{X}$ . Además,  $\xi_{it} = \{v(\mu_{it}) [g'(\mu_{it})]^2\}^{-1}$ .

La ecuación (5.20) se puede expresar en notación matricial como

$$\frac{\partial \log f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} = \frac{1}{\phi} \mathbf{X}^t \boldsymbol{\Xi} \boldsymbol{\Delta} (\mathbf{y}_{st} - \boldsymbol{\mu}_{st})\quad (5.21)$$

donde  $\boldsymbol{\Xi} = \text{diag}(\xi_{it})$  y  $\boldsymbol{\Delta} = \text{diag}(g'(\mu_{it}))$ ,  $i = 1, \dots, n$  y  $j = 1, \dots, T$ .

Entonces, sustituyendo (5.21) en (5.19), se obtiene

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\beta}^*} &= \int_{\mathbb{R}^{nT}} \frac{1}{\phi} \mathbf{X}^t \boldsymbol{\Xi} \boldsymbol{\Delta} (\mathbf{y}_{st} - \boldsymbol{\mu}_{st}) f(\boldsymbol{\varepsilon}_{st} | \mathbf{y}_{st}, \boldsymbol{\theta}) d\boldsymbol{\varepsilon}_{st} \\ &= \mathbf{X}^t E(\boldsymbol{\Xi}^* | \mathbf{y}_{st}) - \mathbf{X}^t E(\boldsymbol{\Xi}^* \boldsymbol{\mu}_{st} | \mathbf{y}_{st})\end{aligned}\quad (5.22)$$

donde  $\Xi^* = \text{diag}(\{\phi v(\mu_{it})g'(\mu_{it})\}^{-1})$  y  $E(\cdot | \mathbf{y}_{st})$  es el valor esperado condicional dado  $\mathbf{y}_{st}$ .

La ecuación de verosimilitud para  $\beta^*$  es por lo tanto

$$\mathbf{X}^t E(\Xi^* | \mathbf{y}_{st}) = \mathbf{X}^t E(\Xi^* \boldsymbol{\mu}_{st} | \mathbf{y}_{st}) \quad (5.23)$$

Típicamente, estas ecuaciones son funciones no lineales de  $\beta^*$ , y así, (5.23) no se puede solucionar analíticamente.

La solución a las ecuaciones de máxima verosimilitud (maximum likelihood, ML), (5.23), para  $\beta^*$  se lleva a cabo mediante un proceso iterativo ponderado de mínimos cuadrados. Esto se puede obtener como un ejemplo del método de Fisher-Scoring (Searle et al. 1987, p. 295). Este método se utiliza para maximizar una verosimilitud y tiene la siguiente forma

$$\begin{aligned} \beta^{*(m+1)} &= \beta^{*(m)} - \left\{ E \left[ \frac{\partial^2 \ln f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \beta^{*(m)})}{\partial \beta^* \partial \beta^{*t}} \middle| \mathbf{y}_{st} \right] \right\}^{-1} \\ &\quad \times E \left[ \frac{\partial \ln f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \beta^{*(m)})}{\partial \beta^*} \middle| \mathbf{y}_{st} \right] \\ &= \beta^{*(m)} + (\mathbf{X}^t \Xi^* \mathbf{X})^{-1} \mathbf{X}^t [E(\Xi^* | \mathbf{y}_{st}) - E(\Xi^* \boldsymbol{\mu}_{st} | \mathbf{y}_{st})] \end{aligned} \quad (5.24)$$

### Parámetros para los efectos aleatorios

Una vez estimados los parámetros fijos en el modelo (5.11),  $\beta^*$ , se está listo para estimar los parámetros  $\boldsymbol{\theta}$ . De la ecuación (5.11), se encuentra que  $\hat{\boldsymbol{\varepsilon}}_{st} = \boldsymbol{\eta}_{st} - \mathbf{X}\hat{\boldsymbol{\beta}}^*$ . Entonces, utilizando el modelo presentado en (5.12) se obtiene

$$\hat{\boldsymbol{\varepsilon}}_{st} = \hat{\boldsymbol{\varepsilon}}_{st}^* \boldsymbol{\psi} + \mathbf{e}_{st} \quad (5.25)$$

donde  $\mathbf{e}_{st} \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_T)$  con  $\boldsymbol{\Sigma}_T = \sigma^2 \mathbf{R}(\boldsymbol{\vartheta})$  una matriz simétrica  $T \times T$  que cumple con el requisito de ser una matriz de covarianzas.  $\mathbf{R}(\boldsymbol{\vartheta})$  es una matriz  $T \times T$  que cumple con el requisito de ser una matriz de covarianzas y  $\boldsymbol{\vartheta}$  es un vector  $s \times 1$  que caracteriza plenamente  $\mathbf{R}(\boldsymbol{\vartheta})$ . Entonces, para el modelo (5.25), los estimadores de los parámetros obtenidos con el método de máxima verosimilitud están dados por

$$\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\varepsilon}}_{st}^{*t} \hat{\boldsymbol{\varepsilon}}_{st}^*)^{-1} \hat{\boldsymbol{\varepsilon}}_{st}^{*t} \hat{\boldsymbol{\varepsilon}}_{st} \quad (5.26)$$

$$\hat{\boldsymbol{\Sigma}}_T = \frac{1}{n} (\hat{\mathbf{E}} - \hat{\mathbf{E}}_{\hat{\boldsymbol{\psi}}})^t (\hat{\mathbf{E}} - \hat{\mathbf{E}}_{\hat{\boldsymbol{\psi}}}) \quad (5.27)$$

donde  $\hat{\mathbf{E}} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_T)$  es una matriz  $n \times T$  y  $\hat{\mathbf{E}}_{\hat{\boldsymbol{\psi}}} = (\psi_1 \mathbf{W}_2 \boldsymbol{\varepsilon}_1, \dots, \psi_T \mathbf{W}_2 \boldsymbol{\varepsilon}_T)$  es una matriz  $n \times T$ .

Como el estimador dado en la ecuación (5.27) es sesgado, se puede utilizar el siguiente estimador insesgado para  $\Sigma_T$ ,

$$\widehat{\Sigma}_T = \frac{1}{n-T} (\widehat{\mathbf{E}} - \widehat{\mathbf{E}}_{\hat{\psi}})^t (\widehat{\mathbf{E}} - \widehat{\mathbf{E}}_{\hat{\psi}}) \quad (5.28)$$

Sin embargo, se puede elegir cualquier otro específico  $\Sigma_T$ , diferente de (5.28). Para definir la estructura que sigue esta matriz de covarianza, es necesario tener en cuenta la variación entre los tiempos. Además, se necesita obtener el tipo de relación entre las observaciones (matriz de covarianza muestral y/o matriz de correlación muestral). Algunas de estas matrices de covarianza y correlación se pueden encontrar en Diggle et al. (2002), Molenberghs & Verbeke (2005) y Davidian (2005).

Un resultado similar a (5.19) se puede derivar de las ecuaciones MLs para los parámetros en la distribución de  $f(\boldsymbol{\varepsilon}_{st} \mid \mathbf{y}_{st}, \boldsymbol{\theta})$ , donde  $\boldsymbol{\theta}$  denota los parámetros, de modo que

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\theta}} &= \int_{\mathbb{R}^{nT}} \frac{\partial \log f(\boldsymbol{\varepsilon}_{st} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\boldsymbol{\varepsilon}_{st} \mid \mathbf{y}_{st}, \boldsymbol{\theta}) d\boldsymbol{\varepsilon}_{st} \\ &= \text{E} \left[ \frac{\partial \log f(\boldsymbol{\varepsilon}_{st} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mid \mathbf{y}_{st} \right] \end{aligned} \quad (5.29)$$

Como  $f(\boldsymbol{\varepsilon}_{st} \mid \boldsymbol{\theta})$  sigue una función de distribución normal multivariada, algunas simplificaciones son posibles. Sin embargo, se prefiere utilizar la expresión (5.29) por su simplicidad.

### 5.3.2 Algoritmo de máxima verosimilitud vía Monte Carlo para DBGLSTARAR

La aplicación de métodos basados en verosimilitud para modelos de vectores autorregresivos generalizados espacio-tiempo no gaussianos basados en distancia, se ve obstaculizada por las dificultades computacionales que surgen de la gran dimensionalidad del vector aleatorio no observado  $\boldsymbol{\varepsilon}_{st}$  en el modelo presentado en (5.11). En esta subsección se considera el método MCMC para máxima verosimilitud (Geyer & Thompson 1992, Geyer 1994, Højbjerg 2003, Christensen 2004) en modelos de vector autorregresivos generalizados espacio-tiempo.

La integral en la ecuación (5.15) es también la constante normalizada en la función de densidad condicional de  $\boldsymbol{\varepsilon}_{st}$  dado  $\mathbf{y}_{st}$ ,

$$f(\boldsymbol{\varepsilon}_{st} \mid \mathbf{y}_{st}, \boldsymbol{\beta}^*, \boldsymbol{\theta}) \propto \prod_{i=1}^n \prod_{t=1}^T f\{y(\mathbf{s}_i, t); g^{-1}[\mathbf{x}(\mathbf{s}_i, t), \boldsymbol{\varepsilon}(\mathbf{s}_i, t); \boldsymbol{\beta}^*]\} f(\boldsymbol{\varepsilon}_{st} \mid \boldsymbol{\theta}) \quad (5.30)$$



Así, MCMC provee un método para la simulación de (5.30) y aproximación de (5.15).

La integral en la ecuación (5.15) tiene una alta dimensión, y consecuentemente, esta es intratable para encontrar las estimaciones de ML por maximización directa. Por consiguiente, la función de verosimilitud (5.15) se puede escribir como

$$\begin{aligned}
L(\boldsymbol{\beta}^*, \boldsymbol{\theta}) &= \int_{\mathbb{R}^{nT}} f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}) f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta}) d\boldsymbol{\varepsilon}_{st} \\
&= \int_{\mathbb{R}^{nT}} \frac{f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) f(\boldsymbol{\varepsilon}_{st} | \mathbf{X}; \boldsymbol{\theta})}{\tilde{f}(\mathbf{y}_{st}, \boldsymbol{\varepsilon}_{st})} \tilde{f}(\mathbf{y}_{st}, \boldsymbol{\varepsilon}_{st}) d\boldsymbol{\varepsilon}_{st} \\
&\propto \int_{\mathbb{R}^{nT}} \frac{f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta})}{f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}_0^*) \tilde{f}(\boldsymbol{\varepsilon}_{st})} \tilde{f}(\boldsymbol{\varepsilon}_{st} | \mathbf{y}_{st}) d\boldsymbol{\varepsilon}_{st} \\
&= \tilde{\mathbb{E}} \left[ \frac{f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*) f(\boldsymbol{\varepsilon}_{st} | \boldsymbol{\theta})}{f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}_0^*) \tilde{f}(\boldsymbol{\varepsilon}_{st})} \middle| \mathbf{y}_{st} \right] \tag{5.31}
\end{aligned}$$

donde  $\tilde{f}(\mathbf{y}_{st}, \boldsymbol{\varepsilon}_{st}) = f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}_0^*) \tilde{f}(\boldsymbol{\varepsilon}_{st})$  y  $\tilde{f}(\boldsymbol{\varepsilon}_{st})$  es alguna función de densidad con soporte en  $\mathbb{R}^{nT}$ , la función de densidad condicional  $\tilde{f}(\boldsymbol{\varepsilon}_{st} | \mathbf{y}_{st}) \propto f(\mathbf{y}_{st} | \boldsymbol{\varepsilon}_{st}) \tilde{f}(\boldsymbol{\varepsilon}_{st})$ , y  $\tilde{\mathbb{E}}(\cdot | \mathbf{y}_{st})$  denota la esperanza con respecto a  $\tilde{f}(\cdot | \mathbf{y}_{st})$  y depende de una estimación inicial de  $\boldsymbol{\beta}^*$ ,  $\boldsymbol{\beta}_0^*$ . Los MLEs se pueden calcular maximizando la aproximación (5.31) vía Monte Carlo,

$$L_r(\boldsymbol{\beta}^*, \boldsymbol{\theta}) = \frac{1}{r} \sum_{j=1}^r \frac{f(\mathbf{y}_{st}(j) | \mathbf{X}, \boldsymbol{\varepsilon}_{st}(j); \boldsymbol{\beta}^*) f(\boldsymbol{\varepsilon}_{st}(j) | \boldsymbol{\theta})}{f(\mathbf{y}_{st}(j) | \mathbf{X}, \boldsymbol{\varepsilon}_{st}(j); \boldsymbol{\beta}_0^*) \tilde{f}(\boldsymbol{\varepsilon}_{st}(j))} \tag{5.32}$$

donde los  $\boldsymbol{\varepsilon}_{st}(j)$ 's ( $j = 1, \dots, r$ ) son muestreados por MCMC de la función de distribución  $\tilde{f}(\cdot | \mathbf{y}_{st})$ . Como se ha señalado en (5.31) se debe elegir  $\tilde{f}(\cdot)$  cercano a  $f(\cdot | \hat{\boldsymbol{\theta}})$ , donde  $\hat{\boldsymbol{\theta}}$  es el MLE de  $\boldsymbol{\theta}$ , porque de lo contrario uno o muy pocos de los términos  $f(\boldsymbol{\varepsilon}_{st}(j) | \boldsymbol{\beta}^*, \boldsymbol{\theta}) / \tilde{f}(\boldsymbol{\varepsilon}_{st}(j))$ ,  $j = 1, \dots, r$  pueden dominar los otros en  $L_r(\boldsymbol{\beta}^*, \boldsymbol{\theta})$ , lo que hace que la aproximación sea menos útil.

A continuación se presenta un procedimiento numérico para maximizar la aproximación de Monte Carlo (5.32). Sea  $(\boldsymbol{\beta}^*, \boldsymbol{\theta}) = (\boldsymbol{\beta}^*, \boldsymbol{\psi}, \boldsymbol{\Sigma}_T)$ , la maximización de  $L_r$  con respecto a  $\boldsymbol{\beta}^*$  dados  $\boldsymbol{\psi}$  y  $\boldsymbol{\Sigma}_T$  es bastante sencilla, debido a que la primera y segunda derivadas de las densidades normales  $f(\boldsymbol{\eta}_{st}(j) | \mathbf{X}; \boldsymbol{\beta}^*, \boldsymbol{\psi}, \boldsymbol{\Sigma}_T)$ ,  $j = 1, \dots, r$ , con respecto a estos parámetros son simples, lo que hace factible y computacionalmente rápido un procedimiento iterativo como el de Newton-Raphson. Por ello, un procedimiento iterativo de valores iniciales adecuados son

$$\boldsymbol{\beta}^*(j) = [\mathbf{X}^t \text{Var}^{-1}(\boldsymbol{\varepsilon}_{st}) \mathbf{X}]^{-1} \mathbf{X}^t \text{Var}^{-1}(\boldsymbol{\varepsilon}_{st}) \boldsymbol{\eta}_{st}(j)$$

$j = 1, \dots, r$ , lo cual corresponde a los estimadores de máxima verosimilitud para las densidades normales  $f(\boldsymbol{\eta}_{st}(j) | \mathbf{X}; \boldsymbol{\beta}^*, \boldsymbol{\psi}, \boldsymbol{\Sigma}_T)$ .

Los valores de  $\beta^*$  que maximizan  $L_r(\beta^*, \theta)$  para un valor fijo de  $\vartheta$ ,  $\hat{\beta}^*(\psi, \Sigma_T)$  está conectado a  $L_r$ , y así, se obtiene  $\tilde{L}_r(\psi, \Sigma) = L_r(\hat{\beta}^*(\psi, \Sigma_T), \psi, \Sigma_T)$ . Esta función es maximizada con respecto a  $\psi$  y  $\Sigma_T$  para una función de correlación dada utilizando optimización numérica. Los parámetros  $\psi$  y  $\Sigma_T$  se ingresan en  $\tilde{L}_r$  via la matriz  $\text{Var}(\epsilon_{st})$  y ya que la inversa de esta matriz es computacionalmente exigente, la maximización podría ser relativamente lenta. La maximización puede también ser sensitiva a los valores iniciales en este proceso debido a que la aproximación  $L_r$  puede ser multimodal. Por lo tanto, el resultado debe ser investigado cuidadosamente, considerando una variedad de valores iniciales.

### 5.3.3 Ecuaciones de estimación generalizadas espacio-tiempo

La dificultad en la evaluación de la verosimilitud para un modelo tal como el presentado en (5.11) y el hecho que la maximización numérica no tan simple ha llevado a recurrir a las dos aproximaciones alternativas presentadas anteriormente y a investigar caminos computacionales efectivos para maximizar las verosimilitudes. La aproximación por GEEs comienza por plantear un GLM marginal para la media de  $\mathbf{y}_{st}$  en función de los predictores.

Liang & Zeger (1986) desarrollaron la aproximación GEE, la cual es una extensión de los GLMs. Cuando las respuestas se miden repetidamente a través del tiempo o en el espacio, lo que ocurre en el modelo DBGLSTARAR propuesto, se necesita estimar la correlación entre los tiempos en la misma ubicación y la correlación entre las regiones en un mismo tiempo. El método GEE toma en cuenta las correlaciones dentro de los clusters de unidades de muestreo por medio de una matriz de correlación parametrizada, mientras que las correlaciones entre los clusters se suponen iguales a cero. En un contexto espacio-tiempo, tales clusters pueden ser interpretados como regiones geográficas medidas a lo largo de tiempo, si las distancias entre las diferentes regiones medidas a lo largo del tiempo son lo suficientemente grandes (Albert & McShane 1995). En este capítulo se hace una pequeña modificación del enfoque de Liang & Zeger (1986) para utilizar estos modelos GEE en el espacio-tiempo. Hay que recordar que afortunadamente, las estimaciones de los parámetros de regresión son bastante robustos frente a errores de mala especificación de la matriz de correlación (Dobson 2002). La aproximación GEE es especialmente adecuada para la estimación de parámetros en vez de la predicción (Augustin et al. 2002), pero puede también ser utilizada para la predicción debido a que se está ajustando un modelo.

Primero, se presenta un estimador  $\hat{\beta}_I^*$  de  $\beta^*$  en el modelo (5.11) el cual surge bajo la hipótesis de que las observaciones a lo largo del tiempo en diferentes

lugares son independientes. Bajo la hipótesis de trabajo de independencia, como  $\mathbf{Y}_{st}$  es el vector de respuestas cuya distribución se puede escribir en la forma (5.2) con valor esperado  $\boldsymbol{\mu}_{st}$  enlazada al predictor lineal  $\boldsymbol{\eta}_{st} = \mathbf{X}\boldsymbol{\beta}^*$  por la expresión  $g(\boldsymbol{\mu}_{st}) = \boldsymbol{\eta}_{st}$ , entonces la función log-verosimilitud dada en (5.21) se puede escribir como

$$\begin{aligned} U_I &= \frac{\partial \log f(\mathbf{y}_{st} | \mathbf{X}, \boldsymbol{\varepsilon}_{st}; \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} = \frac{1}{\phi} \mathbf{X}^t \boldsymbol{\Xi} \boldsymbol{\Delta} (\mathbf{y}_{st} - \boldsymbol{\mu}_{st}) \\ &= \frac{1}{\phi} \mathbf{D}^t \mathbf{A}^{-1} (\mathbf{y}_{st} - \boldsymbol{\mu}_{st}) \end{aligned} \quad (5.33)$$

donde  $\mathbf{D} = \partial \boldsymbol{\mu}_{st} / \partial \boldsymbol{\beta}^*$  y  $\mathbf{A} = \text{diag}(v(\mu_{it}))$  es una matriz diagonal  $nT \times nT$  con el  $it$ -ésimo elemento dado por  $v(\mu_{it})$  para  $i = 1, \dots, n$  y  $t = 1, \dots, T$ . El estimador  $\hat{\boldsymbol{\beta}}_I$  se define como la solución de (5.33) utilizando el método de Fisher-Scoring.

Con el fin de ver la correlación entre las observaciones para el mismo individuo y entre los individuos, se incorpora la estructura de correlación mediante la selección de una matriz de correlación  $\mathbf{R}(\boldsymbol{\theta})$ , utilizando la siguiente expresión

$$\boldsymbol{\Gamma} = \mathbf{A}^{1/2} \mathbf{R}(\boldsymbol{\theta}) \mathbf{A}^{1/2}$$

donde  $\mathbf{R}(\boldsymbol{\theta}) = [\text{diag}(\text{Var}(\varepsilon_{it}))]^{-1/2} \text{Var}(\boldsymbol{\varepsilon}_{st}) [\text{diag}(\text{Var}(\varepsilon_{it}))]^{-1/2}$  con  $\text{Var}(\boldsymbol{\varepsilon}_{st}) = [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1} (\boldsymbol{\Sigma}_T \otimes \mathbf{I}_n) \{[\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1}\}^t$ ,  $\text{diag}(\text{Var}(\varepsilon_{it}))$  una matriz diagonal  $nT \times nT$  y  $\boldsymbol{\Sigma}_T = \sigma^2 \mathbf{R}(\boldsymbol{\vartheta})$ .

Entonces, el GEEs para el vector de parámetros,  $\boldsymbol{\beta}^*$ , toma la siguiente forma

$$\mathbf{U}(\boldsymbol{\beta}^*(\mathbf{R}(\boldsymbol{\theta})), \phi) = \frac{1}{\phi} \mathbf{D}^t \boldsymbol{\Gamma}^{-1} (\mathbf{y}_{st} - \boldsymbol{\mu}_{st}) \quad (5.34)$$

Liang & Zeger (1986) mostraron que el vector solución de la ecuación (5.34),  $\hat{\boldsymbol{\beta}}^*$ , sigue una función de distribución normal multivariada con media  $\boldsymbol{\beta}^*$  y matriz de varianzas y covarianzas dada por

$$\mathbf{H}_1^{-1}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi) = \phi \left( \widehat{\mathbf{D}}^t \widehat{\boldsymbol{\Gamma}}^{-1} \widehat{\mathbf{D}} \right)^{-1} \quad (5.35)$$

La consistencia de la estimación presentados en (5.35) depende de la correcta especificación de la función de enlace utilizada en (5.11). Para remediar este problema, se utiliza frecuentemente como matriz de varianzas y covarianzas de  $\hat{\boldsymbol{\beta}}^*$  la siguiente expresión

$$\begin{aligned} \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta}))) &= nT \left[ \mathbf{H}_1(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi) \right]^{-1} \mathbf{H}_2(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi) \\ &\quad \times \left[ \mathbf{H}_1(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi) \right]^{-1} \end{aligned} \quad (5.36)$$

donde

$$\mathbf{H}_2(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi) = \frac{1}{\phi} \left[ \widehat{\mathbf{D}}^t \widehat{\boldsymbol{\Gamma}}^{-1} (\mathbf{y}_{st} - \widehat{\boldsymbol{\mu}}_{st}) (\mathbf{y}_{st} - \widehat{\boldsymbol{\mu}}_{st})^t \widehat{\boldsymbol{\Gamma}}^{-1} \widehat{\mathbf{D}} \right]$$

Para resolver el sistema de ecuaciones dado en (5.34), se utiliza el método de Fisher-Scoring con la matriz  $H_1^{-1}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi)$  y el vector  $\mathbf{U}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi)$ . La  $m$ -ésima iteración del proceso está dado por

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{*(m+1)} &= \hat{\boldsymbol{\beta}}^{*(m)} + \left[ \mathbf{H}_1^{(m)}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi) \right]^{-1} \mathbf{U}^{(m)}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\boldsymbol{\theta})), \phi) \\ &= \hat{\boldsymbol{\beta}}^{*(m)} + \left( \widehat{\mathbf{D}}^t \widehat{\boldsymbol{\Gamma}}^{-1} \widehat{\mathbf{D}} \right)^{-1} \widehat{\mathbf{D}}^t \boldsymbol{\Gamma}^{-1} (\mathbf{y}_{st} - \widehat{\boldsymbol{\mu}}_{st}^{(m)}) \end{aligned} \quad (5.37)$$

o equivalentemente, se puede expresar como

$$\hat{\boldsymbol{\beta}}^{*(m+1)} = \left( \widehat{\mathbf{D}}^t \widehat{\boldsymbol{\Gamma}}^{-1} \widehat{\mathbf{D}} \right)^{-1} \widehat{\mathbf{D}}^t \boldsymbol{\Gamma}^{-1} \left[ \widehat{\mathbf{D}} \hat{\boldsymbol{\beta}}^{*(m)} + (\mathbf{y}_{st} - \widehat{\boldsymbol{\mu}}_{st}^{(m)}) \right] \quad (5.38)$$

y así, este proceso iterativo para calcular  $\boldsymbol{\beta}^*$  es equivalente a la realización de una regresión lineal iterativa ponderada de  $\left[ \widehat{\mathbf{D}} \hat{\boldsymbol{\beta}}^{*(m)} + (\mathbf{y}_{st} - \widehat{\boldsymbol{\mu}}_{st}^{(m)}) \right]$  en  $\widehat{\mathbf{D}}$  con matriz de pesos  $\boldsymbol{\Gamma}^{-1}$ .

Una vez estimados los parámetros fijos en el modelo (5.11),  $\boldsymbol{\beta}^*$ , se está preparado para estimar los parámetros  $\boldsymbol{\theta}$ . Los parámetros de efectos aleatorios se estiman utilizando el mismo procedimiento presentado en la Subsección 5.3.1 para los efectos aleatorios. En la Sección 5.3.4, se presentan algunas estructuras que puede asumir  $\mathbf{R}(\boldsymbol{\vartheta})$ , donde  $\boldsymbol{\vartheta}$  está involucrado en  $\boldsymbol{\theta}$ . Además, el parámetro de escala,  $\phi$  puede ser estimado por la siguiente expresión

$$\hat{\phi} = \frac{1}{nT - p^*} \sum_{i=1}^n \sum_{t=1}^T r_{it}^2 \quad (5.39)$$

donde

$$r_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{v(\hat{\mu}_{it})}}$$

#### 5.3.4 Elecciones específicas de $\mathbf{R}(\boldsymbol{\vartheta})$

En esta subsección se discuten algunas elecciones específicas de  $\mathbf{R}(\boldsymbol{\vartheta})$  presentadas en Liang & Zeger (1986). El número de parámetros de perturbación y el estimador de  $\boldsymbol{\vartheta}$  varía de caso a caso.

**Ejemplo 5.3. Independencia entre tiempos.** Cuando  $\mathbf{R}(\boldsymbol{\vartheta}) = \mathbf{R}_0 = \mathbf{I}_T$ , la matriz identidad, se obtiene la ecuación de estimación de independencia a lo largo del tiempo pero no en el espacio. Observe que para cualquier especificación  $\mathbf{R}_0$ , no requiere ningún conocimiento sobre  $\phi$  en la estimación de  $\boldsymbol{\beta}^*$  y  $\text{Var}(\boldsymbol{\beta}^*)$ .

**Ejemplo 5.4. l-dependencia.** Sea  $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_{T-1})^t$  donde  $\vartheta_t = \text{Corr}(y_{it}, y_{i(t+1)})$  para  $t = 1, \dots, T-1$ . Un estimador natural de  $\vartheta_t$  dados  $\boldsymbol{\beta}^*$  y  $\phi$  es

$$\hat{\vartheta}_t = \frac{1}{\phi(n-p^*)} \sum_{i=1}^n \hat{r}_{it} \hat{r}_{i(t+1)}, \quad t = 1, \dots, T-1$$

Como un caso especial, sea  $s = 1$  y  $\vartheta_t = \vartheta$  ( $t = 1, \dots, T-1$ ). Entonces el  $\vartheta$  común se puede estimar por

$$\hat{\vartheta} = \frac{1}{n-1} \sum_{t=1}^{T-1} \hat{\vartheta}_t$$

Una extensión a l-dependencia es directa.

**Ejemplo 5.5. Simétrica compuesta.** Sea  $s = 1$  y suponga que  $\text{Corr}(y_{it}, y_{it'}) = \vartheta$  para todo  $t \neq t'$ . Esta es la estructura de correlación intercambiable obtenida a partir de un modelo de efectos aleatorios con un nivel aleatorio para cada sujeto (Laird & Ware 1982). Dado  $\phi$ ,  $\vartheta$  se puede estimar por

$$\hat{\vartheta} = \frac{1}{\phi \left[ \frac{1}{2}T(T-1)n - p^* \right]} \sum_{i=1}^n \sum_{t > t'} \hat{r}_{it} \hat{r}_{it'}$$

Note que es posible un número arbitrario de observaciones y observaciones en el tiempo para cada sujeto con este supuesto.

**Ejemplo 5.6. Autorregresivo de primer orden.** Sea  $\text{Corr}(y_{it}, y_{it'}) = \vartheta^{|t-t'|}$ , ya que bajo esta estructura,  $E(\hat{r}_{it} \hat{r}_{it'}) \approx \vartheta^{|t-t'|}$ , se puede estimar  $\vartheta$  mediante la pendiente de la regresión de  $\log(\hat{r}_{it} \hat{r}_{it'})$  sobre  $\log|t-t'|$ . Observe que un número arbitrario y espaciado de observaciones se pueden acomodar con esta estructura de trabajo.

**Ejemplo 5.7. No estructurada.** Sea  $\mathbf{R}(\boldsymbol{\vartheta})$  totalmente no especificada, es decir se tiene  $s = \frac{1}{2}T(T-1)$  parámetros desconocidos. En este caso,  $\mathbf{R}(\boldsymbol{\vartheta})$  se puede estimar mediante

$$\frac{1}{n\phi} \sum_{i=1}^n \mathbf{A}_i^{-1/2} (\boldsymbol{\mu}_i - \mathbf{y}_i) (\boldsymbol{\mu}_i - \mathbf{y}_i)^t \mathbf{A}_i^{-1/2}$$

donde  $\mathbf{A}_i = \text{diag}(v(\bar{\boldsymbol{\mu}}_i))$  es una matriz diagonal  $T \times T$  con  $\bar{\boldsymbol{\mu}}_i = \frac{1}{T} \sum_{t=1}^T \boldsymbol{\mu}_{it}$ ,  $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{iT})^t$  y  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})^t$ .

Como las GEEs dependen de los parámetros  $\boldsymbol{\beta}^*$  y  $\boldsymbol{\theta}$ , se requieren algunos pasos en la estimación del proceso:

1. Obtenga una estimación inicial de  $\boldsymbol{\beta}^*$ ,  $\hat{\boldsymbol{\beta}}^{*(0)}$ , mediante GLM asumiendo independencia entre las observaciones y los tiempos, o de forma equivalente, usando la ecuación (5.38) con  $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{I}_{nT}$  y  $\phi = 1$ .

2. Utilice  $\hat{\beta}^{*(0)}$  obtenido en la etapa anterior para calcular  $\hat{\epsilon}_{st}$  empleando la ecuación (5.11).
3. Utilice la ecuación (5.25) para encontrar el vector de parámetros  $\psi$  y la estructura de correlación apropiada de  $\Sigma_T$  (ver ejemplos 5.3 a 5.7). En este paso se pueden utilizar las ecuaciones (5.26) y (5.27).
4. Estime el parámetro de escala usando la ecuación (5.39).
5. Obtenga  $\mathbf{R}(\theta)$  y estime  $\beta^*$  utilizando la ecuación (5.38).
6. Repita los pasos 3 a 5 hasta lograr la convergencia deseada en los parámetros  $\beta^*$  y  $\theta$ .

## 5.4 Selección, validación y predicción del modelo ajustado utilizando GEE para espacio-tiempo

Cuando se utiliza el método GEE en espacio-tiempo, se debe tener en cuenta que por construcción GEE es un método que no se basa en el uso de la función de verosimilitud. Esto genera que muchas herramientas diseñadas para la construcción de modelos en el campo de verosimilitud no se puedan utilizar en el contexto de GEE. Por lo tanto, el bien conocido criterio de información Akaike (AIC) no se puede aplicar directamente ya que el AIC se basa en la estimación de máxima verosimilitud, mientras GEE no se basa en verosimilitud. Pan (2001) presenta un enfoque que es una modificación al AIC, donde la verosimilitud se sustituye por la cuasi-verosimilitud y se hace un ajuste apropiado para el término de penalización. Esta modificación está dada por

$$\text{QSTIC}(\mathbf{R}(\hat{\theta}), \phi) \equiv -2Q(\hat{\beta}^*(\mathbf{R}(\hat{\theta})), \hat{\phi}, \mathbf{I}_{nT}) + 2\text{trace}(\hat{\Omega}_T \hat{\Gamma}) \quad (5.40)$$

donde

$$Q(\hat{\beta}^*(\mathbf{R}(\hat{\theta})), \hat{\phi}, \mathbf{R}(\hat{\theta}) = \mathbf{I}_{nT}) = \sum_{i=1}^n \sum_{t=1}^T \int_{y_{it}}^{\mu_{it}} \frac{y_{it} - \tau}{\phi v(\tau)} d\tau, \quad (5.41)$$

$$\hat{\Omega}_T = -\partial^2 Q(\beta^*(\mathbf{R}(\hat{\theta})), \hat{\phi}; \mathbf{I}_{nT}) / \partial \beta^* \partial \beta^{*t} |_{\beta^* = \hat{\beta}^*},$$

y  $\hat{\Gamma}$  fue presentada anteriormente. Esto es la cuasi-verosimilitud espacio-tiempo bajo el criterio de modelo de independencia (quasi-likelihood space-time under the independence model criterion, QSTIC) para GEE. Este criterio también se puede utilizar para seleccionar la mejor estructura de la matriz  $\mathbf{R}(\theta)$  de

acuerdo a los datos, o para seleccionar variables que se deben considerar en el predictor lineal del modelo.

Otro punto de vista de vital importancia en el proceso de inferencia son las pruebas de hipótesis sobre el modelo propuesto, que se pueden realizar de acuerdo a la hipótesis lineal general dada por

$$H_0 : \mathbf{L}\boldsymbol{\beta}^* = \mathbf{0} \quad \text{vs} \quad H_0 : \mathbf{L}\boldsymbol{\beta}^* \neq \mathbf{0}$$

donde  $\mathbf{L}$  es una matriz  $l \times p^*$  que es conocida. Uno de los criterio más utilizados e implementados en diferentes paquetes de análisis de datos es el estadístico de Wald, este enfoque se puede usar para la selección de variables explicativas en el predictor lineal. En este caso, el estadístico de Wald está dado

$$W = (\mathbf{L}\hat{\boldsymbol{\beta}}^*)^t \left[ \mathbf{L}\widehat{\text{Var}}\left(\hat{\boldsymbol{\beta}}^* \left(\mathbf{R}(\hat{\boldsymbol{\theta}})\right)\right) \mathbf{L}^t \right]^{-1} (\mathbf{L}\hat{\boldsymbol{\beta}}^*) \quad (5.42)$$

donde  $\widehat{\text{Var}}\left(\hat{\boldsymbol{\beta}}^* \left(\mathbf{R}(\hat{\boldsymbol{\theta}})\right)\right)$  fue dado en (5.36). Este estadístico tiene una distribución asintótica  $\chi^2$  con  $\text{rank}(\mathbf{L}) = l$  grados de libertad.

Después de ajustar el DBGLSTARAR, es importante llevar a cabo un análisis de diagnóstico para verificar la bondad de ajuste del modelo estimado.

### 5.4.1 Medida de bondad de ajuste

Una medida global de la variación explicada se obtiene mediante el cálculo del pseudo  $R_k^2$  definido como

$$R_k^2 = r^2(\hat{\boldsymbol{\eta}}_{st}, g(\mathbf{y}_{st})) \quad 0 \leq R_k^2 \leq 1 \quad (5.43)$$

donde  $r(\hat{\boldsymbol{\eta}}_{st}, g(\mathbf{y}_{st}))$  es el coeficiente de correlación muestral entre  $\hat{\boldsymbol{\eta}}_{st}$  y  $g(\mathbf{y}_{st})$ . Cuando  $R_k^2 = 1$ , hay una perfecta concordancia entre  $\hat{\boldsymbol{\eta}}_{st}$  y  $g(\mathbf{y}_{st})$ , y por lo tanto, entre  $\hat{\boldsymbol{\mu}}_{st}$  y  $\mathbf{y}_{st}$ .

### 5.4.2 Análisis residual

Como es bien conocido, el objetivo del análisis residual es identificar observaciones atípicas y/o mala especificación del modelo. Ésta se puede basar en los residuos ordinarios o en la deviance de los residuales. De esta manera, los residuos son medidas de concordancia entre los datos y el modelo ajustado. La mayoría de los residuos están basados en las diferencias entre las respuestas observadas y la media condicional ajustada, por ejemplo, los residuales de Pearson, los cuales se pueden expresar como

$$r_{P_{it}} = r_P(\mathbf{s}_i, t) = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{\phi v(\hat{\mu}_{it})}} \quad (5.44)$$

donde  $v(\hat{\mu}_{it})$  es una función varianza.

Por otra parte, basados en la función de cuasi-verosimilitud presentada en (5.41), se puede reescribir ésta como

$$Q(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\hat{\boldsymbol{\theta}})), \hat{\phi}, \mathbf{I}_{nT}) = \sum_{i=1}^n \sum_{t=1}^T Q_{it}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\hat{\boldsymbol{\theta}})), \hat{\phi}, \mathbf{I}_{nT}) \quad (5.45)$$

donde  $Q_{it}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\hat{\boldsymbol{\theta}})), \hat{\phi}, \mathbf{I}_{nT}) = \int_{y_{it}}^{\mu_{it}} \frac{y_{it} - \tau}{\phi v(\tau)} d\tau$ . Análogamente al GLM, se puede construir una función de deviance basada en la función cuasi-verosimilitud. Para una sola observación esta función se define como

$$r_{D_{it}} = r_D(\mathbf{s}_i, t) = -2\phi Q_{it}(\hat{\boldsymbol{\beta}}^*(\mathbf{R}(\hat{\boldsymbol{\theta}})), \hat{\phi}, \mathbf{I}_{nT}) = 2 \int_{\mu_{it}}^{y_{it}} \frac{y_{it} - \tau}{v(\tau)} d\tau \quad (5.46)$$

que es utilizada como una medida de discrepancia entre el valor ajustado,  $\mu_{it}$ , y el valor observado,  $y_{it}$ .

### 5.4.3 Selección de las coordenadas principales para el modelo DBGLSTARAR reducido

Para evitar el problema de obtener un pseudo  $R_k^2 \simeq 1$  cuando el rango de  $\mathbf{X}$  es  $k = nT - 1$ , es necesario considerar únicamente los vectores propios más correlacionados de  $\mathbf{B}$ , dados por  $\mathbf{1}, \mathbf{X}_1, \dots, \mathbf{X}_k$ , con la variable respuesta espacio-tiempo  $\mathbf{y}_{st}$ , es decir, las coordenadas principales significativamente más correlacionadas con  $\mathbf{y}_{st}$ .

Inicialmente, se pueden seleccionar solo  $k$  coordenadas principales. Una buena aproximación para seleccionar las columnas de  $\mathbf{X}$  consiste en clasificarlos en función de su pseudo coeficiente de correlación con respecto a  $\mathbf{y}_{st}$ , es decir,

$$R_1^2(\mathbf{X}_1) > \dots > R_1^2(\mathbf{X}_k) > \dots > R_1^2(\mathbf{X}_{nT-1})$$

donde  $R_1^2(\mathbf{X}_j)$  es el coeficiente de pseudo correlación entre la  $\mathbf{X}_j$ -ésima coordenada principal ( $j = 1, \dots, k, \dots, nT - 1$ ) y  $\mathbf{y}_{st}$ . Esto se hace dejando la misma en función de enlace  $g$  en los diferentes modelos ajustados. Con este procedimiento, las coordenadas principales menos correlacionadas con  $\mathbf{y}_{st}$  se eliminan en la matriz de coordenadas principales  $\mathbf{X}$ , es decir, no se consideran en el modelo final  $nT - k - 1$  coordenadas principales.

Otra opción para la selección de coordenadas principales se lleva a cabo de una manera similar a la selección del número de variables en la regresión multivariada, utilizando el estadístico llamado  $C_p - Mallows$ . Esto es, se hace una gráfica que represente los puntos  $(j, 1 - c(j))$   $j = 1, \dots, k, k + 1, \dots, nT - 1$ ,



y luego, se determinan los puntos con un descenso significativo en la falta de predictibilidad, dada por  $1 - c(j)$ . La predictibilidad  $c(j)$  está dada por (ver Cuadras et al. (1996) para más detalles)

$$c(0) = 0, \quad c(j) = \frac{\sum_{l=1}^j R_1^2(\mathbf{X}_l) \lambda_l}{\sum_{l=1}^{n-1} R_1^2(\mathbf{X}_l) \lambda_l}, \quad j = 1, \dots, k, \dots, nT - 1$$

donde  $\lambda_l$  es el  $l$ -ésimo valor propio asociado a  $\mathbf{X}_l$ ,  $l = 1, \dots, k, \dots, nT - 1$ . Por tanto, se deben eliminar las coordenadas principales  $\mathbf{X}_{k+1}, \dots, \mathbf{X}_{nT-1}$ .

Además, una coordenada principal  $\mathbf{X}_j$  se puede eliminar si la hipótesis nula  $\beta_j = 0$  no se rechaza. Una prueba estadística se puede basar en el estadístico de Wald, el cual está dado por

$$W_j = \hat{\beta}_j^2 \left[ \widehat{\text{Var}} \left( \hat{\beta}_j \left( \mathbf{R}(\hat{\boldsymbol{\theta}}) \right) \right) \right]^{-1}, \quad j = 1, \dots, nT - 1$$

o equivalentemente

$$w_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{\text{Var}} \left( \hat{\beta}_j \left( \mathbf{R}(\hat{\boldsymbol{\theta}}) \right) \right)}}, \quad j = 1, \dots, nT - 1 \quad (5.47)$$

donde  $\hat{\beta}_i$  es la  $i$ -ésima componente de  $\hat{\boldsymbol{\beta}}$ .  $w_j$  es distribuida asintóticamente normal con media cero y varianza 1. Finalmente, ordenando los  $w_j$ 's de mayor a menor valor, se obtienen las  $k$  coordenadas principales más significativas. Este último método fue el utilizado en la aplicación.

Una vez elegidas las  $k$  coordenadas principales, se esta preparado para discutir las técnicas espacio-tiempo para predecir el valor de un campo aleatorio en una ubicación temporal dada de observaciones cercanas a los datos observados.

#### 5.4.4 Predicción espacio-tiempo de un nuevo individuo

Las coordenadas  $\mathbf{x}_{(k)}(\mathbf{s}_0)$  se obtienen asumiendo que se tienen las observaciones de las variables explicativas mixtas para un nuevo individuo, esto es, se conocen  $\mathbf{v}_{10}^t = \mathbf{v}_1^t(\mathbf{s}_0) = (1, v_{01}, \dots, v_{0p_1})$  y  $\mathbf{v}_{20t}^t = \mathbf{v}_2^t(\mathbf{s}_0, t) = (v_{0t1}, \dots, v_{0tp_{2t}})$ . Entonces, existe una configuración de puntos  $\mathbf{v}_0 = (\mathbf{v}_1^t(\mathbf{s}_0), \mathbf{v}_2^t(\mathbf{s}_0, 1), \dots, \mathbf{v}_2^t(\mathbf{s}_0, T)) = (1, v_{01}, \dots, v_{0p_1}, v_{011}, \dots, v_{01p_{2_1}}, \dots, v_{0T1}, \dots, v_{0Tp_{2_T}})^t \in \mathbb{R}^{p^*}$ . Así, se deben calcular las distancias entre el nuevo individuo y cada uno de los individuos involucrados en el modelo propuesto en (5.1), es decir,  $d_{0i} = d(\mathbf{v}_0, \mathbf{v}_i)$ ,  $i = 1, \dots, nT$ . A partir de estas distancias, una predicción puede realizarse utilizando un resultado propuesto por Gower (1968) y Cuadras y Arenas (Cuadras

& Arenas 1990, Section 3.3), que se relaciona el vector  $\mathbf{d}_0 = (d_{01}^2, \dots, d_{0(nT)}^2)^t$  de distancias al cuadrado y el vector  $\mathbf{x}_{0(k)} = (x_{01}, \dots, x_{0k})^t$  de coordenadas principales asociadas al nuevo individuo como sigue

$$d_{0i}^2 = (\mathbf{x}_{0(k)} - \mathbf{x}_{i(k)})^t (\mathbf{x}_{0(k)} - \mathbf{x}_{i(k)}) \quad (5.48)$$

donde  $\mathbf{x}_{i(k)} = (x_{i1}, \dots, x_{ik})^t$  con  $i = 1, \dots, nT$ . Entonces, se tiene que

$$\mathbf{x}_{0(k)} = \frac{1}{2} \Lambda_{(k)}^{-1} \mathbf{X}_{st}^t (\mathbf{b} - \mathbf{d}_0) \quad (5.49)$$

donde  $\Lambda_{(k)}$  es una matriz diagonal con  $k$  valores propios asociados a los  $k$  vectores propios  $\mathbf{X}_{st}$  seleccionados anteriormente,  $\mathbf{b} = (b_{11}, \dots, b_{(nT)(nT)})^t$  es un vector conformado por los elementos de la diagonal de  $\mathbf{B}$ , con  $b_{ii} = \mathbf{x}_{i(k)}^t \mathbf{x}_{i(k)}$ ,  $i = 1, \dots, nT$ .

### Predicción espacio-tiempo

Específicamente, se trata de interpolar el valor  $\mathbf{y}_0 = (y(\mathbf{s}_{n+1}, t_1), \dots, y(\mathbf{s}_{n+n'}, t_{n'}))^t$  de un campo aleatorio  $\mathbf{Y}_0$  a partir de las observaciones  $y_{it} = y(\mathbf{s}_i, t)$ ,  $i = 1, \dots, n$  y  $t = 1, \dots, T$  en  $l$  puntos espacio-tiempo predefinidos donde  $1 \leq t_l \leq T$  con  $l = 1, \dots, n'$ . Así, el interés se centra en la interpolación de efectos aleatorios sobre un espacio-tiempo cuando las observaciones no son gaussianas.

Por consiguiente, sea  $\boldsymbol{\eta}_{st}^0 = (\eta(\mathbf{s}_{n+1}, t_1), \dots, \eta(\mathbf{s}_{n+n'}, t_{n'}))^t$  la predicción funcional y sea  $f(\boldsymbol{\eta}_{st}^0, \boldsymbol{\eta}_{st})$  la función de densidad conjunta  $\boldsymbol{\eta}_{st}$  y el vector  $\boldsymbol{\eta}_{st}^0$ . Si se limita el interés a pseudo predictores lineales insesgados de la forma

$$\tilde{\boldsymbol{\eta}}_{st} = \mathbf{p} + \mathbf{Q}\boldsymbol{\eta}_{st} \quad (5.50)$$

para algún vector conforme a  $\mathbf{p}$  y una matriz  $\mathbf{Q}$  (McCulloch et al. 2008). Por lo tanto, minimizando el error cuadrático medio de la predicción, se encuentra que el mejor pseudo predictor lineal insesgado esta dado por

$$\tilde{\boldsymbol{\eta}}_{st} = \mathbf{X}^0 \boldsymbol{\beta}^* + \text{Cov}^t(\boldsymbol{\eta}_{st}, \boldsymbol{\eta}_{st}^0) \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{st}}^{-1} [\boldsymbol{\eta}_{st} - \mathbf{X}\boldsymbol{\beta}^*]$$

donde  $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{st}} = [\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1} (\boldsymbol{\Sigma}_T \otimes \mathbf{I}_n) \{[\mathbf{I}_{nT} - (\boldsymbol{\Psi} \otimes \mathbf{W}_2)]^{-1}\}^t$ ,  $\mathbf{X}^0$  es una matriz de  $k$  de coordenadas principales para  $n'$  nuevos elementos espacio-tiempo incluyendo un vector de 1's, es decir,  $\mathbf{1}_{n'}$  de tamaño  $n' \times 1$ .

La matriz de covarianzas para la predicción tiene la siguiente forma general

$$\text{Var}(\tilde{\boldsymbol{\eta}}_{st} | \mathbf{y}_{st}) \approx \boldsymbol{\Sigma}_0 + \text{Cov}^t(\boldsymbol{\eta}_{st}, \boldsymbol{\eta}_{st}^0) \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{st}}^{-1} \text{Cov}(\boldsymbol{\eta}_{st}, \boldsymbol{\eta}_{st}^0)$$

donde  $\boldsymbol{\Sigma}_0 = \text{Var}(\boldsymbol{\eta}_{st}^0) - \text{Cov}^t(\boldsymbol{\eta}_{st}, \boldsymbol{\eta}_{st}^0) \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{st}}^{-1} \text{Cov}(\boldsymbol{\eta}_{st}, \boldsymbol{\eta}_{st}^0)$ .

## 5.5 Aplicación

En este estudio se utilizan variables sociales, económicas, geográficas y de presencia del Estado para los 32 departamentos colombianos, con la finalidad de ver la relación entre el número de acciones armadas (AA) estandarizadas por cada 1000  $km^2$  que han realizado los grupos irregulares o guerrillas de las FARC-EP y el ELN durante los años 2003 a 2009, con las variables exógenas: i) tasa de atención a víctimas de la violencia (AVV) estandarizadas por cada 1000  $km^2$  (en términos de 40 salarios mínimos mensuales legales vigentes “SMMLV”), esta variable esta asociada al número de hogares que sufren perjuicios en su vida, o grave deterioro en su integridad personal o en sus bienes, por razón de atentados terroristas, combates, secuestros, ataques y masacres, entre otros, ii) desplazamiento forzado - hogares expulsados (DFHE), es el número de hogares expulsados por cada 1000  $km^2$ , iii) desplazamiento forzado - hogares recibidos (DFHR), es el número de hogares recibidos por cada 1000  $km^2$ , iv) total de confrontaciones armadas por año (CAA) por cada 1000  $km^2$ , v) número de AA por parte de las fuerzas militares (FM) por cada 1000  $km^2$  y vi) porcentaje de personas que residen el área urbana (URB) con respecto a la población total del departamento.

Toda la anterior información es tomada en 1000  $km^2$  de área oficial; esta área se considera con respecto a las zonas donde operan estos grupos y no corresponde al área del departamento. Además, en todas las variables exógenas se cuenta con información desde el 2003 a 2009, es decir que cada una de estas variables va ir por año. Las fuentes de información son el Departamento Administrativo para la Prosperidad Social (variables AVV, DFHE y DFHR), Vicepresidencia de la República (variables AA, CAA y FM) y Departamento Administrativo Nacional de Estadística (variable URB), todas entidades gubernamentales de Colombia. Las páginas web donde se puede encontrar dicha información son: <http://sigotn.igac.gov.co/sigotn/> y <http://sigotn.igac.gov.co/sigotn/default.aspx>.

Por las características y estructura de la información, se tiene un conjuntos de datos panel y por ello, se proponen los modelos GLSTARAR y DBGLSTARAR para ajustar el número de AA por departamento a lo largo del período de estudio. Con el primer modelo se pueden hallar las posibles causas de la presencia y expansión de estos actores armados para los años 2003-2009, y con el segundo modelo, se esperan mejorar las predicciones en comparación al primero, por las ventajas mostradas en los capítulos anteriores del método DB. En este punto, hay que destacar la gran utilidad del segundo modelo sobre el primero, ya que en el modelo GLSTARAR puede haber multicolinealidad y además sobre-parametrización por el exceso de variables explicativas en estudio, mientras que en el modelo DBGLSTARAR lo anterior no sucede porque se trabajan con coordenadas principales obtenidas a partir de la descomposición

espectral realizadas a las distancias entre individuos.

La finalidad de hacer este estudio tiene como propósito llamar la atención a las entidades colombianas gubernamentales para que tomen las acciones sociales que correspondan en las leyes de reparación a víctimas y solución del conflicto. Se estiman diversos modelos econométricos, que buscan hallar las variables explicativas y las coordenadas principales significativas como causantes de un mayor o menor número de AA de las FARC y del ELN en los distintos departamentos de Colombia. Los modelos que se desarrollan son de tipo espacio-temporal al introducir variables que capturan los efectos generados por la autocorrelación espacio-tiempo y la heterogeneidad espacio-tiempo.

Las FARC-EP y el ELN son grupos guerrilleros que operan en Colombia y en las regiones fronterizas de Brasil, Ecuador, Panamá, Perú y Venezuela. Son partícipes del conflicto armado colombiano desde su conformación. Sus acciones consisten en narcotráfico, guerra de guerrillas, así como técnicas terroristas como la implantación de minas antipersona, el asesinato de civiles, miembros del gobierno, policías y militares, el secuestro con fines políticos o extorsivos, atentados con bombas y armas no convencionales (cilindros de gas, animales bomba) y actos que han provocado desplazamientos forzados de civiles (Dudley 2004).

Pese a los 11128 millones de dólares que invirtió Colombia en gasto militar en 2009 lo ubicaron como el país de América Latina que destinaba mayor porcentaje de Producto Interno Bruto (PIB) a este rubro en Latinoamérica (3.9 puntos del PIB), la cifra está muy por debajo de Brasil, que en el 2009 destinó 34334 millones (Sipri 2012). Aunque para el año 2012 el gasto militar se mantuvo (11446 millones de dólares, 3.9 puntos del PIB), se sigue dedicando una buena parte del PIB a combatir los grupos guerrilleros en Colombia.

### 5.5.1 Análisis descriptivo

Dejando de lado la contextualización del problema presentada anteriormente, en primer lugar se hace un análisis descriptivo de la información con la finalidad de ver el comportamiento de los datos y determinar las posibles matrices de pesos  $\mathbf{W}_1$  y  $\mathbf{W}_2$  que intervienen en el modelo DBGLSTARAR. En la Figura 5.1 se presenta la división político y administrativa departamental de Colombia, en éste se presenta a las islas de San Andrés y Providencia (parte superior izquierda); sin embargo, este departamento no se considerará en el análisis ya que el número de presencia de acciones armadas fue nula durante todo el período de estudio. Además, por lo alejado de la costa terrestre en el Atlántico, no es conveniente analizarlo ya que se puede generar ruido.

La Figura 5.2 muestra los quintiles del número de AA por departamento, estos tipos de mapas son útiles para analizar agregaciones espaciales y



FIGURA 5.1: División político administrativa departamental de Colombia.  
 Fuente de información: Cartografía base IGAC. DEM 90 m SRTM-NASA. Escala de elaboración 1:500000. Escala de presentación 1:7000000.

espacio-temporales. Así regiones con valores similares en el número de AA corresponderán en general a un mismo quintil, simultáneamente se puede ver el comportamiento a través del tiempo. Por consiguiente, los colores oscuros indican valores altos para el número de AA y los colores claros valores bajos. En general, se encuentran valores bajos en la región de la amazonia, es decir los departamentos del Putumayo, Caquetá, Amazonas, Vaupés y Guainía, ubicados en la parte inferior derecha. A la vez, se encuentran valores altos en la cordillera de los Andes, específicamente para el 2009 en la cordillera oriental. En la costa Atlántica se encuentran valores altos para los departamentos de Sucre y Bolívar en los años 2003 y 2004. Con respecto a la región Pacífica, hay valores altos en el período 2006 a 2009 para los departamentos de Nariño, Cauca y Valle del Cauca, unidos a los departamentos del Huila, Tolima y Cundinamarca en la región Andina. A través del tiempo es clara una disminución del número de AA en la costa Atlántica y simultáneamente un incremento en la cordillera oriental.

Por otro lado, las estructuras de dependencia espacial pueden ser generadas a partir de las relaciones espaciales entre los departamentos y estas relaciones permiten generar la matriz de pesos espacial. Son tres las formas más comunes de establecer estas estructuras; la primera se denomina efecto reina, que corresponde a la primera fila de mapas en la Figura 5.3, encontrando en la primera columna de izquierda a derecha la contigüidad de orden 1, la cual se construye a partir de los centroides de cada polígono (o centro de cada departamento), considerando vecino a aquel polígono con el que comparte un borde. Las contigüidades de orden 2 y 3 (columnas segunda y tercera de izquierda a derecha); la contigüidad de orden 2 se refiere a los vecinos de los vecinos de cada uno de los polígonos, y la contigüidad de orden 3, se refiere a los vecinos de los vecinos de los vecinos de cada polígono. Lo aconsejable es considerar sólo el primer orden ya que ordenes superiores son un poco difíciles de interpretar y es posible que se pueda asociar la contigüidad de orden 1 con las vías terrestres que conectan a los diferentes departamentos.

La segunda estructura es a partir de las distancias entre los diferentes polígonos, se encuentran los  $k$ -vecinos más cercanos a cada uno de los polígonos, en la Figura 5.3 corresponde a la segunda fila; se han considerado tres casos: 1, 2 y 4 vecinos más cercanos, dada la topografía colombiana es muy difícil que todas las regiones se conecten, debido a que no hay suficientes vías terrestres construidas las conexiones entre una región y otra se dan más por las vías aéreas. Por último, la tercera fila de mapas de la Figura 5.3 es construida a partir de una distancia crítica, denominada umbral, la cual es creada con las distancias espaciales generadas a partir de los centroides de cada polígono. De esta manera, se establecieron los vecinos que se encuentran a una distancia menor del umbral con respecto a cada uno de los polígonos; las distancias consideradas en la Figura 5.3 son 100, 200 y 300 *km*.

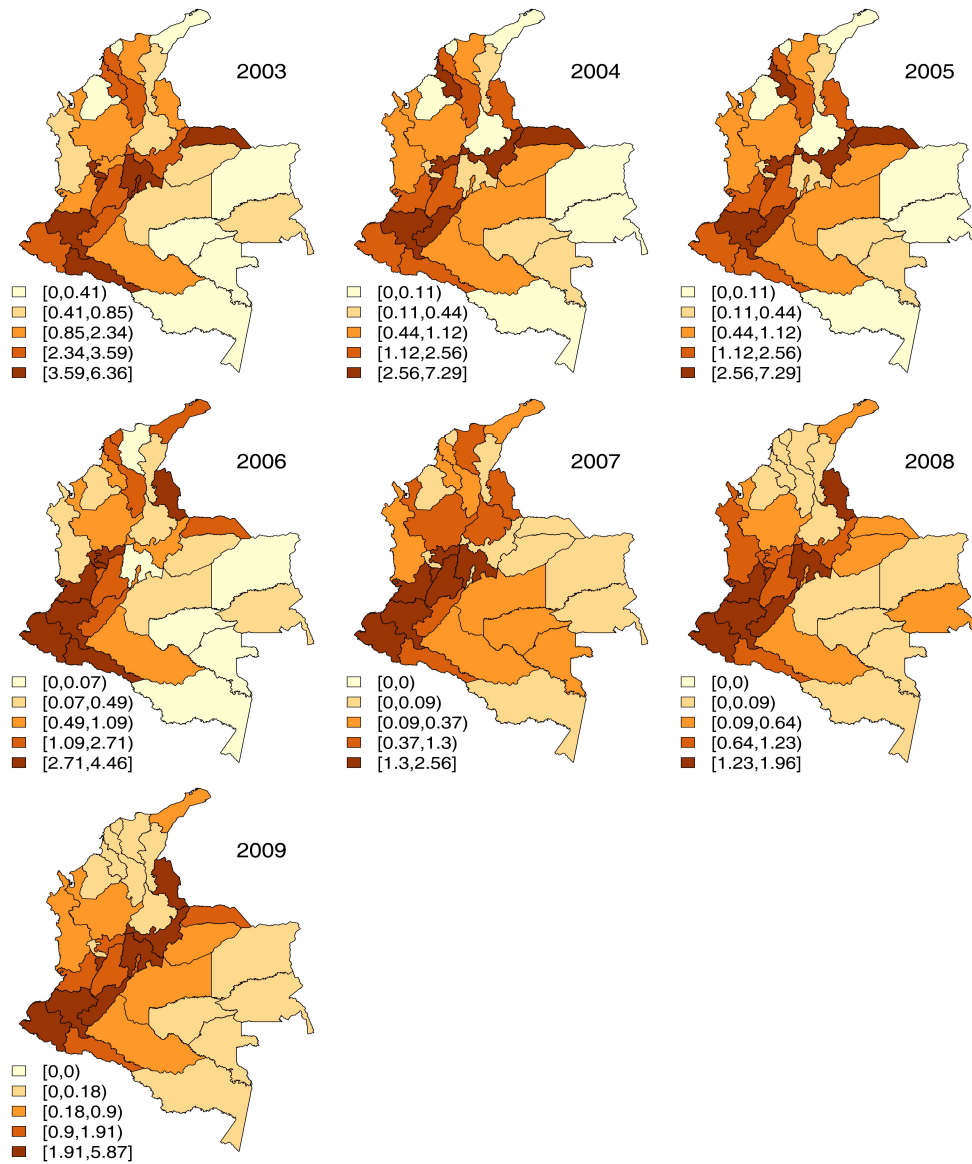


FIGURA 5.2: Mapas de los quintiles del número de acciones armadas por departamento para el período 2003 a 2009

En la Figura 5.4 se muestran los dispersogramas del  $I$  de Moran para el número de AA, los cuales se obtienen a partir de la siguiente ecuación

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} [y(\mathbf{s}_i, t) - \bar{y}_{\bullet t}] [y(\mathbf{s}_j, t) - \bar{y}_{\bullet t}]}{S_0 \sum_{i=1}^n [y(\mathbf{s}_i, t) - \bar{y}_{\bullet t}]^2}, \quad t = 1, \dots, T$$

donde  $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$  y  $\bar{y}_{\bullet t} = \sum_{i=1}^n y(\mathbf{s}_i, t)$ .

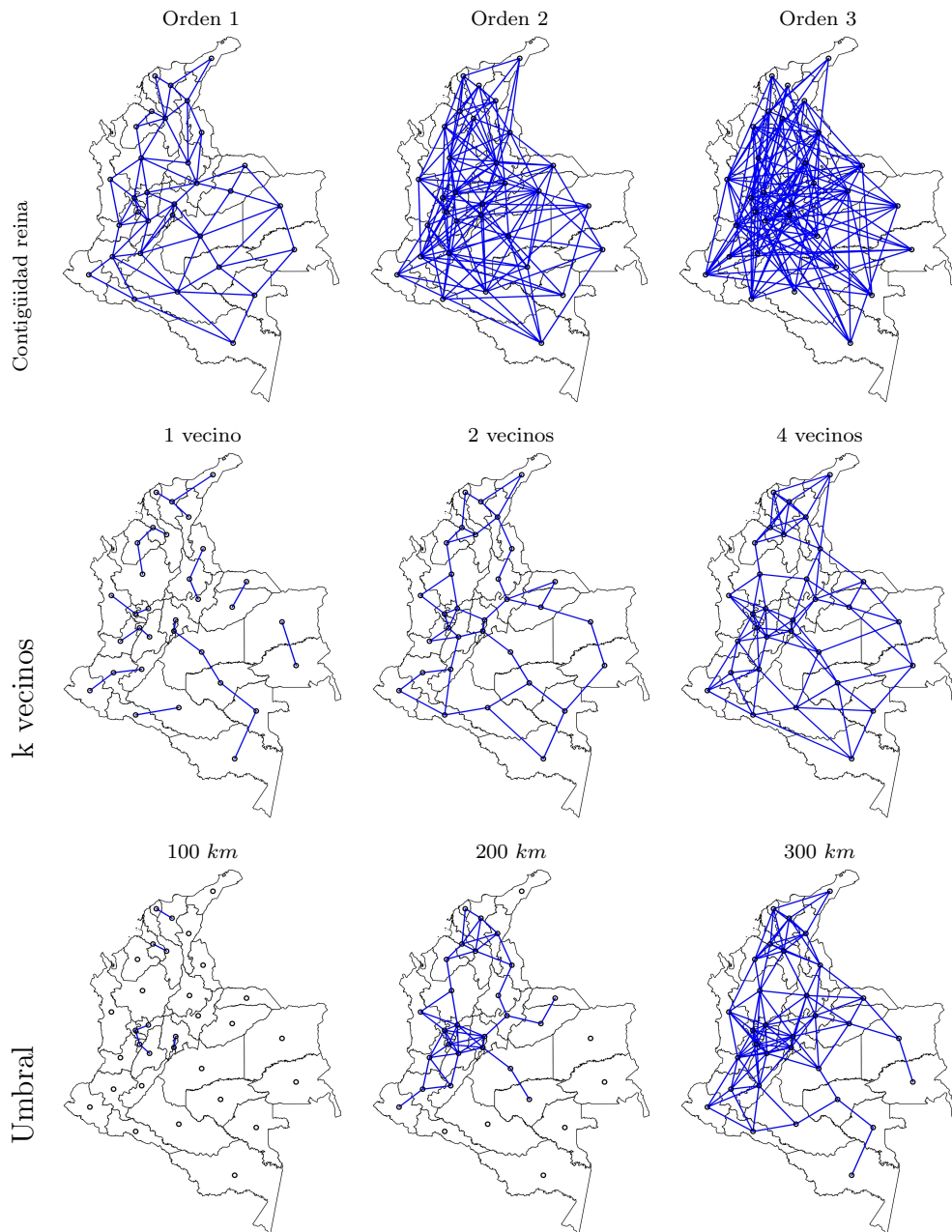


FIGURA 5.3: Mapas de estructuras de vecindarios para la construcción de pesos espaciales

El dispersograma se construye considerando el número de AA y su rezago espacial (spatial lag, SL). Además, la Figura 5.4 permite identificar influyentes espaciales y espacio-temporales. Los influyentes se obtienen luego de calcular los estadísticos  $dffits$  y la razón de covarianza ( $cov.r$ ) de los residuales obtenidos del modelo de regresión lineal entre el número de AA y su rezago espacial. El



dispersograma de Moran en general muestra las interrelaciones entre cada uno de los polígonos y sus vecinos; la pendiente del modelo de regresión mencionado se suele denominar I de Moran. Luego si la pendiente, o en otras palabras el I de Moran, es positivo indicará un predominio de concentraciones espaciales de valores similares para la variable analizada, bien sean altos rodeados de altos o bajos rodeados de bajos. Y si la pendiente es negativa indicará una concentración de valores disimilares de la variable analizada, así se tendrá valores bajos rodeados de altos o altos rodeados de bajos. Por ejemplo, en el año 2003, la Figura 5.4 indica una autocorrelación espacial negativa y muestra a Bogotá y Magdalena como influyentes, ya que predomina en el espacio una concentración de valores disimilares. Además, tanto Bogotá como Magdalena, se asocian en concentraciones de valores similares, alto rodeado de altos y bajo rodeado de bajos, respectivamente.

De los períodos analizados tan sólo en el año 2006 no se observa una pendiente significativamente diferente de 0, lo cual indicaría ausencia de autocorrelación espacial para tal año. Para los demás años, los departamentos que constantemente se muestran como influyentes son Chocó, Cauca y Quindío.

Una vez se han establecido los influyentes espacialmente, se determinan los valores más altos “H” y más bajos “L” para cada uno de los polígonos en cada año. Posteriormente, se cruzan los rezagos espaciales asociados a su vecindario: si una región presenta un valor alto y esta rodeada por valores altos en promedio para las regiones vecinas entonces se denota por “HH”, si una región presenta un valor bajo y a la vez esta rodeada por valores bajos en promedio para las regiones vecinas entonces se denota por “LL”, si una región presenta un valor alto y a la vez esta rodeada por valores bajos en promedio para las regiones vecinas entonces se denota por “HL” y si una región presenta un valor bajo y a la vez esta rodeada por valores altos en promedio para las regiones vecinas entonces se denota por “LL”. Por consiguiente, se observa que en los años 2003 y 2009 hay dos influyentes por cuadrante de dispersión de Moran (Bogotá y Magdalena). Además, en los años 2004 y 2005, los influyentes son Chocó y Quindío, para estos 2 años y los años 2006, 2007 y 2008, los influyentes corresponde a valores disimilares, es decir, “HL” o “LH”. Estos resultados se muestran en la Figura 5.5.

La Figura 5.6 muestra el comportamiento del número de AA para el período 2003 a 2009. Este período evidencia una disminución de los niveles de las AA, lo cual indica un mayor control por parte de la fuerza pública (policía y fuerzas militares) y políticas más acertadas en el control de los grupos insurgentes como las FARC y ELN.

A continuación se proceden a ajustar los modelos GLSTARAR y DBDGLSTARAR. En primer lugar se presentan los resultados obtenidos al emplear el modelo GLSTARAR, y posteriormente, se presentaran los resultados del mo-

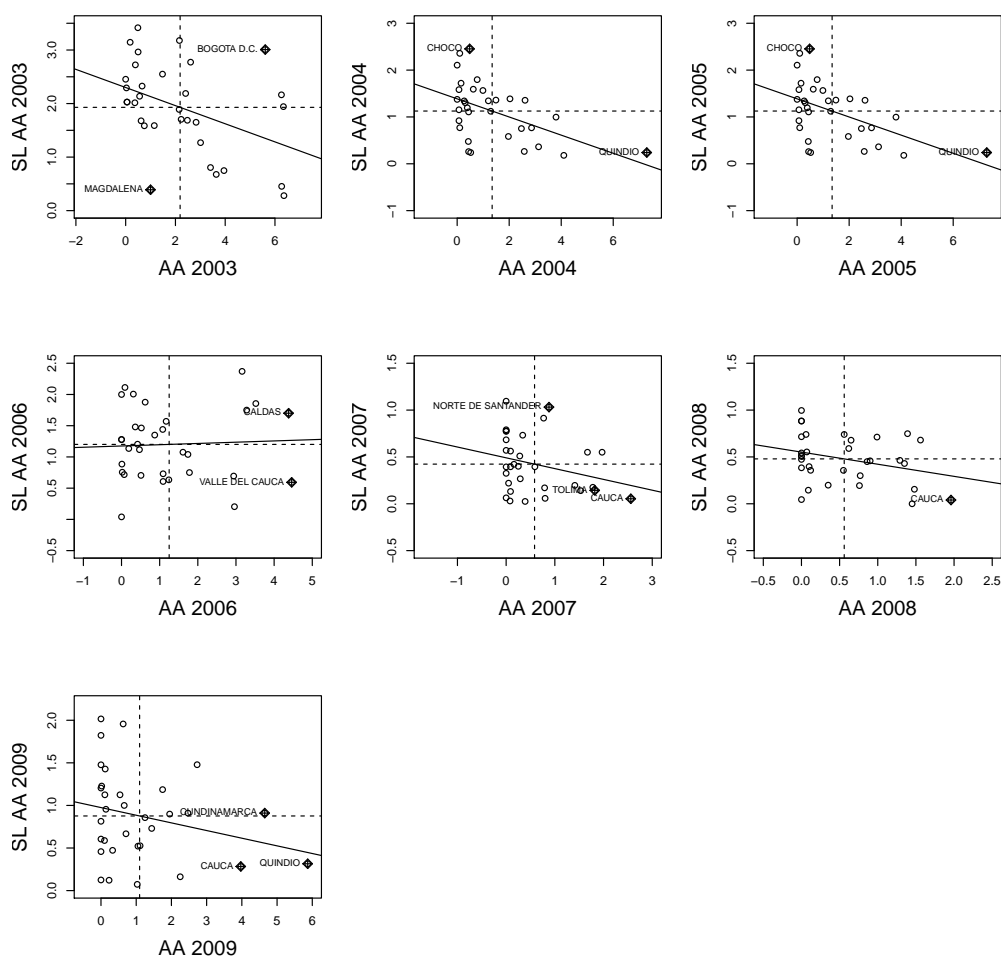


FIGURA 5.4: Gráficos de dispersión de Moran del número de acciones armadas por departamento, período 2003 a 2009

delo DBGLSTARAR.

### 5.5.2 Modelo GLSTARAR

La dependencia residual espacial se maneja de la siguiente manera: sea  $y(\mathbf{s}_i, t)$  que denota el número de AA en el  $\mathbf{s}_i$ -ésimo departamento en el  $t$ -ésimo tiempo ( $s_i = 1, \dots, 32$  y  $t = 1, \dots, 7$ ). En el modelo se supone que los  $y(\mathbf{s}_i, t)$ 's son variables poisson independientes dado el proceso estocástico espacial no observado  $\varepsilon_{it}$ , y además, que la respuesta media en  $(\mathbf{s}_i, t)$  depende de las variables explicativas observadas en la ubicación  $\mathbf{s}_i$  y tiempo  $t$ . Específicamente,

se puede escribir el modelo GLSTARAR como

$$\begin{aligned}\log y(\mathbf{s}_i, t) &= \gamma_0 + \mathbf{v}_{2it}^t \boldsymbol{\gamma}_t + \pi_t \sum_{i'=1}^{32} w_{ii'}^{(1)} \log y(\mathbf{s}_{i'}, t) + \varepsilon_{it} \\ \varepsilon_{it} = \varepsilon(\mathbf{s}_i, t) &= \psi_t \sum_{i'=1}^{32} w_{ii'}^{(2)} \varepsilon_{i't} + e_{it}\end{aligned}\tag{5.51}$$

con  $i = 1, \dots, 32$ ,  $t = 1, \dots, 7$ , donde  $\gamma_0$  es el parámetro del intercepto desconocido,  $\mathbf{v}_{2it}^t = \mathbf{v}_2^t(\mathbf{s}_i, t) = (AVV2003_{i1}, FM2003_{i1}, \dots, DFHR2009_{i7})$  es un vector que contiene las variables explicativas asociadas al espacio-tiempo en la  $\mathbf{s}_i$ -ésima ubicación y  $\boldsymbol{\gamma}_t = (\gamma_{11}, \dots, \gamma_{7p_{27}})^t$  es un vector de parámetros de regresión espacio-tiempo desconocidos. Además,  $w_{ii'}^{(1)} = w_{ii'}^{(2)}$  son los pesos utilizando un rezago espacial de orden uno o efecto reina de orden uno (ver detalles en (Anselin 1988)),  $\pi_t$  es el coeficiente autorregresivo espacial en el  $t$ -ésimo período de tiempo,  $\varepsilon_{it}$  refleja el término de error autocorrelacionado espacial para la  $\mathbf{s}_i$ -ésima ubicación en el  $t$ -ésimo tiempo,  $\psi_t$  es llamado el coeficiente de autocorrelación espacial en el  $t$ -ésimo período de tiempo y  $e_{it} = e(\mathbf{s}_i, t)$  es un error normal idénticamente distribuido asociado a la  $\mathbf{s}_i$ -ésima ubicación en el  $t$ -ésimo tiempo con media cero y covarianza  $E(e_{it}, e_{i't'}) = \sigma_{tt'}$  para  $t, t' = 1, \dots, 7$ , con  $E(e_{it}, e_{i't}) = 0$  para  $i, i' = 1, \dots, 32$ .

En las Tablas 5.2 y 5.3 se presentan los resultados de la estimación de los parámetros del modelo GLSTARAR presentado en la ecuación (5.51) después de realizar el proceso iterativo del final de la Sección 5.3 utilizando el método GEE, pero sin emplear el método de distancias, sino las variables en su escala natural. En este modelo se consideraron todas las estructuras de correlación de los errores  $e_{it}$  expuestas en la Sección 5.3.4; sin embargo, ninguno de los parámetros de las matrices de correlación fueron significativos al 5%, por lo cual se podrían considerar estos errores independientes. El pseudo- $R^2$  en este caso fue igual a 69.85%, lo cual indica un aparente buen ajuste del modelo.

En la Tabla 5.2 se observa que todos los rezagos espaciales de los errores fueron significativos al 5%. Además, como los coeficientes de rezago espacial entre tiempos no son los mismos, se puede decir que hay una interacción entre espacio y tiempo en el modelo de error espacial. Sin embargo, estos coeficientes se parecen un poco, por lo que se debería realizar una prueba de igualdad de pendientes; en este capítulo no se desarrollo dicha prueba, pero en un futuro trabajo se podría realizar una prueba asintótica para juzgar la igualdad entre parámetros. Los signos de todos los parámetros son positivos, pero no se observa un crecimiento o decrecimiento de los mismos a lo largo del tiempo, el signo positivo se puede interpretar como que un departamento con alto/bajo error esta rodeado de departamentos con alto/bajo error. El parámetro de escala en este caso corresponde a  $\sigma^2 = 0.837$ , lo cual significa que la dispersión es baja.

En la Tabla 5.3 se observa que todos los rezagos espaciales del log del

TABLA 5.2: Estimación de los parámetros del modelo GLSTARAR para el error espacial mediante GEE.

Variable de Rezago espacial	Coefficientes estimados	Error estándar	Wald	$Pr(>  W )$
WE1.lag	0.1932	0.0227	72.4	<2e-16
WE2.lag	0.1839	0.0214	74.0	<2e-16
WE3.lag	0.1897	0.0268	50.2	1.4e-12
WE4.lag	0.1917	0.0171	125.1	<2e-16
WE5.lag	0.1659	0.0316	27.6	1.5e-07
WE6.lag	0.1980	0.0158	156.2	<2e-16
WE7.lag	0.1737	0.0279	38.6	5.2e-10
Parámetro de escala estimado (Intercepto)	0.837	0.434		

número de AA fueron significativos al 5%. Al igual que en el modelo de error espacial, como los coeficientes de rezago espacial entre tiempos no son los mismos, hay una interacción entre espacio y tiempo en el modelo de rezago espacial ya que todos los signos de los coeficientes varían a través del tiempo. Esto quiere decir que en un momento en el tiempo, por ejemplo, para el año 2009 (WY7.lag), un determinado departamento tiene relación positiva ( $1.15e-01$ ) con sus vecinos, es decir que un departamento tiene un número alto/bajo de AA de las FARC y ELN y sus vecinos alrededor presentan un número alto/bajo de AA. Sin embargo, en el año 2008 (WY6.lag), esta relación fue negativa ( $-4.49e-02$ ), lo cual significa que un departamento con un número alto de AA de las FARC y ELN tiene vecinos alrededor que presentan un número bajo de AA o viceversa.

El parámetro de escala es  $\phi = 0.615$ , el cual significa que la dispersión es baja y no hay problemas fuertes aparentemente de sobredispersión. De otra parte, todas las variables explicativas en los diferentes años, resultaron significativas al 5%. En este caso, como los coeficientes de las variables explicativas cambian a lo largo del tiempo, también hay interacción entre tiempo y espacio para las diferentes variables explicativas consideradas. No hay un comportamiento especial de los coeficientes de las variables a lo largo del tiempo, ya que muchas cambian sus signos a lo largo del tiempo, en ciertos momentos tienen efecto positivo y en otro un efecto negativo. Una variable que tiene un signo positivo a lo largo del tiempo es la variable total de confrontaciones armadas por año, lo cual significa que a mayor cantidad de confrontaciones armadas mayor es el número de AA por cada  $1000 \text{ km}^2$  de los grupos guerrilleros de las FARC y el ELN.

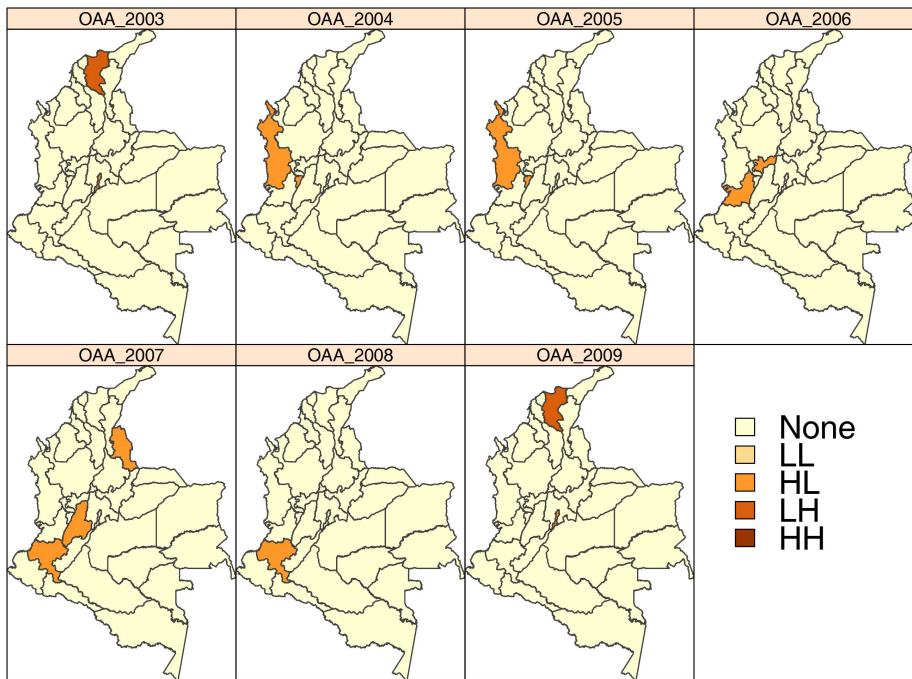


FIGURA 5.5: Mapas de influentes por cuadrante de dispersión de Moran del número de acciones armadas por departamentos, período 2003 a 2009

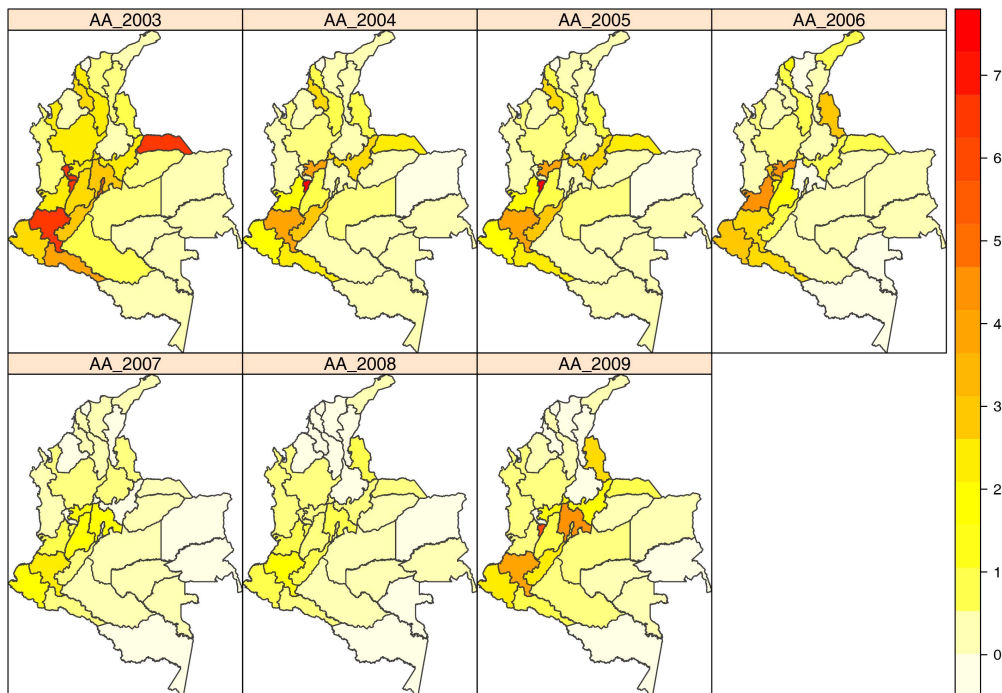


FIGURA 5.6: Mapas del número de acciones armadas por departamento, período 2003 a 2009

TABLA 5.3: Estimación de los parámetros del modelo GLSTARAR para el rezago espacial mediante GEE.

Variable	Coefficientes estimados	Error estándar	Wald	$Pr(>  W )$
(Intercepto)	-8.48e-01	1.93e-07	1.94e+13	<2e-16
WY1.lag	4.28e-02	2.85e-09	2.24e+14	<2e-16
WY2.lag	1.19e-02	3.15e-09	1.43e+13	<2e-16
WY3.lag	2.11e-02	6.12e-09	1.19e+13	<2e-16
WY4.lag	-9.95e-03	5.56e-09	3.20e+12	<2e-16
WY5.lag	-2.93e-02	4.16e-07	4.97e+09	<2e-16
WY6.lag	-4.49e-02	2.18e-08	4.23e+12	<2e-16
WY7.lag	1.88e-02	1.24e-08	2.29e+12	<2e-16
AVV2003	1.15e-01	5.58e-09	4.27e+14	<2e-16
FM2003	-1.57e-01	7.07e-09	4.93e+14	<2e-16
CAA2003	2.02e-01	4.57e-09	1.96e+15	<2e-16
DFHE2003	1.60e-03	1.10e-10	2.11e+14	<2e-16
DFHR2003	-1.73e-04	2.73e-11	4.05e+13	<2e-16
URB2003	-6.21e-01	1.97e-07	9.97e+12	<2e-16
AVV2004	-1.50e-03	1.42e-10	1.12e+14	<2e-16
FM2004	-6.63e-01	2.96e-08	5.02e+14	<2e-16
CAA2004	5.87e-01	1.84e-08	1.02e+15	<2e-16
DFHE2004	4.15e-03	3.98e-10	1.08e+14	<2e-16
DFHR2004	-1.38e-03	7.02e-11	3.89e+14	<2e-16
URB2004	-4.65e-01	1.99e-07	5.44e+12	<2e-16
AVV2005	-3.93e-02	1.32e-09	8.87e+14	<2e-16
FM2005	1.41e-01	4.87e-08	8.34e+12	<2e-16
CAA2005	9.05e-02	2.69e-08	1.13e+13	<2e-16
DFHE2005	7.51e-03	2.98e-10	6.33e+14	<2e-16
DFHR2005	3.06e-04	2.23e-10	1.89e+12	<2e-16
URB2005	-1.31e+00	8.74e-08	2.23e+14	<2e-16
AVV2006	4.77e-02	3.21e-09	2.21e+14	<2e-16
FM2006	-1.95e-01	5.39e-10	1.31e+17	<2e-16
CAA2006	2.91e-01	7.26e-10	1.61e+17	<2e-16
DFHE2006	5.08e-03	3.16e-10	2.59e+14	<2e-16
DFHR2006	-3.04e-04	8.46e-13	1.29e+17	<2e-16
URB2006	-2.25e+00	6.23e-08	1.30e+15	<2e-16
AVV2007	1.00e-02	6.89e-09	2.11e+12	<2e-16
FM2007	-6.97e-01	1.30e-06	2.88e+11	<2e-16
CAA2007	9.07e-01	1.37e-06	4.37e+11	<2e-16
DFHE2007	-8.32e-04	3.85e-09	4.66e+10	<2e-16
DFHR2007	-1.60e-03	4.03e-08	1.57e+09	<2e-16
URB2007	-2.80e+00	7.40e-06	1.43e+11	<2e-16
AVV2008	-1.88e-02	5.02e-08	1.40e+11	<2e-16
FM2008	-5.07e-01	1.78e-07	8.09e+12	<2e-16
CAA2008	8.41e-01	1.46e-07	3.34e+13	<2e-16
DFHE2008	2.86e-03	1.37e-09	4.35e+12	<2e-16
DFHR2008	-9.97e-04	2.91e-10	1.18e+13	<2e-16
URB2008	-2.38e+00	3.86e-07	3.79e+13	<2e-16

TABLA 5.4: Estimación de los parámetros del modelo GLSTARAR para el rezago espacial mediante GEE (Continuación Tabla 5.3).

Variable	Coefficientes estimados	Error estándar	Wald	$Pr(>  W )$
AVV2009	-4.59e-02	6.41e-09	5.13e+13	<2e-16
FM2009	1.13e-01	3.19e-08	1.25e+13	<2e-16
CAA2009	1.14e-01	1.60e-08	5.06e+13	<2e-16
DFHE2009	1.12e-02	1.69e-09	4.40e+13	<2e-16
DFHR2009	-5.82e-04	7.49e-10	6.04e+11	<2e-16
Parámetro de escala estimado				
(Intercepto)	0.615	3.24e-08		

Debido a los posibles problemas de multicolinealidad de las variables explicativas en el modelo ajustado GLSTARAR de acuerdo a la ecuación (5.51) y por el exceso de parámetros que se ocasiona en el mismo, se realiza una reducción de variables explicativas utilizando el método DB.

### 5.5.3 Modelo DBGLSTARAR

Para el ajuste del modelo DBSTARAR, se emplea la distancia euclidiana entre individuos dada en (5.7) porque todas las variables explicativas son continuas. De este modo, se obtuvieron las distancias empleando las variables explicativas centradas y la matriz  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$ . Entonces, se construyeron las coordenadas principales utilizando las variables explicativas: tasa de atención a víctimas de la violencia (AVV), desplazamiento forzado - hogares expulsados (DFHE), desplazamiento forzado - hogares recibidos (DFHR), total de confrontaciones armadas por año (CAA), número de AA por parte de las fuerzas militares (FM) y porcentaje de personas que residen el área urbana (URB). Para seleccionar las coordenadas principales a incluir en el modelo, se observó las más significativas de acuerdo al valor más alto de Wald, presentado en la ecuación (5.47).

Específicamente, para este estudio se seleccionaron siete coordenadas principales ( $\mathbf{X}_1$ ,  $\mathbf{X}_5$ ,  $\mathbf{X}_6$ ,  $\mathbf{X}_7$ ,  $\mathbf{X}_{16}$ ,  $\mathbf{X}_{27}$  y  $\mathbf{X}_{34}$ ), lo cual fue realizado con la finalidad de mostrar el buen funcionamiento del método DB y observar la gran reducción de variables explicativas en comparación al modelo GLSTARAR.

Con los mismos supuestos del modelo (5.51) sobre la variable respuesta y

los errores, se puede escribir el modelo DBGLSTARAR como

$$\begin{aligned}\log y(\mathbf{s}_i, t) &= \mathbf{x}_{it}^t \boldsymbol{\beta} + \pi_t \sum_{i'=1}^{32} w_{ii'}^{(1)} \log y(\mathbf{s}_{i'}, t) + \varepsilon_{it} \\ \varepsilon_{it} &= \psi_t \sum_{i'=1}^{32} w_{ii'}^{(2)} \varepsilon_{i't} + e_{it}\end{aligned}\tag{5.52}$$

con  $i = 1, \dots, 32$  y  $t = 1, \dots, 7$ , y donde  $\mathbf{x}_{it}^t = (1, x_{i0}, \dots, x_{i7})$  y  $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \dots, \beta_7)$  es un vector de parámetros desconocido. Los demás términos y supuestos son iguales que en el modelo (5.51).

En las Tablas 5.4 y 5.5 se presentan los resultados de la estimación de los parámetros del modelo DBGLSTARAR presentado en la ecuación (5.52) después de realizar el proceso iterativo del final de la Sección 5.3 utilizando el método GEE. En este modelo se consideraron todas las estructuras de correlación de los errores  $e_{it}$  expuestas en la Sección 5.3.4; sin embargo, al igual que en el modelo GLSTARAR (5.51) ningún parámetro de la matriz de correlación fue significativo al 5%, por lo cual se podrían considerar estos errores independientes. El pseudo- $R^2$  en este caso fue igual a 45.63%, lo cual indica un aparente buen ajuste del modelo; sin embargo este no es tan bueno como en el modelo GLSTARAR, lo cual en principio es debido a haber tomado menos número de coordenadas principales que número de variables explicativas.

En la Tabla 5.4 se observa, al igual que en modelo GLSTARAR, que todos los rezagos espaciales de los errores fueron significativos al 5%. Además, como los coeficientes de rezago espacial entre tiempos no son los mismos a lo largo del tiempo, se puede decir que hay una interacción entre espacio y tiempo en el modelo de error espacial. Los signos de todos los parámetros son positivos, lo cual significa que un departamento con alto/bajo error esta rodeado de departamentos con alto/bajo error. El parámetro de escala es  $\sigma^2 = 0.428$ , lo cual significa que la dispersión es baja y que este modelo es menos preciso que el modelo GLSTARAR ( $\sigma^2 = 0.837$ ), debido a que el  $\sigma^2$  del modelo DBGLSTARAR es más bajo que el del modelo GLSTARAR. Este comportamiento del método DB en el modelo de error espacial se presenta porque se tienen mucho menos variables (coordenadas principales) que variables explicativas en el modelo GLSTARAR.

En la Tabla 5.5 se observa que todos los rezagos espaciales del log del número de AA fueron significativos al 5%. Al igual que en el modelo de error espacial, como los coeficientes de rezago espacial entre tiempos no son los mismos (los signos y los parámetros estimados varían a través del tiempo), hay una interacción entre espacio y tiempo en el modelo de rezago espacial. Esto significa que en un momento en el tiempo, por ejemplo, para el año 2009 (WY7.lag), un determinado departamento tiene relación positiva ( $2.19\text{e-}02$ ) con sus departamentos vecinos, es decir que un departamento tiene un número



TABLA 5.5: Estimación de los parámetros del modelo DBGLSTARAR para el error espacial mediante GEE.

Variable de Rezago espacial	Coefficientes estimados	Error estándar	Wald	$Pr(>  W )$
WE1.lag	0.2064	0.0234	77.9	<2e-16
WE2.lag	0.1894	0.0336	31.8	1.7e-08
WE3.lag	0.1835	0.0308	35.4	2.7e-09
WE4.lag	0.1991	0.0260	58.8	1.7e-14
WE5.lag	0.1847	0.0242	58.1	2.5e-14
WE6.lag	0.1913	0.0181	112.0	<2e-16
WE7.lag	0.1814	0.0233	60.7	6.8e-15
Parámetro de escala estimado (Intercepto)	0.428	0.0332		

alto/bajo de AA de las FARC y ELN y sus departamentos vecinos alrededor presentan un número alto/bajo de AA. Sin embargo, en el año 2006 (WY4.lag), esta relación fue negativa (-2.71e-03), lo cual significa que un departamento con un número alto de AA de las FARC y ELN tiene departamentos vecinos alrededor que presentan un número bajo de AA o viceversa.

El parámetro de escala es  $\phi = 0.945$ , el cual significa que la dispersión es baja y no hay problemas fuertes aparentemente de sobredispersión. Además, como este coeficiente es más alto que el obtenido en el modelo GLSTARAR de rezago espacial (0.615), entonces se puede decir que el modelo DBGLSTARAR de rezago espacial es más preciso que el modelo GLSTARAR. De otra parte, todas las coordenadas principales resultaron significativas al 5%, lo cual es muy bueno ya que estas coordenadas no están correlacionadas entre sí y recogen la información de las variables explicativas originales. Las coordenadas principales  $\mathbf{X}5$  y  $\mathbf{X}6$  afectan de forma positiva al número de AA por cada 1000  $km^2$  de los grupos guerrilleros FARC y ELN, es decir, valores altos de estas coordenadas producen un mayor número de AA. Mientras las coordenadas principales,  $\mathbf{X}1$ ,  $\mathbf{X}7$ ,  $\mathbf{X}16$ ,  $\mathbf{X}27$  y  $\mathbf{X}34$ , afectan de forma negativa al número de AA, es decir, valores altos de estas coordenadas producen un menor número de AA de los grupos guerrilleros.

#### 5.5.4 Validación de los supuestos sobre los modelos GLSTARAR y DBGLSTARAR

En la Figura 5.6 se encuentra que el valor observado para el departamento de Cundinamarca esta alrededor de 5 y la predicción con los modelos GLSTARAR

TABLA 5.6: Estimación de los parámetros del modelo DBGLSTARAR para el rezago espacial mediante GEE.

Variable	Coefficientes estimados	Error estándar	Wald	$Pr(>  W )$
(Intercepto)	-3.66e-01	2.63e-08	1.94e+14	<2e-16
WY1.lag	1.71e-03	4.73e-10	1.31e+13	<2e-16
WY2.lag	1.50e-02	2.92e-09	2.66e+13	<2e-16
WY3.lag	1.32e-02	6.20e-08	4.50e+10	<2e-16
WY4.lag	-2.71e-03	2.19e-09	1.53e+12	<2e-16
WY5.lag	1.18e-01	3.80e-09	9.58e+14	<2e-16
WY6.lag	8.80e-02	1.42e-08	3.85e+13	<2e-16
WY7.lag	2.19e-02	3.73e-09	3.45e+13	<2e-16
X1	-3.59e+00	2.18e-08	2.70e+16	<2e-16
X5	2.42e+00	2.49e-08	9.51e+15	<2e-16
X6	5.62e+00	1.86e-08	9.10e+16	<2e-16
X7	-2.63e+01	1.09e-07	5.82e+16	<2e-16
X16	-1.01e+01	1.42e-06	5.07e+13	<2e-16
X27	-1.52e+01	2.61e-08	3.41e+17	<2e-16
X34	-1.26e+01	2.62e-07	2.30e+15	<2e-16
Parámetro de escala estimado				
(Intercepto)	0.945	1.2e-07		

y DBGLSTARAR es de aproximadamente 2 (ver Figuras 5.7 y 5.10), es decir hay un error aproximado de 3. Este error se debe a que en el dispersograma de Moran (ver Figura 5.4), Cundinamarca es un influyente, luego dadas las características espaciales de contigüidad de este tipo de modelos, no se pueden remover outliers; sin embargo, es un tema interesante a considerar en futuros trabajos afines. En general, las predicciones generadas con el modelo GLSTARAR producen buenos resultados tal como se observa en la Figura 5.7 y en los mapas de residuales de Pearson y deviance asociados a tal modelo (ver Figuras 5.8 y 5.9). En estos mapas, para el año 2009, se observa residuales cercanos a 3 (residual de Pearson) y 2 (residual de deviance) para el departamento de Cundinamarca, lo cual concuerda con lo mencionado previamente.

Con respecto al modelo DBGLSTARAR, se observan también en la Figura 5.10 que las predicciones del número de AAs disminuyen con el tiempo. En general, se tienen buenas predicciones con un valor máximo de 6.34 para el número de AA, mientras que en el modelo GLSTARAR la predicción máxima esta en 8.72, las medias son muy similares 1.16 y 1.15, respectivamente; al compararlas con los valores observados para el número de AA (media 1.19 y máximo igual a 7.29), se puede decir que los dos modelos presentan comportamientos similares en términos de las medias, pero para los valores máximos,

el modelo DBGLSTARAR generó un valor inferior al observado, mientras que el modelo GLSTARAR generó un valor superior al observado.

De acuerdo a los mapas de residuales de Pearson y deviance (ver Figuras 5.11 y 5.12) asociados al modelo DBGLSTARAR, se observa que no hay aparentemente un grave problema de atipicidades en los residuos para los diferentes departamentos a lo largo del tiempo, ya que en el caso de los residuales de Pearson, éstos no se salen en su gran mayoría del intervalo  $(-3, 3)$  y en el caso de los residuales de deviance entre  $(-2, 2)$ . En el caso del departamento de Cundinamarca para el año 2009 sucede lo mismo que en el ajuste del modelo GLSTARAR, éste departamento si podría ser un dato atípico para el año 2009.

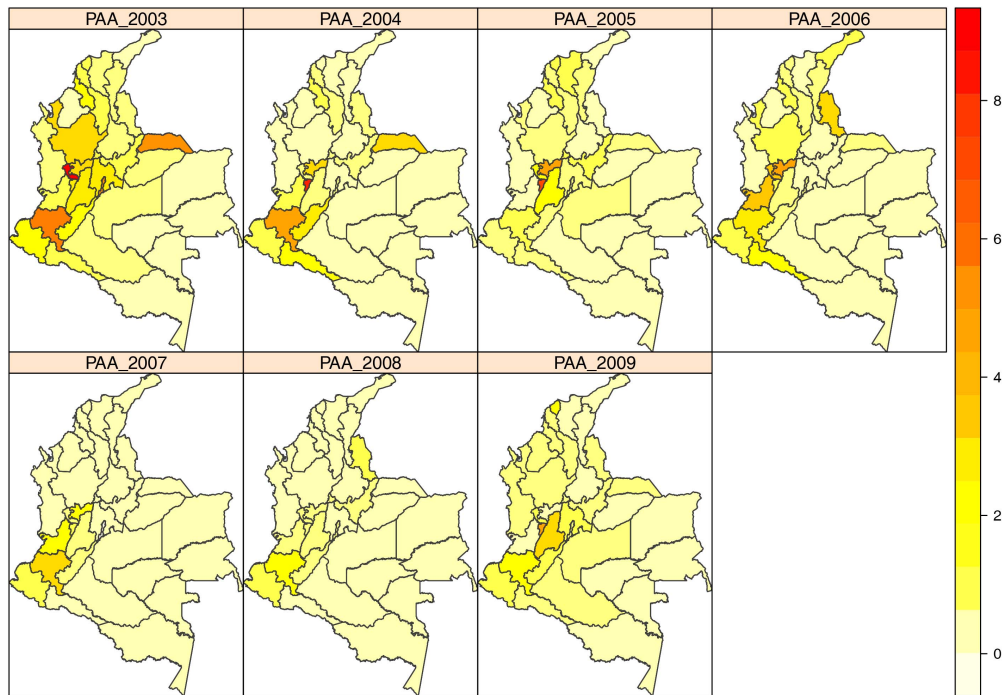


FIGURA 5.7: Mapas de la predicción del número de acciones armadas por departamento, bajo el modelo GLSTARAR para el período 2003 a 2009

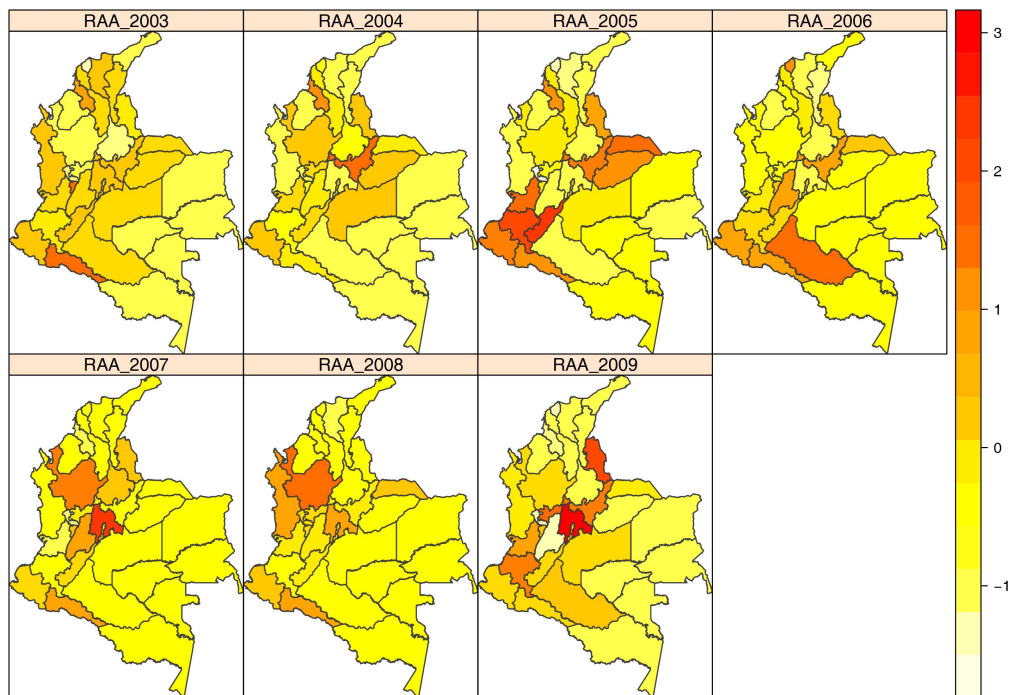


FIGURA 5.8: Mapas de residuos de Pearson por departamento bajo el modelo GLS-TARAR para el período 2003 a 2009

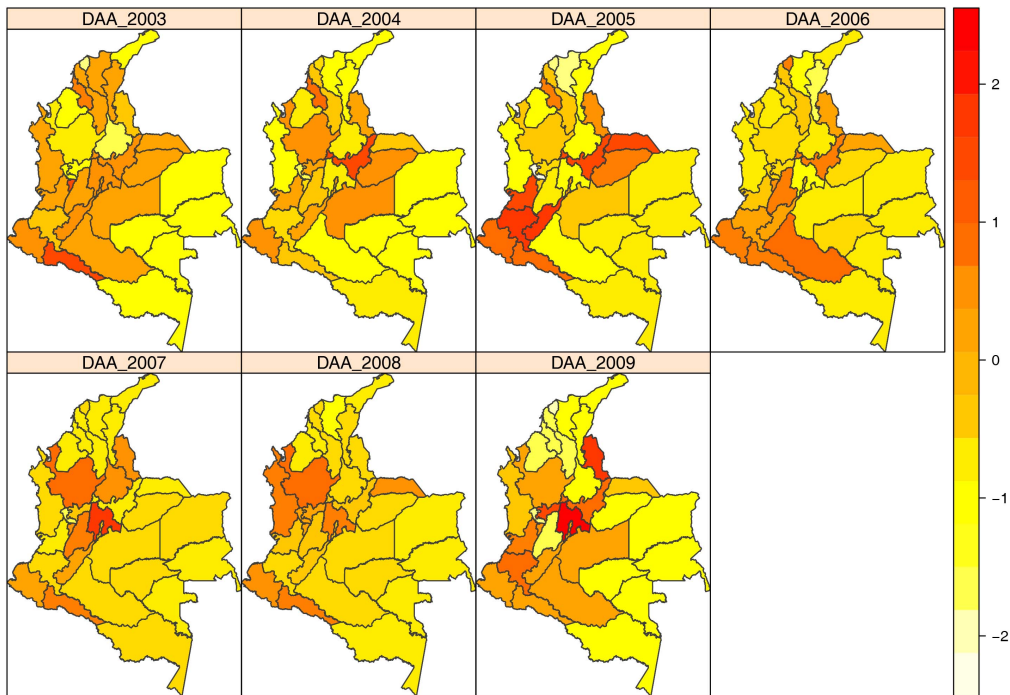


FIGURA 5.9: Mapas de deviancias por departamento bajo el modelo GLSTARAR para el período 2003 a 2009

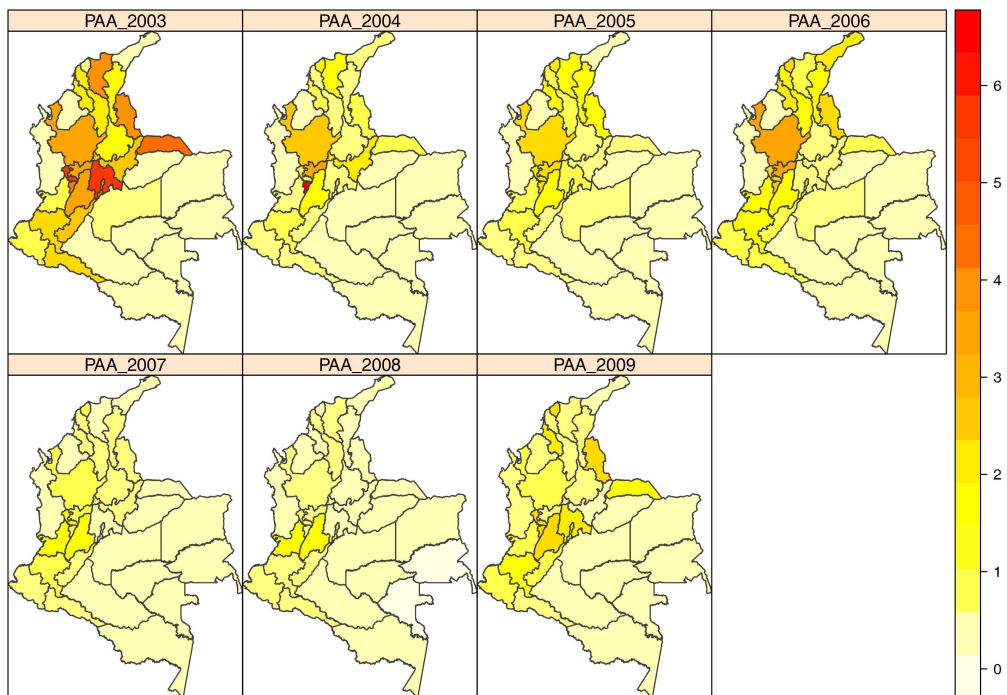


FIGURA 5.10: Mapas de la predicción del número de acciones armadas por departamento bajo el modelo DBGLSTARAR para el período 2003 a 2009

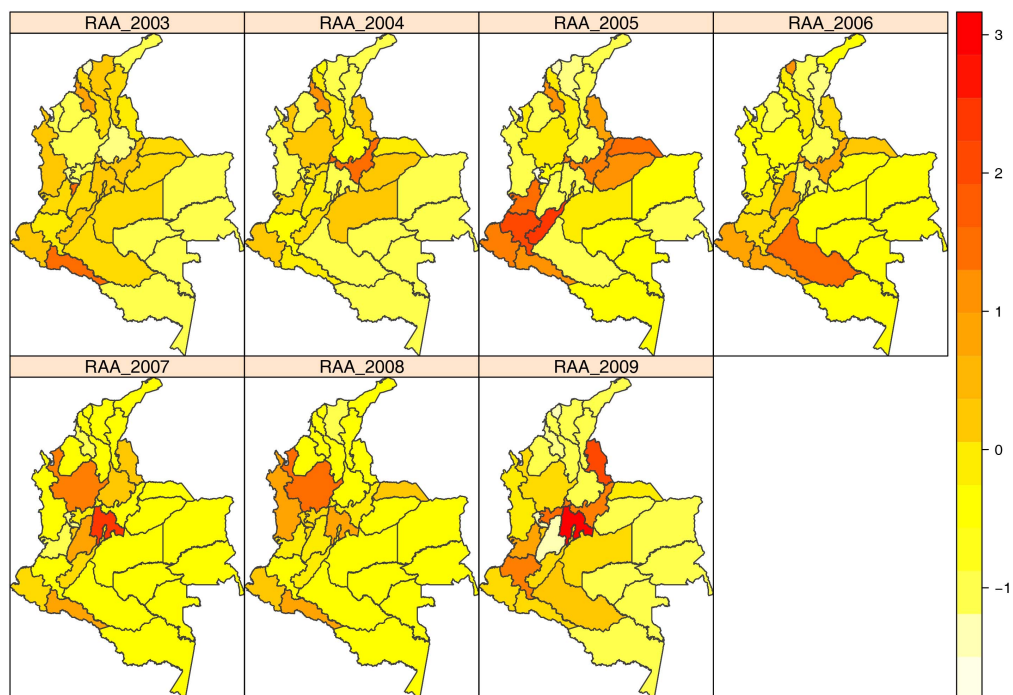


FIGURA 5.11: Mapas de residuos de Pearson por departamento bajo el modelo DBGLSTARAR para el período 2003 a 2009

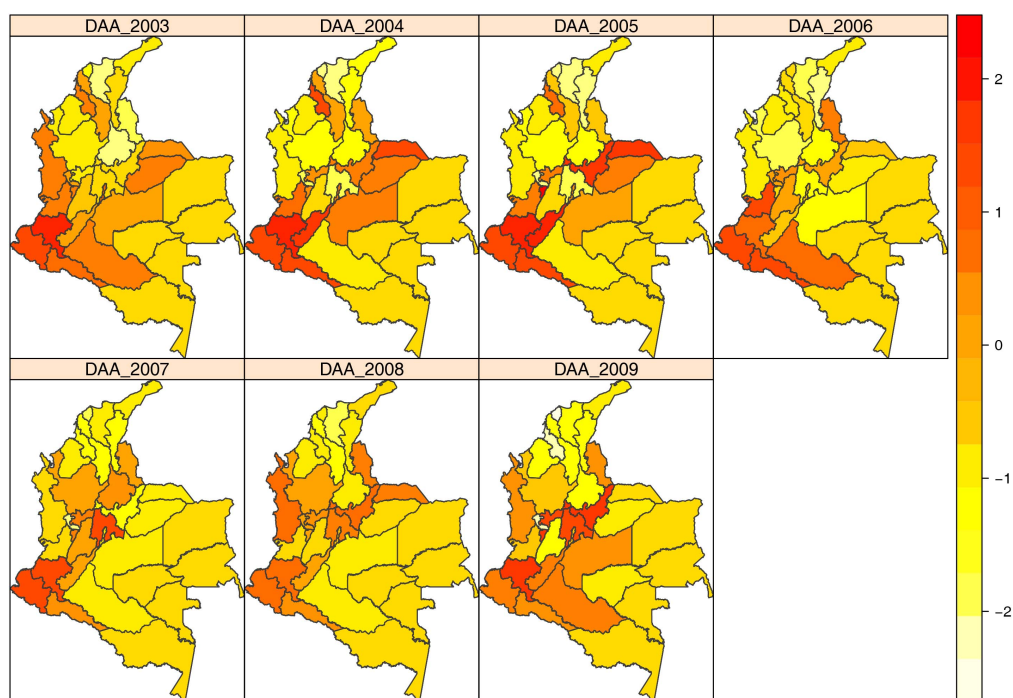


FIGURA 5.12: Mapas de deviancias por departamento, bajo el modelo DBGLSTARAR para el período 2003 a 2009



# Capítulo 6

## Vector autorregresivo espacial lineal generalizado mixto basado en distancias incorporando la dinámica tanto espacial como temporal

### 6.1 Introducción

En este capítulo se presenta un modelo autorregresivo espacial lineal generalizado mixto utilizando el método basado en distancias propuesto por Cuadras (1989). Este modelo incluye retrasos tanto espaciales como temporales entre vectores de variables de estado estacionarias. Por lo tanto, este modelo contiene perturbaciones que están correlacionadas tanto espacialmente como temporalmente. Aunque los parámetros estructurales no están completamente identificados en este modelo, los coeficientes de rezago espacial contemporáneos se pueden identificar mediante las variables explicativas de estado. Se utiliza la dinámica espacial de los datos econométricos tipo panel para estimar el modelo propuesto. Por lo tanto, los parámetros involucrados en el modelo se estimaron utilizando el método MCMC mediante máxima verosimilitud, la cual es una técnica factible y útil en este caso. Además, se discute en este capítulo la interacción entre estacionariedad temporal y espacial, y se derivan las respuestas de impulso para el modelo propuesto, lo cual naturalmente dependerá de la dinámica temporal y espacial del modelo.

En los últimos años, la literatura de datos espaciales ha exhibido un creciente interés en la especificación y estimación de relaciones por ejemplo sociales y econométricas basada en paneles espaciales. Los paneles espaciales suelen refe-



rirse a los datos que contienen observaciones de series de tiempo de un número de unidades espaciales (códigos postales, municipios, regiones, estados, jurisdicciones, países, etc.). Este interés puede ser explicado por el hecho que los datos panel ofrecen a los investigadores extender las posibilidades de modelamiento en comparación con el ajuste de una sola ecuación de corte transversal, lo cual fue el enfoque principal de la literatura social y econométrica espacial por mucho tiempo (Hsiao 2003). Un grupo considerablemente más pequeño (Elhorst 2003, Lee 2004, Elhorst 2005, Elhorst 2008) se ocupa del análisis social y econométrico de la dinámica espacial en los modelos de datos panel temporalmente estáticos.

En muchos estudios que tienen una variable respuesta continua se han aplicado estas estrategias para estimar modelos: regionales del mercado de trabajo, de crecimiento económico, de gasto público o de ajuste fiscal, y agrícolas. Sin embargo, cuando la variable respuesta es un conteo, una tasa o una respuesta binaria, no hay demasiada literatura que resuelva el problema utilizando el modelo autorregresivo espacial e incluyendo rezagos tanto temporales como espaciales entre vectores de variables de estado estacionarias. Por lo tanto, no solo muchas de estas aplicaciones han dado lugar a nuevas ideas, desarrollos y ampliaciones, sino también nuevos cuestionamientos. Los científicos han demostrado que la dependencia espacial de los datos sociales o económicos pueden alterar, e incluso revertir, los resultados de los modelos estándares de series de tiempo.

La econometría espacial como tal es un subcampo de la econometría que se ocupa de la interacción espacial (autocorrelación espacial) y la estructura espacial (heterogeneidad espacial) en modelos de regresión transversal y de datos panel (Paelinck & Klaassen 1979, Anselin 1988). Este enfoque en la localización y la interacción espacial recientemente ha ganado un lugar más central no sólo en lo aplicado sino también en la econometría teórica. En el pasado, los modelos que incorporaban explícitamente espacio o localización geográfica se encontraban principalmente en campos especializados tales como ciencia regional, urbana, economía inmobiliaria y geografía económica (Anselin & Florax 1995, Anselin & Kelejian 1997, Pace et al. 1998, Anselin et al. 2004). Sin embargo, recientemente, los métodos de econometría y social espacial cada vez se han aplicado en una amplia gama de investigaciones empíricas en los campos más tradicionales de la economía, incluyendo entre otros, estudios en análisis de demanda, economía internacional, economía laboral, economía pública, finanzas públicas locales, agricultura y ciencias económicas ambientales.

Estos y otros estudios establecen la importancia de integrar rezagos espaciales y temporales en el análisis económico de datos regionales. Sin embargo, la literatura en modelos con dinámica espacial y temporal solo han presentado algunos progresos, pero en muchos casos por separado. En este capítulo, se

juntan las dos metodologías y se presenta una solución a problemas en donde la variable respuesta es un conteo, una razón o una respuesta binaria (dicotómica) utilizando un refinado modelo lineal generalizado mixto espacial y temporal dinámico basado en distancias. En este último aspecto, se incorporan algunas medidas generales de distancia/disimilitud que se pueden aplicar a variables explicativas: numéricas, categóricas, o una mezcla de ellas. El concepto geométrico de distancia entre individuos o poblaciones puede ser usado, ya que el concepto de distancia es una herramienta útil en pruebas de hipótesis, estimación de parámetros, regresión, etc. (Cuadras & Arenas 1990, Cuadras et al. 1996, Arenas & Cuadras 2002).

Por lo tanto, la propuesta en este capítulo es considerar cómo se pueden utilizar los datos de panel espaciales con variable respuesta no gaussiana para estimar conjuntamente modelos que tienen tanto dinámica espacial como temporal. Estos modelos específicamente se llamarán vectores autorregresivos lineales generalizados espaciales mixtos basados en distancias (distance-based spatial generalised linear mixed vector autoregressions, DBSGLMVARs). Éstos incorporan tanto la dinámica espacial como temporal y difiere de los modelos espaciales porque incorporan la dinámica temporal. Por consiguiente, el DBSGLMVARs contiene dos tipos de dinámicas espaciales. Las variables en el tiempo  $t$  pueden depender de los retrasos espaciales contemporáneos como sucede en los modelos espaciales para datos transversales. Además, las variables en el tiempo  $t$  pueden depender de los rezagos espaciales en el tiempo  $t - l$  ( $l > 0$ ), lo cual se llamará “rezagos espaciales rezagados”. En ausencia de rezago espacial, el DBSGLMVAR es un vector autorregresivo lineal generalizado basado en distancia (distance-based generalised linear mixed vector autoregression, DBGLMVAR), y en la ausencia de rezago temporal, el DBSGLMVAR es idéntico al modelo de panel espacial.

Una pregunta que surge es si en el DBSGLMVAR se pueden identificar todos los parámetros estructurales a estimar. Estos parámetros incluyen los parámetros subyacentes del modelo y los coeficientes de rezago espacial y temporal. En este capítulo se diferencia entre los DBSGLMVARs con y sin autocorrelación espacial (spatial autocorrelation, SAC) en los residuales. Además también se compara el DBSGLMVAR en el que no hay rezago espacial pero donde los residuales están espacialmente correlacionados, con el DBSGLMVAR en el que hay rezago espacial pero los residuales están espacialmente no correlacionados. El primero está anidado en el último y una prueba de factor común se puede utilizar para distinguir empíricamente entre ellos. También se muestra que las respuestas de impulso del DBSGLMVAR con autocorrelación espacial son una simple transformación de las respuestas de impulso donde los impactos regionales se asumen no correlacionados.

Este capítulo está dividido de la siguiente manera: en la Sección 6.2 se desarrolla la metodología propuesta, se presenta el modelo lineal generaliza-

do dinámico espacio-tiempo utilizando el método DB y se presentan algunos ejemplos de casos particulares del modelo propuesto. En la Sección 6.3 se estiman los parámetros asociados al modelo DBSGLMVAR mediante el uso del algoritmo de máxima verosimilitud MCMC. Finalmente, en la Sección 6.4 se presentan algunas medidas de bondad de ajuste y se realiza la predicción espacial de un nuevo individuo.

## 6.2 Modelo lineal generalizado dinámico espacio-tiempo utilizando el método basado en distancia

Sea  $\{y(\mathbf{s}, t), \mathbf{s} \in D, t \in T\}$  un proceso estocástico espacio-temporal, el conjunto de índices  $D$  están en una superficie continua o un conjunto finito de ubicaciones y  $T \subseteq \mathbb{Z}$ . De este modo, el modelo desarrollado es adecuado para tiempo discreto. Una distribución pertenece a la familia exponencial si su función de densidad es de la siguiente forma (McCullagh & Nelder 1989)

$$f(y(\mathbf{s}, t); \alpha_{st}) = h_1(y(\mathbf{s}, t)) \exp\{\eta(\alpha_{st})h_2(y(\mathbf{s}, t) - b(\alpha_{st}))\}$$

donde  $\eta(\alpha_{st})$ ,  $b(\alpha_{st})$ ,  $h_1(y(\mathbf{s}, t))$  y  $h_2(y(\mathbf{s}, t))$  son funciones que toman valores en la recta real.

La propuesta de interpolación está construida para un modelo no gaussiano aleatorio de dinámica temporal y espacial, considerando específicamente variables continuas e indicadoras en el modelo de tendencia. Los datos generados por el mecanismo condicional sobre la señal del modelo siguen un modelo lineal generalizado clásico como el descrito por McCullagh & Nelder (1989). Específicamente, nos enfocamos en modelos de rezago espacial y rezago en el error, en el que se utiliza rezago tanto en el espacio como en el tiempo en la variable dependiente. También, se utiliza la dependencia espacial en ubicaciones vecinas en un diferente período del tiempo. Por lo tanto, el punto de partida es el siguiente modelo

$$\begin{aligned} \eta_{it} = \eta(\mathbf{s}_i, t) = g(\mu_{it}) = \mathbf{v}_i^t \boldsymbol{\gamma} + \sum_{l=1}^{q_1} \varrho_l \eta_{i(t-l)} + \rho_0 \sum_{i'=1}^n w_{ii'}^{(1)} \eta_{i't} \\ + \sum_{l=1}^{q_1} \sum_{i'=1}^n \rho_l w_{ii'}^{(1)} \eta_{i'(t-l)} + \varepsilon_{it} \end{aligned} \quad (6.1)$$

$$\varepsilon_{it} = \varepsilon(\mathbf{s}_i, t) = \sum_{l'=1}^{q_2} \varsigma_{l'} \varepsilon_{i(t-l')} + \varphi_0 \sum_{i'=1}^n w_{ii'}^{(2)} \varepsilon_{i't} + \sum_{l'=1}^{q_2} \sum_{i'=1}^n \varphi_{l'} w_{ii'}^{(2)} \varepsilon_{i'(t-l')} + e_{it} \quad (6.2)$$

con  $i = 1, \dots, n$ ,  $t = 1, \dots, T$ ,  $|\varrho_l| \leq 1$ ,  $|\rho_l| \leq 1$ ,  $|\varsigma_{l'}| \leq 1$  y  $|\varphi_{l'}| \leq 1$ , donde  $\mu_{it} = \mu(\mathbf{s}_i, t) = E[y(\mathbf{s}_i, t) | \mathbf{v}_i, y(\mathbf{s}_i, t-1), \dots, y(\mathbf{s}_i, t-q_1), \varepsilon_{it}]$ ,  $g(\cdot)$

es una función enlace que es invertible y continua,  $\mathbf{v}_i^t \boldsymbol{\gamma}$  es la tendencia,  $\mathbf{v}_i^t = \mathbf{v}^t(\mathbf{s}_i) = (1, v_1(\mathbf{s}_i), \dots, v_p(\mathbf{s}_i))$  es un vector que contiene variables explicativas asociadas a la ubicación espacial  $\mathbf{s}_i$ -ésima y  $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_p)^t$  es un vector de parámetros de regresión espacial desconocidos. Además,  $\varrho_l$  es el coeficiente del  $l$ -ésimo rezago temporal ( $l = 1, \dots, q_1$ ),  $\eta_{i(t-l)}$  es la  $\mathbf{s}_i$ -ésima ubicación en el  $l$ -ésimo rezago temporal aplicado a la función de enlace de las observaciones sobre la variable dependiente,  $\rho_0$  es llamado el coeficiente de rezago espacial,  $\rho_l$  es el coeficiente de “rezago espacial rezagado” en el  $l$ -ésimo rezago ya que puede haber un rezago temporal en el rezago espacial,  $\varepsilon_{it} = \varepsilon_t(\mathbf{s}_i)$  refleja el término de error espacial autocorrelacionado para la  $\mathbf{s}_i$ -ésima ubicación en el  $t$ -ésimo tiempo,  $\varsigma_{l'}$  es el coeficiente de autocorrelación temporal en el  $l'$ -ésimo previo período de tiempo ( $l' = 1, \dots, q_2$ ),  $\varepsilon_{i(t-l')}$  es la  $i$ -ésima ubicación del error espacial en el  $l'$ -ésimo previo período de tiempo,  $\varphi_0$  es llamado el coeficiente SAC,  $\varphi_{l'}$  es llamado el coeficiente rezagado en el  $l'$ -ésimo previo período de tiempo y  $e_{it} = e(\mathbf{s}_i, t)$  es un término de error que es independiente y distribuido idénticamente normal para la  $\mathbf{s}_i$ -ésima ubicación en el  $t$ -ésimo tiempo con media cero y varianza  $\sigma^2$ .

Las ecuaciones (6.1) y (6.2) se pueden reescribir como

$$\varrho(\mathbf{B})\eta_{it} = \mathbf{v}_i^t \boldsymbol{\gamma} + \sum_{j=1}^n w_{ij}^{(1)} \rho^*(\mathbf{B})\eta_{jt} + \varepsilon_{it} \quad (6.3)$$

$$\varsigma(\mathbf{B})\varepsilon_{it} = \sum_{j=1}^n w_{ij}^{(2)} \varphi^*(\mathbf{B})\varepsilon_{it} + e_{it} \quad (6.4)$$

donde  $\varrho(\mathbf{B}) = 1 - \varrho_1 \mathbf{B} - \varrho_2 \mathbf{B}^2 - \dots - \varrho_{q_1} \mathbf{B}^{q_1}$ ,  $\rho^*(\mathbf{B}) = \rho_0 + \rho_1 \mathbf{B} + \dots + \rho_{q_1} \mathbf{B}^{q_1}$ ,  $\varsigma(\mathbf{B}) = 1 - \varsigma_1 \mathbf{B} - \varsigma_2 \mathbf{B}^2 - \dots - \varsigma_{q_2} \mathbf{B}^{q_2}$  y  $\varphi^*(\mathbf{B}) = \varphi_0 + \varphi_1 \mathbf{B} + \dots + \varphi_{q_2} \mathbf{B}^{q_2}$ , con  $\mathbf{B}^l$  un operador de rezago tal que  $\mathbf{B}^l \eta_{it} = \eta_{i(t-l)}$  y  $\mathbf{B}^{l'} \varepsilon_{it} = \varepsilon_{i(t-l')}$ .

Dentro del campo de modelos lineales, es usual trabajar con el modelo en su forma canónica ( $\eta_{it}(\alpha(\mathbf{s}_i, t)) = \alpha(\mathbf{s}_i, t) = \alpha_{it}$ ,  $h_2(y(\mathbf{s}_i, t)) = y(\mathbf{s}_i, t)$ ), que incluye un parámetro de dispersión  $\phi > 0$ . Específicamente, condicionando sobre las variables explicativas ( $\mathbf{v}_{it}$ ), el error espacial no observado  $\varepsilon_{it}$  y los tiempos previos ( $y(\mathbf{s}_i, t-l)$ ,  $l = 1, \dots, q_1$ ),  $y(\mathbf{s}_i, t)$  sigue una distribución de la familia exponencial, es decir,

$$\begin{aligned} & y(\mathbf{s}_i, t) \mid (\mathbf{v}(\mathbf{s}_i, t), y(\mathbf{s}_i, t-1), \dots, y(\mathbf{s}_i, t-q_1), \varepsilon_{it}) \stackrel{ind}{\sim} \\ & f(y(\mathbf{s}_i, t) \mid \mathbf{v}_{it}, y(\mathbf{s}_i, t-1), \dots, y(\mathbf{s}_i, t-q_1), \varepsilon_{it}) \\ & f(y(\mathbf{s}_i, t) \mid \mathbf{v}_{it}, y(\mathbf{s}_i, t-1), \dots, y(\mathbf{s}_i, t-q_1), \varepsilon_{it}) = \\ & \exp \left\{ \frac{1}{\phi} [y(\mathbf{s}_i, t)\alpha_{it} - b(\alpha_{it})] - c(y(\mathbf{s}_i, t), \phi) \right\} \end{aligned} \quad (6.5)$$

donde  $\phi$  es un parámetro de extra-variación y  $c(\cdot)$  es una función específica. La media condicional,  $\mu_{it}$ , se relaciona con  $\alpha_{it}$  a través de la identidad  $\mu_{it} = \frac{\partial b(\alpha_{it})}{\partial \alpha_{it}}$ .

Esta media puede ser modelada, después de una transformación adecuada, mediante el modelo lineal generalizado presentado en las ecuaciones (6.1) y (6.2). La Tabla 5.1 muestra cómo algunas de las más populares distribuciones se pueden reescribir en la forma (6.5).

Los modelos (6.1) y (6.2) se pueden reformular en una forma vectorial compacta como

$$\begin{aligned}\boldsymbol{\eta}_t &= g(\boldsymbol{\mu}_t) = g\{E(\mathbf{y}_t \mid \mathbf{V}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-q_1}, \boldsymbol{\varepsilon}_t)\} \\ &= \mathbf{V}\boldsymbol{\gamma} + \sum_{l=1}^{q_1} \varrho_l \boldsymbol{\eta}_{t-l} + \rho_0 \mathbf{W}_1 \boldsymbol{\eta}_t + \sum_{l=1}^{q_1} \rho_l \mathbf{W}_1 \boldsymbol{\eta}_{t-l} + \boldsymbol{\varepsilon}_t\end{aligned}\quad (6.6)$$

$$\boldsymbol{\varepsilon}_t = \sum_{l'=1}^{q_2} \varsigma_{l'} \boldsymbol{\varepsilon}_{t-l'} + \varphi_0 \mathbf{W}_2 \boldsymbol{\varepsilon}_t + \sum_{l'=1}^{q_2} \varphi_{l'} \mathbf{W}_2 \boldsymbol{\varepsilon}_{t-l'} + \mathbf{e}_t\quad (6.7)$$

donde  $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_n)^t$ ,  $\boldsymbol{\eta}_t = (\eta_{1t}, \dots, \eta_{nt})^t$  es un vector  $n \times 1$ ,  $\boldsymbol{\mu}_t = E(\mathbf{y}_t \mid \mathbf{V}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-q_1})$ ,  $\boldsymbol{\mu}_t = (\mu_{1t}, \dots, \mu_{nt})^t$  es un vector  $n \times 1$ ,  $\mathbf{y}_t = (y(\mathbf{s}_1, t), \dots, y(\mathbf{s}_n, t))^t$  es un vector  $n \times 1$  y  $\mathbf{W}_1$  y  $\mathbf{W}_2$  describen el rezago espacial y el error espacial de las unidades observadas en la muestra, respectivamente.  $\mathbf{V} = \mathbf{H}\mathbf{V}^*$  es una matriz  $n \times (p+1)$  con  $\mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t$  una matriz centrada,  $\mathbf{I}_n$  es la matriz identidad de tamaño  $n \times n$ ,  $\mathbf{1}_n$  es un vector de unos de tamaño  $n \times 1$  y  $\mathbf{V}^*$  es una matriz de variables explicativas originales; observe que  $\mathbf{V}^*$  puede tener en sus componentes variables continuas, categóricas y binarias, o incluso una mezcla de las ellas. Por lo tanto  $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \dots, \varepsilon_{nt})^t$  es un vector  $n \times 1$  y  $\mathbf{e}_t = (e_{1t}, \dots, e_{nt})^t$  es un vector  $n \times 1$ .

Los modelos dados en las ecuaciones (6.6) y (6.7) son los vectores autorregresivos espaciales generalizados (spatial generalised vector autoregressions, SGVARs) utilizando el autorregresivo espacio-tiempo de orden  $q_1$  y perturbaciones autorregresivas espacio-tiempo de orden  $q_2$ , es decir un SGVAR( $q_1, q_2$ ), que incluye regresores exógenos e incorpora dinámicas tanto en lo espacial como en lo temporal. Las interacciones espacio-tiempo son modeladas a través de rezagos de espacio-tiempo y errores espacio-tiempo. Los modelos permiten interacciones espacio-tiempo en la variable dependiente, las variables exógenas y las perturbaciones.

Los modelos (6.3) y (6.4) se pueden expresar como

$$\begin{aligned}\varrho(\mathbf{B})\boldsymbol{\eta}_t &= \mathbf{V}\boldsymbol{\gamma} + \mathbf{W}_1 \rho^*(\mathbf{B})\boldsymbol{\eta}_t + \boldsymbol{\varepsilon}_t \\ \varsigma(\mathbf{B})\boldsymbol{\varepsilon}_t &= \mathbf{W}_2 \varphi^*(\mathbf{B})\boldsymbol{\varepsilon}_t + \mathbf{e}_t\end{aligned}\quad (6.8)$$

los cuales se pueden reescribir como

$$\begin{aligned}[\varrho(\mathbf{B})\mathbf{I}_n - \rho^*(\mathbf{B})\mathbf{W}_1]\boldsymbol{\eta}_t &= \mathbf{V}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t \\ [\varsigma(\mathbf{B})\mathbf{I}_n - \varphi^*(\mathbf{B})\mathbf{W}_2]\boldsymbol{\varepsilon}_t &= \mathbf{e}_t\end{aligned}\quad (6.9)$$

Luego de algunos desarrollos algebraicos, se obtiene

$$[\varrho(\mathbf{B})\mathbf{I}_n - \rho^*(\mathbf{B})\mathbf{W}_1]\boldsymbol{\eta}_t = \mathbf{V}\boldsymbol{\gamma} + [\varsigma(\mathbf{B})\mathbf{I}_n - \varphi^*(\mathbf{B})\mathbf{W}_2]^{-1}\mathbf{e}_t \quad (6.10)$$

Esta forma del modelo es válida siempre que  $[\varsigma(\mathbf{B})\mathbf{I}_n - \varphi^*(\mathbf{B})\mathbf{W}_2]$  sea una matriz no singular.

### 6.2.1 Vectores autorregresivos generalizados espaciales basados en distancias

Para construir el modelo propuesto, se utilizan las medidas de distancias y disimilaridad presentadas en la Sección 3.2.1 para variables explicativas continuas, categóricas, o una mezcla de ellas. Una vez seleccionada algunas de las distancias o disimilaridades presentadas en dicha sección se realiza la descomposición espectral obteniendo  $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^t = \mathbf{X}\mathbf{X}^t$ , donde  $\mathbf{A} = (-d_{ii}^2/2)$ ,  $\boldsymbol{\Lambda}$  es una matriz diagonal que contiene los valores propios de  $\mathbf{B}$  y  $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}$  es una matriz  $n \times n$  de coordenadas principales y de rango  $n - 1$  porque tiene un vector propio igual a  $\mathbf{1}_n$ , y  $\mathbf{U}$  contiene las coordenadas principales estandarizadas.

Como ya se ha mencionado anteriormente, uno de los peligros potenciales en el modelo basado en distancias es la enorme sobreparametrización, ya que el rango de  $\mathbf{B}$  puede ser tan grande como  $n - 1$ . Entonces, el número de coordenadas principales (columnas de  $\mathbf{X}$ ) puede ser excesivo, lo que permite un modelo arbitrariamente sobre-ajustado. Con el fin de evitar este tipo de problemas, se deben seleccionar únicamente las coordenadas principales más significativas utilizando cualquier método de la Sección 6.4.2. Por lo tanto, en forma matricial, el modelo DBSGLMVAR en forma reducida se puede escribir como

$$\begin{aligned} \boldsymbol{\eta}_t &= g(\boldsymbol{\mu}_t) = g\{\mathbf{E}(\mathbf{y}_t \mid \mathbf{X}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-q_1}, \boldsymbol{\varepsilon}_t)\} \\ &= \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{q_1} \varrho_l \boldsymbol{\eta}_{t-l} + \rho_0 \mathbf{W}_1 \boldsymbol{\eta}_t + \sum_{l=1}^{q_1} \rho_l \mathbf{W}_1 \boldsymbol{\eta}_{t-l} + \boldsymbol{\varepsilon}_t \end{aligned} \quad (6.11)$$

$$\boldsymbol{\varepsilon}_t = \sum_{l'=1}^{q_2} \varsigma_{l'} \boldsymbol{\varepsilon}_{t-l'} + \varphi_0 \mathbf{W}_2 \boldsymbol{\varepsilon}_t + \sum_{l'=1}^{q_2} \varphi_{l'} \mathbf{W}_2 \boldsymbol{\varepsilon}_{t-l'} + \mathbf{e}_t \quad (6.12)$$

donde  $\boldsymbol{\mu}_t = \mathbf{E}(\mathbf{y}_t \mid \mathbf{X}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-q_1}, \boldsymbol{\varepsilon}_t)$ ,  $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \dots, \beta_k)$  es un vector de parámetros desconocido,  $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ ,  $k \leq n - 1$ ,  $\mathbf{X}(k) = (\mathbf{1}_n, \mathbf{X}_1, \dots, \mathbf{X}_k)$  contiene un subconjunto de las  $k$  columnas más relevantes de  $\mathbf{X}$  que son las coordenadas significativamente correlacionadas con  $\boldsymbol{\eta}_t$ .

Los modelos presentados en las ecuaciones (6.11) y (6.12) son los modelos DBSGLMVARs utilizando el autorregresivo espacio-tiempo de orden  $q_1$

y la perturbación autorregresiva espacio-tiempo de orden  $q_2$ , es decir, se obtiene un DBSGLMVAR( $q_1, q_2$ ), que incluye regresores exógenos e incorpora tanto dinámicas espaciales como temporales. Las interacciones espacio-tiempo son modeladas a través de rezagos espacio-tiempo y errores espacio-tiempo. Además, los modelos tienen en cuenta las interacciones espacio-tiempo en la variable dependiente, las variables exógenas y las perturbaciones.

Los modelos alternativos (6.11) y (6.12) se pueden expresar también como

$$\begin{aligned}\eta_{it} &= \mathbf{x}_i^t \boldsymbol{\beta} + \sum_{l=1}^{q_1} \varrho_l \eta_{i(t-l)} + \rho_0 \sum_{i'=1}^n w_{ii'}^{(1)} \eta_{i't} + \sum_{l=1}^{q_1} \sum_{i'=1}^n \rho_l w_{ii'}^{(1)} \eta_{i'(t-l)} + \varepsilon_{it} \\ \varepsilon_{it} &= \sum_{l'=1}^{q_2} \varsigma_{l'} \varepsilon_{i(t-l')} + \varphi_0 \sum_{i'=1}^n w_{ii'}^{(2)} \varepsilon_{i't} + \sum_{l'=1}^{q_2} \sum_{i'=1}^n \varphi_{l'} w_{ii'}^{(2)} \varepsilon_{i'(t-l')} + e_{it}\end{aligned}$$

con  $i = 1, \dots, n$ , y donde  $\mathbf{x}_i^t = (1, x_{i0}, \dots, x_{ik})$ .

Los modelos (6.11) y (6.12) son también equivalentes a

$$\begin{aligned}[\varrho(\mathbf{B})\mathbf{I}_n - \rho^*(\mathbf{B})\mathbf{W}_1] \boldsymbol{\eta}_t &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \\ [\varsigma(\mathbf{B})\mathbf{I}_n - \varphi^*(\mathbf{B})\mathbf{W}_2] \boldsymbol{\varepsilon}_t &= \mathbf{e}_t\end{aligned}$$

Luego haciendo algunos desarrollos algebraicos, se llega a

$$\begin{aligned}[\varrho(\mathbf{B})\mathbf{I}_n - \rho^*(\mathbf{B})\mathbf{W}_1] \boldsymbol{\eta}_t &= \mathbf{X}\boldsymbol{\beta} + [\varsigma(\mathbf{B})\mathbf{I}_n - \varphi^*(\mathbf{B})\mathbf{W}_2]^{-1} \mathbf{e}_t \\ \boldsymbol{\eta}_t &= \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{q_1} \varrho_l \boldsymbol{\eta}_{t-l} + \rho_0 \mathbf{W}_1 \boldsymbol{\eta}_t + \sum_{l=1}^{q_1} \rho_l \mathbf{W}_1 \boldsymbol{\eta}_{t-l} + \boldsymbol{\varepsilon}_t \\ &= \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}_t\end{aligned}\tag{6.13}$$

donde  $\mathbf{X} = (\mathbf{X}, \boldsymbol{\eta}_{t-1}, \dots, \boldsymbol{\eta}_{t-q_1}, \mathbf{W}_1 \boldsymbol{\eta}_t, \mathbf{W}_1 \boldsymbol{\eta}_{t-1}, \dots, \mathbf{W}_1 \boldsymbol{\eta}_{t-q_1})$ ,  $\boldsymbol{\alpha} = (\boldsymbol{\beta}^t, \varrho_1, \dots, \varrho_{q_1}, \rho_0, \rho_1, \dots, \rho_{q_1})^t$  y  $\boldsymbol{\varepsilon}_t \sim MN(\mathbf{0}, \mathbf{Q}_\varepsilon \mathbf{Q}_\varepsilon^t \sigma^2)$  con  $\mathbf{Q}_\varepsilon = [\varsigma(\mathbf{B})\mathbf{I}_n - \varphi^*(\mathbf{B})\mathbf{W}_2]^{-1}$ .

A continuación se presentan algunos ejemplos particulares de los modelos DBSGLMVARs.

**Ejemplo 6.1.** *Modelo de vectores autorregresivos generalizados con perturbaciones basados en distancias (distance-based generalised linear mixed vector autoregressive disturbances, DBGLMVAR) es un modelo DBSGLMVAR con los parámetros  $\rho_0 = \rho_1 = \dots = \rho_{q_1} = 0$  y  $\varphi_0 = \varphi_1 = \dots = \varphi_{q_2} = 0$  en los modelos (6.11) y (6.12), respectivamente, es decir, se obtiene un modelo DBGLMVAR( $q_1, q_2$ ) dado por*

$$\boldsymbol{\eta}_t = \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{q_1} \varrho_l \boldsymbol{\eta}_{t-l} + \boldsymbol{\varepsilon}_t\tag{6.14}$$

$$\boldsymbol{\varepsilon}_t = \sum_{l'=1}^{q_2} \varsigma_{l'} \boldsymbol{\varepsilon}_{t-l'} + \mathbf{e}_t\tag{6.15}$$

donde  $\mathbf{e}_t$  es un vector de error distribuido normal con media cero y matriz de varianza y covarianza  $\sigma^2 \mathbf{I}_n$ . Observe que  $\boldsymbol{\eta}_t$  y  $\boldsymbol{\varepsilon}_t$  no son independientes.

Los modelos (6.14) y (6.15) se pueden expresar como

$$\boldsymbol{\eta}_t = \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{q_1} \varrho_l \boldsymbol{\eta}_{t-l} + \varsigma^{-1}(\mathbf{B})\mathbf{e}_t$$

**Ejemplo 6.2.** Modelo transversal espacial generalizado basado en distancias (*distance-based spatial generalised cross sectional, DBGCS*) es un modelo DBSGLMVAR con parámetros  $\varrho_1 = \dots = \varrho_{q_1} = 0$ ,  $\rho_1 = \dots = \rho_{q_1} = 0$ ,  $\varsigma_1 = \dots = \varsigma_{q_2} = 0$  y  $\varphi_1 = \dots = \varphi_{q_2} = 0$  en los modelos (6.11) y (6.12), es decir, se obtiene un modelo DBGLMVAR(0,0) dado por

$$\boldsymbol{\eta}_t = \mathbf{X}\boldsymbol{\beta} + \rho_0 \mathbf{W}_1 \boldsymbol{\eta}_t + \boldsymbol{\varepsilon}_t \quad (6.16)$$

$$\boldsymbol{\varepsilon}_t = \varphi_0 \mathbf{W}_2 \boldsymbol{\varepsilon}_t + \mathbf{e}_t \quad (6.17)$$

o equivalentemente

$$\begin{aligned} \boldsymbol{\eta}_t &= (\mathbf{I}_n - \rho_0 \mathbf{W}_1)^{-1} \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \rho_0 \mathbf{W}_1)^{-1} (\mathbf{I}_n - \varphi_0 \mathbf{W}_2)^{-1} \mathbf{e}_t \\ &= \mathbf{Q}_{w_1} (\mathbf{X}\boldsymbol{\beta} + \mathbf{Q}_{w_2} \mathbf{e}_t) = \mathbf{X}^* \boldsymbol{\beta} + \mathbf{e}_t^* \end{aligned} \quad (6.18)$$

donde  $\mathbf{Q}_{w_1} = (\mathbf{I}_n - \rho_0 \mathbf{W}_1)^{-1}$ ,  $\mathbf{Q}_{w_2} = (\mathbf{I}_n - \varphi_0 \mathbf{W}_2)^{-1}$ ,  $\mathbf{X}^* = \mathbf{Q}_{w_1} \mathbf{X}$ ,  $\mathbf{e}_t^*$  es un vector de error distribuido normal con vector de media cero y matriz de varianza y covarianza  $(\mathbf{Q}_{w_1} \mathbf{Q}_{w_2}) (\mathbf{Q}_{w_1} \mathbf{Q}_{w_2})^t \sigma^2$ . En este modelo, como los datos son transversales hay únicamente un período de tiempo.

La identificación de la estructura de los parámetros en los modelos (6.16) y (6.17) requiere que las coordenadas principales ( $\mathbf{X}_j$ ,  $j = 1, \dots, k$ ) no sean perfectamente colineales espacialmente, lo cual es válido en este caso ya que son coordenadas principales, y  $w_{ii'} < 1$  (Manski 1993),  $i, i' = 1, \dots, n$ . Si no hay efecto de las coordenadas principales ( $\mathbf{X}$ ), los coeficientes de rezago espacial no se pueden identificar. Por lo tanto, los coeficientes de rezago espacial no son estimables en ausencia de variables exógenas.

En las dos partes del modelo presentado en la ecuación (6.18), el rezago espacial y el error espacial, se requiere estacionariedad, ésta se cumple si  $1/\varpi_{\min} < \rho_0 < 1/\varpi_{\max}$  y  $1/\varpi_{\min}^* < \varphi_0 < 1/\varpi_{\max}^*$ , donde  $\varpi_{\min}$  y  $\varpi_{\max}$  representan los valores propios más pequeño y más grande de la matriz  $\mathbf{W}_1$ , y  $\varpi_{\min}^*$  y  $\varpi_{\max}^*$  representan los valores propios más pequeño y más grande de la matriz  $\mathbf{W}_2$ . Si bien a menudo se sugiere en la literatura para restringir  $\rho_0$  o  $\varphi_0$  al intervalo  $(-1, +1)$ , esto es una restricción a veces innecesaria. Para los pesos espaciales en filas normalizadas, el valor propio más grande es  $+1$ , pero en general los resultados no son válidos para el valor propio más pequeño, donde el límite inferior es típicamente menor de  $-1$ . Como una alternativa a las filas normalizadas, se pueden normalizar  $\mathbf{W}_1$  y  $\mathbf{W}_2$  por columnas de tal



forma que estas sume uno. Estos tipos de normalización se utilizan algunas veces en áreas de la economía y biología, entre otras (Fisher & Getis 2010).

La ecuación (6.18) resuelve el problema para el perfil de respuesta al impulso espacial, que muestra el efecto de  $\mathbf{X}$  en la región  $i$  sobre  $\boldsymbol{\eta}_t$  en la región  $i'$  ( $q_{ii'}$ ). En la ausencia de rezago espacial  $q_{ii'} = 0$  cuando  $i \neq i'$ . En ese caso, los cambios de  $\mathbf{X}$  en la región  $i$  no se propagan más allá de la región  $i$ .

El modelo (6.18) se puede reescribir también como

$$(\mathbf{I}_n - \varphi_0 \mathbf{W}_2) (\mathbf{I}_n - \rho_0 \mathbf{W}_1) \boldsymbol{\eta}_t = (\mathbf{I}_n - \varphi_0 \mathbf{W}_2) \mathbf{X} \boldsymbol{\beta} + \mathbf{e}_t$$

y después de hacer algunos procedimientos algebraicos, se llega a

$$\boldsymbol{\eta}_t = \mathbf{X} \boldsymbol{\alpha}_1 + \mathbf{W}_2 \mathbf{X} \boldsymbol{\alpha}_2 + \alpha_3 \mathbf{W}_2 \boldsymbol{\eta}_t + \alpha_4 \mathbf{W}_1 \boldsymbol{\eta}_t + \alpha_5 \mathbf{W}_2 \mathbf{W}_1 \boldsymbol{\eta}_t + \mathbf{e}_t \quad (6.19)$$

Las restricciones sobre los parámetros en el modelo (6.19) son:  $\boldsymbol{\alpha}_1 = \boldsymbol{\beta}$ ,  $\boldsymbol{\alpha}_2 = -\varphi_0 \boldsymbol{\beta}$ ,  $\alpha_3 = \varphi_0$ ,  $\alpha_4 = \rho_0$  y  $\alpha_5 = -\varphi_0 \rho_0$ .

Los modelos espacialmente estadísticos con errores espacialmente autocorrelacionados son versiones restringidas de modelos espacialmente dinámicos con errores espacialmente independientes. Esto se demuestra tomando el siguiente modelo estadístico espacial

$$\begin{aligned} \boldsymbol{\eta}_t &= \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\varepsilon}_t &= \varphi_0 \mathbf{W}_2 \boldsymbol{\varepsilon}_t + \mathbf{e}_t \end{aligned}$$

que se puede escribir como

$$(\mathbf{I}_n - \varphi_0 \mathbf{W}_2) \boldsymbol{\eta}_t = (\mathbf{I}_n - \varphi_0 \mathbf{W}_2) \mathbf{X} \boldsymbol{\beta} + \mathbf{e}_t \quad (6.20)$$

El modelo espacialmente dinámico es

$$(\mathbf{I}_n - \rho_0 \mathbf{W}_1) \boldsymbol{\eta}_t = \mathbf{X} \boldsymbol{\beta} + \mathbf{e}_t \quad (6.21)$$

La ecuación (6.20) contiene la restricción de factor común  $(\mathbf{I}_n - \varphi_0 \mathbf{W}_2)$ , mientras que la ecuación (6.21) no. Por lo tanto, la ecuación (6.21) es una versión restringida de la ecuación (6.20). Esta restricción de factor común se puede contrastar utilizando una prueba de factor común (Anselin 1988, pp.226-229), pero en este caso, utilizando una adaptación para modelos generalizados.

**Ejemplo 6.3.** Los datos panel espaciales generalizados basados en distancias son un modelo DBSGLMVAR( $q_1, q_2$ ). Los clásicos datos panel se han discutido por algunos autores (Anselin 1988, chapter 10), (Elhorst 2003) y (Lee 2004) en un contexto temporal estático, es decir, modelos en el cual existe dinámicas espaciales pero no dinámicas temporales. Introducir dinámicas espaciales dentro de modelos de datos panel temporalmente estáticos con variable respuesta

no gaussiana es una complicación que no se hace; sin embargo, este problema si está presente altera sustancialmente la teoría de datos panel econométricos.

A modo de introducción se considera el siguiente modelo generalizado con rezagos autorregresivos temporal y espacial de primer orden utilizando perturbaciones autorregresivas espacio-temporales de primer orden, es decir se considera un modelo DBSGLMVAR(1,1). Este modelo está dado por

$$\boldsymbol{\eta}_t = \mathbf{X}\boldsymbol{\beta} + \varrho_1\boldsymbol{\eta}_{t-1} + \rho_0\mathbf{W}_1\boldsymbol{\eta}_t + \rho_1\mathbf{W}_1\boldsymbol{\eta}_{t-1} + \boldsymbol{\varepsilon}_t \quad (6.22)$$

$$\boldsymbol{\varepsilon}_t = \varsigma_1\boldsymbol{\varepsilon}_{t-1} + \varphi_0\mathbf{W}_2\boldsymbol{\varepsilon}_t + \varphi_1\mathbf{W}_2\boldsymbol{\varepsilon}_{t-1} + \mathbf{e}_t \quad (6.23)$$

La identificación de los parámetros en la ecuación (6.27), incluyendo el coeficiente de rezago espacial  $\rho_0$ , requiere que  $\boldsymbol{\eta}_{t-1}$  y  $\mathbf{W}_1\boldsymbol{\eta}_{t-1}$  sean débilmente exógenas. Si éstas no lo son, esas variables no son independientes de  $\boldsymbol{\varepsilon}_t$ . Es fácil demostrar que esas variables son débilmente exógenas cuando  $\varsigma_1 = \varphi_1 = 0$ . Si  $\varsigma_1 \neq 0$ ,  $\boldsymbol{\varepsilon}_{t-1}$  afecta tanto a  $\boldsymbol{\varepsilon}_t$  como a  $\boldsymbol{\eta}_{t-1}$ , en cuyo caso  $\boldsymbol{\eta}_{t-1}$  y  $\boldsymbol{\varepsilon}_t$  no son independientes. Si  $\varphi_1 \neq 0$ ,  $\mathbf{W}_2\boldsymbol{\varepsilon}_{t-1}$  afecta tanto a  $\boldsymbol{\varepsilon}_t$  como a  $\mathbf{W}_1\boldsymbol{\eta}_{t-1}$ , en cuyo caso  $\mathbf{W}_1\boldsymbol{\eta}_{t-1}$  y  $\boldsymbol{\varepsilon}_t$  no son independientes. En resumen, la autocorrelación temporal y/o el SAC rezagado significa que los parámetros del modelo no pueden ser identificados o estimables.

**Ejemplo 6.4.** Así como los DBGLMVARs son utilizados para simular los efectos dinámicos de cambios exógenos sobre las variables de estado, los DBSGLMVARs pueden ser utilizados para simular los efectos dinámicos espacio-temporales de cambios exógenos. El análisis de la respuesta al impulso en los DBSGLMVARs es inevitablemente más compleja que en los DBGLMVARs y modelos espaciales porque los cambios se propagan a través del espacio como también sobre el tiempo. Considerando el modelo (6.11) sin coordenadas principales y el modelo (6.12), en el cual los cambios no están autocorrelacionados espacialmente, se obtiene el siguiente modelo DBSGLMVAR

$$\boldsymbol{\eta}_t = \varrho_1\boldsymbol{\eta}_{t-1} + \rho_0\mathbf{W}_1\boldsymbol{\eta}_t + \rho_1\mathbf{W}_1\boldsymbol{\eta}_{t-1} + \mathbf{e}_t \quad (6.24)$$

Los rezagos espaciales son expresados una vez más por los escalares  $\rho_0$  y  $\rho_1$ . Si estos parámetros son cero, la ecuación (6.24) es un proceso autorregresivo de orden uno. La ecuación (6.24) se puede escribir utilizando el operador de retraso temporal ( $\mathbf{B}$ ) como

$$[(1 - \varrho_1\mathbf{B})\mathbf{I}_T - (\rho_0 + \rho_1\mathbf{B})\mathbf{W}_1]\boldsymbol{\eta}_t = \mathbf{e}_t \quad (6.25)$$

Los perfiles de respuesta al impulso son obtenidos derivando la representación de Wold de la ecuación (6.25), es decir, expresando  $\boldsymbol{\eta}_t$  en términos de los valores actuales y rezagados de  $\mathbf{e}_t$ . Éste se obtiene premultiplicando a ambos lados de la ecuación (6.25) por  $\mathbf{Q}_1^{-1} = [(1 - \varrho_1\mathbf{B})\mathbf{I}_T + (\rho_0 + \rho_1\mathbf{B})\mathbf{W}_1]^{-1}$ , en

cuyo caso la solución para  $\boldsymbol{\eta}_t$  es

$$\boldsymbol{\eta}_t = \mathbf{Q}_1^{-1} \mathbf{e}_t + \sum_{i=1}^n \mathbf{a}_i l_i$$

donde los valores propios se denotan por  $l_i$  y los  $\mathbf{a}_i$ 's son vectores de constantes arbitrarios determinados por las condiciones iniciales. Siempre que los datos sean estacionarios,  $|l_i| < 1$ ; en cuyo caso el término de la sumatoria tiende a cero con el tiempo. Los  $n$  vectores y valores propios se obtienen de la solución a

$$|\mathbf{Q}_1^{-1} - l\mathbf{I}_n| = 0$$

Como  $\mathbf{Q}_1$  depende de  $\varrho_1$  y  $\rho_1$ , es inevitable que los valores propios del DBSGLMVAR dependan de los coeficientes de rezago espacial. Esto también significa que las condiciones de estacionariedad para los DBGLMVARs son diferentes de sus contrapartes en DBSGLMVARs.

Para ilustrar lo anterior, sea  $n = 2$ ,  $q_1 = 1$  y  $w_{12}^{(1)} = w_{21}^{(1)} = 1$ , en cuyo caso el modelo estructural es

$$\begin{aligned}\eta_{1t} &= \varrho_1 \eta_{1(t-1)} + \rho_0 \eta_{2t} + \rho_1 \eta_{2(t-1)} + e_{1t} \\ \eta_{2t} &= \varrho_1 \eta_{2(t-1)} + \rho_0 \eta_{1t} + \rho_1 \eta_{1(t-1)} + e_{2t}\end{aligned}$$

La ecuación característica es

$$(1 - \rho_0^2)l^2 - 2(\varrho_1 + \rho_0\rho_1)l + (\varrho_1^2 - \rho_1^2) \quad (6.26)$$

La ecuación (6.26) tiene dos valores propios  $l_1$  y  $l_2$ , dados por

$$l_1 = \frac{\varrho_1 - \rho_1}{1 + \rho_0} \quad y \quad l_2 = \frac{\varrho_1 + \rho_1}{1 - \rho_0}$$

La estacionariedad requiere que estas dos raíces sean menores a la unidad en valor absoluto. Es obvio que la estacionariedad no depende simplemente de  $\varrho_1$  como sucede en el caso de ausencia de efectos espaciales. En efecto, el valor absoluto de  $\varrho_1$  puede ser menor de la unidad, pero  $\boldsymbol{\eta}_t$  puede sin embargo ser no estacionario. Los siguientes resultados son fácilmente establecidos

- (a) Si  $\varrho_1 = 1$  no hay valores de  $\rho_0$  y  $\rho_1$  que inducen estacionariedad. Por lo tanto, si una variable es temporalmente no estacionaria lo sigue siendo cuando las dinámicas espaciales están presentes.
- (b) Si  $\rho_0 = 0$  los valores propios son  $l = \varrho_1 \pm \rho_1$ . Por lo tanto, si  $\rho_1 = 1$ , la variable debe ser no estacionaria independientemente de  $\varrho_1$ .
- (c) Si  $\varrho_1 = 0$ , los valores propios son  $l = \rho_1(\rho_0 \pm 1)/(1 - \rho_0^2)$ , en cuyo caso la variable puede ser no estacionaria.

(d) Si  $\rho_0 = 1$  y  $\rho_1 = 0$ , hay un sólo valor propio con  $l = 3/2\rho_1$ . En este caso especial, la variable es estacionaria cuando  $\rho_1 < 2/3$ .

Asumiendo estacionariedad, la solución general para  $\eta_{1t}$  es

$$\eta_{1t} = \frac{e_{1t} - \rho_1 e_{1(t-1)} + (\rho_0 e_{2t} + \rho_1) e_{2(t-2)}}{(1 - l_1 \mathbf{B})(1 - l_2 \mathbf{B})} + a_1 l_1 + a_2 l_2 \quad (6.27)$$

donde los  $a_i$ 's se determinan por condiciones iniciales. Ya que las raíces se encuentran en el círculo unitario, estos términos tienden a cero con el tiempo. Invirtiendo los rezagos polinomiales por fracciones parciales en la ecuación (6.27) se obtiene la relación entre  $\eta_{1t}$  y los  $e$ 's actuales y rezagados

$$\begin{aligned} \eta_{1t} = & \frac{1}{l_1 - l_2} \sum_{j=0}^{\infty} [l_1^{1+j} (e_{1(t-j)} - \rho_1 e_{1(t-j-1)}) - l_2^{1+j} (\rho_0 e_{2(t-j)} + \rho_1 e_{2(t-j-1)})] \\ & + c_1 l_1 + c_2 l_2 \end{aligned} \quad (6.28)$$

donde los  $c_i$ 's son constantes arbitrarias determinadas por las condiciones iniciales. De acuerdo a la ecuación (6.28), los cambios actuales y rezagos en la región 2 repercutirán en la región 1. Sin embargo, si no hay dinámicas espaciales ( $\rho_0 = \rho_1 = 0$ ), la ecuación (6.28) se simplifica a

$$\eta_{1t} = \sum_{j=0}^{\infty} \rho_1^j e_{1(t-j)} + c_1 l_1$$

**Ejemplo 6.5.** Considere una relación donde el vector de la variable dependiente transformado en el tiempo  $t$ , denotado por  $\boldsymbol{\eta}_t$ , se determina utilizando un esquema autorregresivo espacial que depende del espacio-tiempo de los valores rezagados de la variable dependiente de observaciones vecinas. Esto lleva a un rezago en la media de los vecinos de la variable dependiente observada durante los períodos previos,  $\mathbf{W}_1 \boldsymbol{\eta}_{t-1}$ . En esta situación, se puede también incluir características de las propias regiones ( $\mathbf{X}$ ) usando DB. Esto sugiere la siguiente relación como una representación para procesos DBSGLMVAR

$$\boldsymbol{\eta}_t = \mathbf{X}\boldsymbol{\beta} + \rho_1 \mathbf{W}_1 \boldsymbol{\eta}_{t-1} + \boldsymbol{\varepsilon}_t \quad (6.29)$$

Observe que se puede reemplazar  $\boldsymbol{\eta}_{t-1}$  en el lado derecho de (6.29) por  $\boldsymbol{\eta}_{t-1} = \mathbf{X}\boldsymbol{\beta} + \rho_1 \mathbf{W}_1 \boldsymbol{\eta}_{t-2} + \boldsymbol{\varepsilon}_{t-1}$ . Por lo tanto, el modelo (6.29) se puede re-escribir como

$$\begin{aligned} \boldsymbol{\eta}_t = & \mathbf{X}\boldsymbol{\beta} + \rho_1 \mathbf{W}_1 (\mathbf{X}\boldsymbol{\beta} + \rho_1 \mathbf{W}_1 \boldsymbol{\eta}_{t-2} + \boldsymbol{\varepsilon}_{t-1}) + \boldsymbol{\varepsilon}_t \\ = & \mathbf{X}\boldsymbol{\beta} + \rho_1 \mathbf{W}_1 \mathbf{X}\boldsymbol{\beta} + \rho_1^2 \mathbf{W}_1^2 \boldsymbol{\eta}_{t-2} + \boldsymbol{\varepsilon}_t + \rho_1 \mathbf{W}_1 \boldsymbol{\varepsilon}_{t-1} \end{aligned} \quad (6.30)$$

Sustituyendo recursivamente por los valores del pasado del vector  $\boldsymbol{\eta}_{t-1}$  en el lado derecho de (6.30) sobre  $q_1$  períodos, se llega a

$$\begin{aligned} \boldsymbol{\eta}_t = & (\mathbf{I}_n + \rho_1 \mathbf{W}_1 + \cdots + \rho_1^{q_1-1} \mathbf{W}_1^{q_1-1}) \mathbf{X}\boldsymbol{\beta} + \rho_1^{q_1} \mathbf{W}_1^{q_1} \boldsymbol{\eta}_{t-q_1} + u_t \\ u_t = & \boldsymbol{\varepsilon}_t + \rho_1 \mathbf{W}_1 \boldsymbol{\varepsilon}_{t-1} + \cdots + \rho_1^{q_1-1} \mathbf{W}_1^{q_1-1} \boldsymbol{\varepsilon}_{t-(q_1-1)} \end{aligned}$$

Estas expresiones se pueden simplificar observando que  $E(\boldsymbol{\varepsilon}_{t-l}) = \mathbf{0}$ ,  $l = 0, \dots, q_1 - 1$ , lo cual implica que  $E(u_t) = \mathbf{0}$ . Además, la magnitud de  $\rho_1^{q_1} \mathbf{W}_1^{q_1} \boldsymbol{\eta}_{t-q_1}$  se vuelve pequeño para  $q_1$  grande, bajo el supuesto usual que  $|\rho_1| < 1$  y suponiendo que  $\mathbf{W}_1$  es por filas estocástica, así la matriz  $\mathbf{W}_1$  tiene un valor propio de 1. Por consiguiente, se puede interpretar la relación transversal observada como el resultado o la expectativa de un equilibrio a largo plazo como se muestra en la siguiente ecuación

$$\lim_{q_1 \rightarrow \infty} E(\boldsymbol{\eta}_t) = (\mathbf{I}_n - \rho_1 \mathbf{W}_1)^{-1} \mathbf{X} \boldsymbol{\beta}$$

donde  $(\mathbf{I}_n - \rho_1 \mathbf{W}_1)^{-1} = \mathbf{I}_n + \rho_1 \mathbf{W}_1 + \rho_1^2 \mathbf{W}_1^2 + \dots + \rho_1^{q_1-1} \mathbf{W}_1^{q_1-1} + \rho_1^{q_1} \mathbf{W}_1^{q_1} + \dots$ , con  $|\rho_1| < 1$ .

Observe que lo anterior provee una motivación dinámica para el proceso de datos generados del modelo transversal DBSGLMVAR que sirve como un caballo de batalla de los modelos de regresión espacial. Esto es, la relación del modelo transversal DBSGLMVAR puede surgir de la dependencia-tiempo de decisiones de los agentes económicos situados en varios puntos en el espacio cuando las decisiones dependen de estos vecinos.

### 6.3 Algoritmo de máxima verosimilitud vía Monte Carlo para DBSGLMVAR

La aplicación de métodos basados en verosimilitud a modelos de vectores autorregresivos generalizados espaciales no gaussianos basados en distancia se ve obstaculizada por las dificultades computacionales que surgen de la gran dimensionalidad del vector aleatorio no observado,  $\boldsymbol{\varepsilon}_t$ , en el modelo presentado en la ecuación (6.13). En esta sección se considera el método MCMC vía máxima verosimilitud (Geyer & Thompson 1992, Geyer 1994, Højbjerg 2003, Christensen 2004) para modelos de vector autorregresivos generalizados espaciales.

Asumiendo que cada  $Y(\mathbf{s}_i, t)$  en el modelo (6.13) tiene una distribución perteneciente a la familia exponencial, y por independencia de  $Y(\mathbf{s}_1, t), \dots, Y(\mathbf{s}_n, t)$  dadas las covariables observadas  $\mathbf{X}$ ,  $\boldsymbol{\varepsilon}_t$  y  $g^{-1}(\cdot)$ , la función de densidad condicional de  $\mathbf{Y}_s = \mathbf{y}_s$  dadas  $\mathbf{X}$  y  $\boldsymbol{\varepsilon}_t$  es

$$f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}) = \prod_{i=1}^n f\{y(\mathbf{s}_i, t); g^{-1}[\mathbf{x}(\mathbf{s}_i, t), y(\mathbf{s}_i, t), \dots, y(\mathbf{s}_i, t - q_1), \boldsymbol{\varepsilon}(\mathbf{s}_i, t); \boldsymbol{\alpha}]\}$$

Desde una perspectiva clásica, la función de verosimilitud basada en las variables aleatorias observadas  $\mathbf{y}_t$  se obtiene marginalizando con respecto a las variables no observadas  $\boldsymbol{\varepsilon}_t$ , dando lugar a la verosimilitud del modelo mixto.

Entonces la función verosimilitud para un modelo de vector autorregresivo generalizado mixto espacial no se puede escribir en forma cerrada sino sólo como una integral de alta dimensión

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\theta}) &= f(\mathbf{y}_t | \boldsymbol{\alpha}, \boldsymbol{\theta}) = \int_{\mathbb{R}^n} f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}) f(\boldsymbol{\varepsilon}_t | \boldsymbol{\theta}) d\boldsymbol{\varepsilon}_t \\ &= \int_{\mathbb{R}^n} \prod_{i=1}^n f\{y(\mathbf{s}_i, t); g^{-1}[\mathbf{x}(\mathbf{s}_i, t), y(\mathbf{s}_i, t), \dots, y(\mathbf{s}_i, t - q_1), \boldsymbol{\varepsilon}(\mathbf{s}_i, t); \boldsymbol{\alpha}]\} \\ &\quad \times f(\boldsymbol{\varepsilon}_t | \boldsymbol{\theta}) d\varepsilon_{1t} \cdots d\varepsilon_{nt} \end{aligned} \quad (6.31)$$

donde  $f(\boldsymbol{\varepsilon}_t | \boldsymbol{\theta})$  denota la función de distribución normal multivariada de  $\boldsymbol{\varepsilon}_t$  dadas las covariables observadas  $\mathbf{X}$ , con  $\boldsymbol{\theta} = (\varsigma_1, \dots, \varsigma_{q_2}, \varphi_0, \dots, \varphi_{q_2}, \sigma^2)$  el vector de parámetros asociados a  $\boldsymbol{\varepsilon}_t$ . La integral anterior es también la constante de normalización de la función de densidad condicional de  $\boldsymbol{\varepsilon}_t$  dado  $\mathbf{y}_t$ ,

$$\begin{aligned} f(\boldsymbol{\varepsilon}_t | \mathbf{y}_t, \boldsymbol{\alpha}, \boldsymbol{\theta}) &\propto \prod_{i=1}^n f\{y(\mathbf{s}_i, t); g^{-1}[\mathbf{x}(\mathbf{s}_i, t), y(\mathbf{s}_i, t), \dots, y(\mathbf{s}_i, t - q_1), \boldsymbol{\varepsilon}(\mathbf{s}_i, t); \boldsymbol{\alpha}]\} \\ &\quad \times f(\boldsymbol{\varepsilon}_t | \boldsymbol{\theta}) \end{aligned} \quad (6.32)$$

La MCMC provee un método para la simulación de (6.32) y una aproximación de (6.31).

La integral tiene una alta dimensión, y consecuentemente, es intratable para encontrar las estimaciones de máxima verosimilitud (MLEs) por maximización directa. La función de verosimilitud (6.31) se puede escribir como

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\theta}) &= \int_{\mathbb{R}^n} f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}) f(\boldsymbol{\varepsilon}_t | \boldsymbol{\theta}) d\boldsymbol{\varepsilon}_t \\ &= \int_{\mathbb{R}^n} \frac{f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}) f(\boldsymbol{\varepsilon}_t | \mathbf{X}; \boldsymbol{\theta})}{\tilde{f}(\mathbf{y}_t, \boldsymbol{\varepsilon}_t)} \tilde{f}(\mathbf{y}_t, \boldsymbol{\varepsilon}_t) d\boldsymbol{\varepsilon}_t \\ &\propto \int_{\mathbb{R}^n} \frac{f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}) f(\boldsymbol{\varepsilon}_t | \boldsymbol{\theta})}{f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}_0) \tilde{f}(\boldsymbol{\varepsilon}_t)} \tilde{f}(\boldsymbol{\varepsilon}_t | \mathbf{y}_t) d\boldsymbol{\varepsilon}_t \\ &= \tilde{\mathbb{E}} \left[ \frac{f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}) f(\boldsymbol{\varepsilon}_t | \boldsymbol{\theta})}{f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}_0) \tilde{f}(\boldsymbol{\varepsilon}_t)} \middle| \mathbf{y}_t \right] \end{aligned} \quad (6.33)$$

donde  $\tilde{f}(\mathbf{y}_t, \boldsymbol{\varepsilon}_t) = f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t; \boldsymbol{\alpha}_0) \tilde{f}(\boldsymbol{\varepsilon}_t)$  y  $\tilde{f}(\boldsymbol{\varepsilon}_t)$  es alguna función de densidad con soporte en  $\mathbb{R}^n$ , la función de densidad condicional  $\tilde{f}(\boldsymbol{\varepsilon}_t | \mathbf{y}_t) \propto f(\mathbf{y}_t | \boldsymbol{\varepsilon}_t) \tilde{f}(\boldsymbol{\varepsilon}_t)$ , y  $\tilde{\mathbb{E}}(\cdot | \mathbf{y}_t)$  denota la esperanza con respecto a  $\tilde{f}(\cdot | \mathbf{y}_t)$  y depende de una estimación inicial de  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\alpha}_0$ . Los MLEs se pueden calcular por maximización vía la aproximación de Monte Carlo presentada en (6.33),

$$L_r(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \frac{1}{r} \sum_{j=1}^r \frac{f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t(j); \boldsymbol{\alpha}) f(\boldsymbol{\varepsilon}_t(j) | \boldsymbol{\theta})}{f(\mathbf{y}_t | \mathbf{X}, \boldsymbol{\varepsilon}_t(j); \boldsymbol{\alpha}_0) \tilde{f}(\boldsymbol{\varepsilon}_t(j))} \quad (6.34)$$

donde los  $\varepsilon_t(j)$ 's ( $j = 1, \dots, r$ ) son muestreados por MCMC de la distribución  $\tilde{f}(\cdot | \mathbf{y}_t)$ . Como se observó en (6.33) se podría escoger  $\tilde{f}(\cdot)$  cercano a  $f(\cdot | \hat{\boldsymbol{\theta}})$ , donde  $\hat{\boldsymbol{\theta}}$  es el MLE de  $\boldsymbol{\theta}$ , ya que de lo contrario uno o muy pocos de los términos  $f(\varepsilon_t(j) | \boldsymbol{\alpha}, \boldsymbol{\theta}) / \tilde{f}(\varepsilon_t(j))$ ,  $j = 1, \dots, r$  puede dominar a los otros en  $L_r(\boldsymbol{\alpha}, \boldsymbol{\theta})$ , lo que hace la aproximación menos útil.

A continuación se presenta un procedimiento numérico para maximizar la aproximación de Monte Carlo (6.34). Sea  $(\boldsymbol{\alpha}, \boldsymbol{\theta}) = (\boldsymbol{\alpha}, \sigma^2, \varsigma_1, \dots, \varsigma_{q_2}, \varphi_0, \dots, \varphi_{q_2})$ , la maximización de  $L_r$  con respecto a  $\boldsymbol{\alpha}$  y  $\sigma^2$  dado  $\boldsymbol{\vartheta} = (\varsigma_1, \dots, \varsigma_{q_2}, \varphi_0, \dots, \varphi_{q_2})$  es bastante sencilla, ya que la primera y segunda derivadas de la función de densidad normal  $f(\boldsymbol{\eta}_t(j) | \mathbf{X}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\vartheta})$ ,  $j = 1, \dots, r$ , con respecto a esos parámetros son simples, lo que hace que un procedimiento iterativo como el de Newton-Raphson sea factible y computacionalmente rápido. Para ello, un procedimiento iterativo de valores iniciales adecuados son

$$\begin{aligned}\boldsymbol{\alpha}(j) &= [\mathbf{X}^t \mathbf{Q}_\varepsilon^{-1} (\mathbf{Q}_\varepsilon^t)^{-1} \mathbf{X}]^{-1} \mathbf{X}^t \mathbf{Q}_\varepsilon^{-1} (\mathbf{Q}_\varepsilon^t)^{-1} \boldsymbol{\eta}_t(j) \\ \sigma^2(j) &= \frac{1}{n} [\boldsymbol{\eta}_j(j) - \mathbf{X}\boldsymbol{\alpha}(j)]^t \mathbf{Q}_\varepsilon^{-1} (\mathbf{Q}_\varepsilon^t)^{-1} [\boldsymbol{\eta}_t(j) - \mathbf{X}\boldsymbol{\alpha}(j)]\end{aligned}$$

$j = 1, \dots, r$ , los cuales corresponden a las estimaciones de máxima verosimilitud para la función de densidad normal  $f(\boldsymbol{\eta}_t(j) | \mathbf{X}; \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\vartheta})$ .

Los valores de  $\boldsymbol{\alpha}$  y  $\sigma^2$  que maximizan  $L_r(\boldsymbol{\alpha}, \boldsymbol{\theta})$  para un valor fijo de  $\boldsymbol{\vartheta}$ ,  $\hat{\boldsymbol{\alpha}}(\boldsymbol{\vartheta})$  y  $\hat{\sigma}^2(\boldsymbol{\vartheta})$  están conectados dentro de  $L_r$ , y por consiguiente, se obtiene  $\tilde{L}_r(\boldsymbol{\vartheta}) = L_r(\hat{\boldsymbol{\alpha}}(\boldsymbol{\vartheta}), \hat{\sigma}^2(\boldsymbol{\vartheta}), \boldsymbol{\vartheta})$ . Esta función es maximizada con respecto a  $\boldsymbol{\vartheta}$  para una función de correlación dada utilizando optimización numérica. Los parámetros  $\boldsymbol{\vartheta}$  ingresan en  $\tilde{L}_r$  vía la matriz  $\mathbf{Q}_\varepsilon$ , y como, la inversa de esta matriz es computacionalmente exigente, la maximización puede ser relativamente lenta. La maximización puede también ser sensible a valores iniciales en este proceso, porque la aproximación  $L_r$  puede ser multimodal. Por tal motivo se deben explorar diferentes valores iniciales con la finalidad de garantizar que las estimaciones obtenidas sean las correctas.

## 6.4 Selección, validación y predicción del modelo ajustado

### 6.4.1 Medidas de bondad de ajuste

Después de ajustar el modelo basado en distancia, es importante llevar a cabo un análisis de diagnóstico para verificar la bondad de ajuste del modelo estimado. Una medida global de la variación explicada es obtenida mediante el

cálculo del pseudo  $R_k^2$  definido como

$$R_k^2 = \frac{l(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})}{l(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\theta}})}, \quad 0 \leq R_k^2 \leq 1$$

donde  $l(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\theta}})$  es la función log-verosimilitud para el modelo saturado evaluado en  $\tilde{\boldsymbol{\alpha}}$  y  $\tilde{\boldsymbol{\theta}}$ , y  $l(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$  denota el valor de la log-verosimilitud para el modelo de interés. Observe que  $l(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\theta}})$  será más grande que cualquier otra función de verosimilitud para esas observaciones, asumiendo la misma función de distribución y función de enlace ya que ésta provee la más completa descripción de los datos.

La discrepancia del modelo ajustado se puede determinar a través de qué tan bien el modelo ajustado es significativamente diferente del modelo saturado, el cual contiene tantos parámetros como observaciones están presentes en el modelo. Para esto, sea

$$D(\mathbf{y}_t, \boldsymbol{\alpha}, \boldsymbol{\theta}) = 2 [l(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\theta}})]$$

Una aproximación a esta cantidad se puede obtener de la manera clásica como

$$D(\mathbf{y}_t, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{i=1}^n (r_D(\mathbf{s}_i, t))^2 \quad (6.35)$$

la cual se conoce como la *deviance* y

$$r_{D_i} = r_D(\mathbf{s}_i, t) = \text{sign} [y(\mathbf{s}_i, t) - \hat{\mu}(\mathbf{s}_i, t)] \\ \times \left\{ 2 [l(y(\mathbf{s}_i, t), \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}) - l(y(\mathbf{s}_i, t), \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\theta}})] \right\}^{1/2}$$

donde  $l(y(\mathbf{s}_i, t), \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\theta}})$  es la máxima log-verosimilitud para el modelo saturado asociado a la  $i$ -ésima observación y  $l(y(\mathbf{s}_i, t), \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$  es el máximo valor de la función de log-verosimilitud para el modelo de interés asociado a la  $i$ -ésima observación.  $r_D(\mathbf{s}_i, t)$  es la  $i$ -ésima deviance residual porque una observación con un valor absoluto grande de  $r_D(\mathbf{s}_i, t)$  se puede ver como una discrepancia. Como se esperaba, la log-verosimilitud asociada con el modelo saturado debe ser más grande que la de un modelo con  $k < n$  parámetros.

Como se sabe, el objetivo del análisis residual es identificar observaciones atípicas y/o una mala especificación del modelo. Para investigar esto, se pueden utilizar los residuos ordinarios o los deviance residuales. De este modo, los residuales son medidas de concordancia entre los datos y el modelo ajustado. La mayoría de los residuales se basan en las diferencias entre la respuesta observada y la media condicional ajustada. Así, por ejemplo los residuales de Pearson están dados por

$$r_{P_i} = r_P(\mathbf{s}_i, t) = \frac{y(\mathbf{s}_i, t) - \hat{\mu}(\mathbf{s}_i, t)}{\sqrt{v(\hat{\mu}(\mathbf{s}_i, t))}}$$



donde  $v(\hat{\mu}(\mathbf{s}_i, t))$  es alguna de las funciones de varianza presentadas en la Tabla 5.1.

Para evitar el problema de tener un pseudo  $R_k^2 \simeq 1$  con el rango de  $\mathbf{X}$  igual a  $k = n - 1$ , es necesario considerar únicamente los vectores propios más correlacionados de  $\mathbf{B}$ , dados por  $\mathbf{1}, \mathbf{X}_1, \dots, \mathbf{X}_k$ , con la variable regionalizada temporalmente  $\mathbf{y}_t$ , es decir, las coordenadas principales más significativas correlacionadas con  $\mathbf{y}_t$ .

### 6.4.2 Selección de las coordenadas principales para el DBSGLMVAR reducido

Inicialmente, el número de variables explicativas se puede elegir como  $k$ . Una buena aproximación para la selección de columnas de  $\mathbf{X}$  consiste en clasificarlos en función de su pseudo coeficiente de correlación con respecto a  $\mathbf{y}_s$ , es decir,

$$R_1^2(\mathbf{X}_1) > \dots > R_1^2(\mathbf{X}_k) > \dots > R_1^2(\mathbf{X}_{n-1})$$

donde  $R_1^2(\mathbf{X}_j)$  es el coeficiente de pseudo correlación entre el  $\mathbf{X}_j$ -ésima coordenada principal ( $j = 1, \dots, k, \dots, n - 1$ ) y  $\mathbf{y}_t$ . Esto se hace dejando la misma función de enlace  $g$  en los diferentes modelos ajustados. Con este procedimiento, las variables menos correlacionadas con  $\mathbf{y}_t$  en la matriz de coordenadas principales  $\mathbf{X}$  se eliminan, es decir,  $n - k - 1$  coordenadas principales no son consideradas en el modelo final.

Un procedimiento similar se obtiene utilizando (6.35), pero con  $r_i(\mathbf{X}_j)$ , que es el residual obtenido a partir de la inclusión en el modelo de  $\mathbf{X}_j$ ,  $j = 1, \dots, k, \dots, n - 1$ . De esta manera, las pseudo-deviances son ordenadas de menor a mayor; los primeros  $k$  pseudo-deviances son entonces

$$D(\mathbf{y}_t; \mathbf{X}_1) < D(\mathbf{y}_t; \mathbf{X}_2) < \dots < D(\mathbf{y}_t; \mathbf{X}_k)$$

Una vez elegidas las coordenadas principales significativas ( $k$ ), se pueden discutir las técnicas espacio-temporales para predecir el valor de un campo aleatorio en una ubicación dada de observaciones cercanas.

### 6.4.3 Predicción espacial de un nuevo individuo

Las coordenadas principales  $\mathbf{x}_{(k)}(\mathbf{s}_0)$  se obtienen asumiendo que las observaciones de las variables explicativas están disponibles para un nuevo individuo, esto es,  $\mathbf{v}(\mathbf{s}_0) = (v_1(\mathbf{s}_0), \dots, v_p(\mathbf{s}_0))^t$  es conocido. Entonces, las distancias entre el nuevo individuo y cada uno de los individuos involucrados en el modelo propuesto en (6.1) se deben calcular, es decir,  $d_{0i} = d(v(\mathbf{s}_0), v(\mathbf{s}_i))$ ,  $i = 1, \dots, n$ .

A partir de estas distancias una predicción se puede realizar utilizando un resultado propuesto por Gower (1968) y Cuadras & Arenas (1990, Section 3.3), que relaciona el vector  $\mathbf{d}_0 = (d_{01}^2, \dots, d_{0n}^2)^t$  de distancias al cuadrado y el vector  $\mathbf{x}_{(k)}(\mathbf{s}_0) = (x_1(\mathbf{s}_0), \dots, x_k(\mathbf{s}_0))^t$  de coordenadas principales asociadas al nuevo individuo como sigue

$$d_{0i}^2 = (\mathbf{x}_{(k)}(\mathbf{s}_0) - \mathbf{x}_{(k)}(\mathbf{s}_i))^t (\mathbf{x}_{(k)}(\mathbf{s}_0) - \mathbf{x}_{(k)}(\mathbf{s}_i))$$

donde  $\mathbf{x}_{(k)}(\mathbf{s}_i) = (x_1(\mathbf{s}_i), \dots, x_k(\mathbf{s}_i))^t$  con  $i = 1, \dots, n$ . Entonces, se tiene que

$$\mathbf{x}_{(k)}(\mathbf{s}_0) = \frac{1}{2} \mathbf{\Lambda}_{(k)}^{-1} \mathbf{X}_{(k)}^t (\mathbf{b} - \mathbf{d}_0) \quad (6.36)$$

donde  $\mathbf{\Lambda}_{(k)}$  es una matriz diagonal con  $k$  valores propios asociados a los  $k$  vectores propios  $\mathbf{X}_{(k)}$  seleccionados anteriormente,  $\mathbf{b} = (b_{11}, \dots, b_{nn})^t$  es un vector formado por los elementos de la diagonal de  $\mathbf{B}$ , con  $b_{ii} = \mathbf{x}_{(k)}^t(\mathbf{s}_i) \mathbf{x}_{(k)}(\mathbf{s}_i)$ ,  $i = 1, \dots, n$ .

### Predicción espacial

Específicamente, se trata de interpolar el valor  $\mathbf{y}_0 = (y(\mathbf{s}_{n+1}, t), \dots, y(\mathbf{s}_{n+n'}, t))^t$  de un campo aleatorio  $\mathbf{Y}_0$  a partir de las observaciones  $y_{it} = y(\mathbf{s}_i, t)$ ,  $i = 1, \dots, n$  y  $t = 1, \dots, T$  en  $l$  ubicaciones predefinidas donde  $l = 1, \dots, n'$ . Así, el interés se centra en la interpolación de efectos aleatorios sobre un área espacial continua cuando las observaciones son no gaussianas. Por lo tanto, sea  $\boldsymbol{\eta}_t^0 = (\eta(\mathbf{s}_{n+1}, t), \dots, \eta(\mathbf{s}_{n+n'}, t))^t$  la función de predicción y sea  $f(\boldsymbol{\eta}_t^0, \boldsymbol{\eta}_t)$  la función de densidad conjunta de  $\boldsymbol{\eta}_t$  y el vector  $\boldsymbol{\eta}_t^0$ . Limitando el interés a pseudo predictores lineales insesgados de la forma

$$\tilde{\boldsymbol{\eta}}_t = \mathbf{h} + \mathbf{Q}\boldsymbol{\eta}_t$$

para algún vector conformable  $\mathbf{h}$  y una matriz  $\mathbf{Q}$  (McCulloch et al. 2008). Por lo consiguiente, minimizando el error cuadrático medio de la predicción, se encuentra que el mejor pseudo predictor lineal insesgado está dado por

$$\tilde{\boldsymbol{\eta}}_t = \mathbf{X}^0 \boldsymbol{\alpha} + \text{Cov}^t(\boldsymbol{\eta}_t, \boldsymbol{\eta}_t^0) \boldsymbol{\Sigma}_{\boldsymbol{\eta}_t}^{-1} [\boldsymbol{\eta}_t - \mathbf{X}\boldsymbol{\alpha}]$$

donde  $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_t} = \mathbf{Q}_\varepsilon \mathbf{Q}_\varepsilon^t \sigma^2$ ,  $\mathbf{X}^0$  es una matriz de  $k$  coordenadas principales para los nuevos  $n'$  sujetos espaciales incluyendo un vector de 1's, es decir,  $\mathbf{1}_{n'}$  es de tamaño  $n' \times 1$ .

La matriz de covarianza para la predicción tiene la siguiente forma general

$$\begin{aligned} \text{Var}(\tilde{\boldsymbol{\eta}}_t \mid \mathbf{y}_t, \dots, \mathbf{y}_{t-q_1}) &\approx \boldsymbol{\Sigma}_0 + \text{Cov}^t(\boldsymbol{\eta}_t, \boldsymbol{\eta}_t^0) \boldsymbol{\Sigma}_{\boldsymbol{\eta}_t}^{-1} [\widetilde{\text{Var}}_r(\boldsymbol{\eta}_t \mid \mathbf{y}_t, \dots, \mathbf{y}_{t-q_1})] \\ &\quad \times \boldsymbol{\Sigma}_{\boldsymbol{\eta}_t}^{-1} \text{Cov}(\boldsymbol{\eta}_t, \boldsymbol{\eta}_t^0) \end{aligned}$$

donde  $\boldsymbol{\Sigma}_0 = \text{Var}(\boldsymbol{\eta}_t^0) - \text{Cov}^t(\boldsymbol{\eta}_t, \boldsymbol{\eta}_t^0) \boldsymbol{\Sigma}_{\boldsymbol{\eta}_t}^{-1} \text{Cov}(\boldsymbol{\eta}_t, \boldsymbol{\eta}_t^0)$  y  $\widetilde{\text{Var}}_r$  es la matriz de covarianza basada en la muestra  $\boldsymbol{\eta}_t(1), \dots, \boldsymbol{\eta}_t(r)$ .



## Capítulo 7

# Conclusiones y recomendaciones

En el Capítulo 2 se propuso una metodología DB para ajustar variables respuesta de tipo beta con dispersión contante y variable. Se desarrolló la estimación de parámetros por máxima verosimilitud, diagnósticos y predicciones; se realizó una comparación con el modelo de regresión beta clásico. Además, cuando el número de coordenadas principales creció, el modelo DBBR mostró mejores resultados en las predicciones que el método de regresión beta tradicional. Sin embargo, si el número de coordenadas principales era igual al número de variables explicativas, los resultados obtenidos fueron similares a los obtenidos con el procedimiento de regresión beta clásico propuesto por Ferrari & Cribari-Neto (2004). Cuando en el modelo propuesto hay multicolinealidad, el método DB se recomienda porque cada una de las coordenadas principales son independientes, lo que no ocurre en la regresión beta clásica con las variables originales.

En las dos aplicaciones presentadas en el Capítulo 2, se ajustó el modelo DBBR con dispersión variable utilizando coordenadas principales, las cuales fueron obtenidas por medio del uso de la distancia de Gower en el caso de los fondos de inversión tanto para el modelo de media como para el modelo de dispersión variable. Sin embargo, en el porcentaje de gasolina, en el modelo de media se utilizó la distancia de Gower y la distancia euclidiana clásica en el modelo de dispersión variable. Estos resultados muestran que cualquier tipo de distancia se puede utilizar, y que no necesariamente, se debe utilizar el mismo tipo de distancia en las dos componentes (modelo de media y modelo de dispersión variable) del modelo DBBR.

El modelo DBBR presentado en (2.6) no es un modelo de regresión beta desde un punto de vista clásico porque la relación entre la variable respuesta con las variables explicativas no se establece bajo este modelo. De esta manera, el modelo DBBR debe ser considerado como un modelo de predicción bastante flexible, ya que sólo se necesita elegir una adecuada distancia. Además, se

puede hacer frente al problema de datos faltantes.

Por otro lado, se ha dado cierta justificación a los métodos propuestos para elegir las dimensiones de predicción adecuadas. Los métodos que se presentaron en la subsección 2.3.2 pueden resolver este problema cuando la variable respuesta se explica mejor por medio de las coordenadas principales con mayor varianza o significancia, tanto para el modelo de media como para el modelo de dispersión variable.

Además, ya que la dimensión aumenta con  $n$ , potencialmente se podría trabajar con muchas covariables. Por lo tanto, se pueden encontrar valores propios muy pequeños, y así, la predicción para un nuevo individuo tanto para el modelo de media como para el modelo de dispersión variable (2.6) podría ser incorrecto, debido a que se deben calcular  $\mathbf{\Lambda}_x^{-1}$  y  $\mathbf{\Lambda}_z^{-1}$ , respectivamente. Entonces, hay dos maneras alternativas para solucionar el anterior problema: o bien eliminar covariables con varianzas pequeñas, o utilizar (2.38), seleccionando las más pequeñas  $\kappa_v$  y  $\kappa_u$  tal que  $(\mathbf{\Lambda}_x + \kappa_v \mathbf{I})^{-1}$  y  $(\mathbf{\Lambda}_z + \kappa_u \mathbf{I})^{-1}$  se puedan calcular con bastante precisión.

El DBSGLMM propuesto utilizando máxima verosimilitud en el MCMC provee una herramienta útil para la selección del modelo en modelos espaciales complicados. Sin embargo, es bien conocido que el algoritmo MCMC en modelos espaciales (basado en campos aleatorios) puede ser muy lento en su convergencia. Esto sin duda sucede en el DBSGLMM, donde una tasa de reducción de 1 en 1000 se aplica a la salida MCMC. En el modelo ajustado, la inferencia se puede realizar con el objetivo de tener un mejor desempeño en las predicciones en puntos no observados. La técnica presentada en esta tesis, se puede también utilizar en otro tipo de modelos geoestadísticos, mas allá de los modelos lineales generalizados espaciales mixtos. Además, incrementando el número de coordenadas principales, el método propuesto produce mejores estimaciones de las variables regionalizadas (ver Sección 3.4). Mientras que si se tiene un número de coordenadas principales igual al número de variables explicativas originales, los resultados obtenidos son similares a los métodos tradicionales.

En el Capítulo 3, se empleo la predicción DB propuesta por Cuadras (1989), Cuadras & Arenas (1990) y Cuadras et al. (1996), en la cual el conjunto de predictores euclidianos disponibles se enmarcan dentro de los obtenidos por coordenadas principales, lo que es equivalente a hacer un escalamiento métrico multidimensional clásico. Boj et al. (2010) clarificaron el papel central que juega el espacio por columnas en las matrices de distancias al cuadrado doblemente centradas. Así desde una perspectiva metodológica, las predicciones presentadas al utilizar el DBSGLMM son intrínsecas porque éstas dependen de las disimilaridades propias más que de una base específica dada. La invariancia permite el uso de otras posibles bases equivalentes que podrían ser numérica-

mente más estables que la solución de coordenadas principales presentada en este trabajo. Por consiguiente, un trabajo como el Boj et al. (2010) podría ser implementado en el DBSGLMM.

La metodología propuesta tiene amplia aplicación en modelos de riesgos ambientales de una enfermedad. En general, cuando hay escasez de datos de campo e incertidumbre inherente en el modelo de entrada, las decisiones tomadas mediante un modelo necesitan incorporar estas incertidumbres en la estructura del modelo. Con respecto a la aplicación presentada en el Capítulo 3, se estudió la variación en la prevalencia de *Loa loa* microfilaria, y además, se consideró la variación no espacial que no se puede atribuir a la distribución del error Binomial, pero este último no fue significativo. Sin embargo, el verdor de la vegetación circundante y la elevación o altura del terreno parecen notablemente afectar la prevalencia de *Loa Loa* en Camerún.

El BSLMM con dispersión variable propuesto utilizando MCMC es útil en situaciones donde la variable respuesta espacial es una razón o una proporción. En el BSLMM propuesto se construyeron dos modelos: SMM y SVDM para ajustar los parámetros involucrados en la función de distribución beta. En los dos modelos se utilizaron variables explicativas para modelar la tendencia y la correlación espacial. Además, se presentaron algunas medidas aproximadas, para hacer inferencia y diagnóstico sobre los parámetros del BSLMM propuesto.

En la aplicación del contenido de arcilla presentada en el Capítulo 4, se encontró que la correlación espacial en el SMM no fue significativa; sin embargo, la correlación espacial sin efecto pepita en el SVDM si lo fue. Esto significa que la precisión no es constante sobre el espacio y es necesaria para mejorar el ajuste del BSLMM. Mientras en la aplicación del contenido de magnesio presentada en el Capítulo 4, se encontró que en ambos modelos, la correlación espacial sin efecto pepita era significativa. La precisión fue alta en ubicaciones con contenido alto de magnesio y baja en ubicaciones con contenido de magnesio ligeramente bajo. En esta última aplicación, los mapas para la media y la dispersión variable se pudieron construir porque se tenía información de las variables explicativas en ubicaciones donde la variable respuesta no fue observada, lo cual no sucede en la primera aplicación de ese capítulo. Este hecho muestra que se puede utilizar el BSLMM para hacer inferencia sobre el efecto de las variables explicativas espaciales en la variable respuesta y para encontrar áreas de alta o baja presencia con sus respectivas precisiones.

De otro lado, aunque el enfoque bayesiano utilizado por Diggle et al. (1998) y Christensen & Waagepetersen (2002) provee una manera natural de incorporar parámetros de incertidumbre en la inferencia predictiva, a-prioris significativas sobre los parámetros estructurales tales como el tipo de correlación y la elección de la función de enlace son en muchos casos muy laboriosos de

construir. Además, no siempre se puede realizar apropiadamente un algoritmo MCMC con actualizaciones de tales parámetros (Christensen 2004); sin embargo, esto se puede investigar en estudios futuros. Las a-prioris informativas en la práctica son difíciles de obtener y no hay un consenso sobre cómo construir a-prioris informativas de referencia para estos modelos.

En el Capítulo 5, se presentaron los modelos lineales generalizados autorregresivos espacio-tiempo basados en distancias con perturbaciones autorregresivas espacio-tiempo. Para este modelo, se desarrollaron la estimación de los efectos fijos y aleatorios y se determinaron sus niveles de significancia. Se amplió la posibilidad de juzgar la especificación de efectos fijos contra la especificación de efectos aleatorios en modelos de datos panel, para incluir la autocorrelación del error espacio-tiempo o una variable dependiente rezagada espacio-tiempo utilizando pruebas de especificación. Se estimó la matriz de varianzas y covarianzas de los parámetros en estos modelos extendidos; el proceso de estimación de los diferentes parámetros se realizó mediante una adaptación del método de ecuaciones de estimación generalizada para espacio-tiempo. Se presentaron dos opciones adicionales de estimación que se pueden emplear: máxima verosimilitud y el método MCMC obtenido mediante máxima verosimilitud. Además, se presentó la selección, validación y predicción del modelo ajustado utilizando el método GEE para espacio-tiempo; se expuso una medida de bondad de ajuste, se dieron algunas medidas para realizar el análisis de residuos, se hizo el proceso de selección de las coordenadas principales y se realizó la predicción espacio-tiempo de un nuevo sujeto.

En la aplicación presentada en el Capítulo 5 sobre el número de acciones armadas de los grupos guerrilleros de las FARC-EP y el ELN en Colombia, se ilustra claramente la metodología planteada para espacio-tiempo y se lleva a cabo un proceso de diagnóstico y diagnóstico para esta clase de modelos que no es presentada habitualmente por tratarse de datos en espacio-tiempo.

En el Capítulo 6 se presentó un modelo autorregresivo espacial lineal generalizado mixto utilizando el método basado en distancias. Este modelo incluyó retrasos tanto espaciales como temporales entre vectores de variables de estado estacionarias. Aunque los parámetros estructurales no se pueden en algunos casos identificar completamente en este modelo, los coeficientes de rezago espacial contemporáneos se pueden identificar mediante las variables explicativas de estado. Se utilizó la dinámica espacial de los datos econométricos tipo panel para estimar el modelo propuesto, es así como los parámetros involucrados en el modelo se estimaron utilizando el método MCMC mediante máxima verosimilitud. Además, se discutió en este capítulo la interacción entre estacionariedad temporal y espacial, y se derivaron las respuestas al impulso para el modelo propuesto, lo cual naturalmente depende de la dinámica temporal y espacial del modelo.

Aunque algunas de las metodologías propuestas en esta investigación trabajan relativamente rápido, la complejidad computacional sigue siendo un problema al utilizar en especial el método MCMC en modelos mixtos lineales generalizados basados en distancias. Es por ello que algunos otros de los métodos propuestos en esta tesis, demandan mucho tiempo y pueden demandar demasiadas horas para generar las predicciones. Esto indica que todavía hay oportunidad para mejorar el procesamiento de datos en los métodos propuestos.

Desde el punto de vista teórico, parece interesante utilizar funcionales tipo kernel, spline o funciones de base radial espaciales o espacio-temporales en modelos espaciales o espacio-temporales con respuesta no normal; comprobar su bondad de ajuste con simulaciones y datos reales, para valorar sus ventajas o desventajas con respecto a los métodos propuestos en esta tesis. Esta temática no se trabajó en esta tesis, ya que la idea era dar un primer paso al utilizar la combinación de metodologías clásicas y entendibles por una gran mayoría de los usuarios en espacio y en espacio-tiempo con variables respuestas no normales.





# Referencias

- Albert, P. S. & McShane, L. M. (1995), 'A generalized estimating equations approach for spatially correlated binary data: with an application to the analysis of neuroimaging data', *Biometrics* **51**, 627–638.
- Anderson, D. R., Sweeney, D. J. & Williams, T. A. (2011), *Statistics for Business and Economics*, South-Western, Cengage Learning, Mason, OH.
- Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic, Dordrecht.
- Anselin, L. & Florax, R. (1995), *New Directions in Spatial Econometrics*, Springer-Verlag, Berlin.
- Anselin, L., Florax, R. J. G. M. & Rey, S. (2004), *Advances in Spatial Econometrics, Methodology, Tools and Applications*, Springer, Berlin.
- Anselin, L. & Kelejian, H. H. (1997), 'Testing for spatial error autocorrelation in the presence of endogenous regressors', *International Regional Science Review* **20**(1-2), 153–182.
- Aranda-Ordaz, F. J. (1981), 'On two families of transformations to additivity for binary response data', *Biometrika* **68**, 357–363.
- Arenas, C. & Cuadras, C. (2002), 'Recent statistical methods based on distances', *Contributions to Science, Institut d'Estudis Catalans Barcelona* **2**(2), 183–191.
- Atkinson, A. (1985), *Plots, Transformations and regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Clarendon Press, Oxford.
- Augustin, N. H., Kublin, E., Metzler, B., Meierjohann, E. & von Wuhlisch, G. (2002), 'Analyzing the spread of beech canker', *Forest Science* **51**(5), 438–448.

- Boj, E., Delicado, P. & Fortiana, J. (2010), 'Distance-based local linear regression for functional predictors', *Computational Statistics and Data Analysis* **54**(2), 429–437.
- Botter, D. A. & Cordeiro, G. M. (1997), 'Bartlett corrections for generalized linear models with dispersion covariates', *Communication in Statistics - Theory Methods* **26**, 279–307.
- Bouma, J., Stoorvogel, J., Van Alphen, B. J. & Booltink, H. W. G. (1999), 'Pedology, precision agriculture, and the changing paradigm of agricultural research', *Soil Science Society of America Journal* **63**, 1763–1768.
- Boussinesq, M., Gardon, J., Kamgno, J., Pion, S. D., Gardon-Wendel, N. & Chippaux, J. P. (2001), 'Relationships between the prevalence and intensity of *Loa loa* infection in the Central province of Cameroon', *Annals of Tropical Medicine and Parasitology* **95**, 495–507.
- Box, G. E. & Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society Series B* **26**, 211–246.
- Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**, 9–25.
- Brunsdon, C. F., Fotheringham, A. S. & Charlton, M. E. (1996), 'Geographically weighted regression: a method for exploring spatial nonstationarity', *Geographical Analysis* **28**(4), 281–298.
- Brunsdon, C., Fotheringham, S. & Charlton, M. (1998), 'Geographically weighted regression-modelling spatial non-stationarity', *The Statistician* **47**, 431–443.
- Bury, K. (1999), *Statistical Distributions in Engineering*, Cambridge University Press, New York.
- Capeche, C. L., Macedo, J. R., Manzatto, H. R. H. & Silva, E. F. (1997), Caracterização pedológica da fazenda angra - pesagro/rio, in 'Informação, globalização, uso do solo', Estação experimental de Campos, Rio de Janeiro.
- Cepeda-Cuervo, E. & Achcar, J. A. (2010), 'Heteroscedastic nonlinear regression models', *Communications in Statistics - Simulation and Computation* **39**(2), 405–419.
- Chamberlain, G. (1984), *Panel Data*, In Handbook of Econometrics, Vol II, edited by Z. Griliches and M. Intriligator, pp. 1247-1318, Amsterdam.

- Christensen, O. F. (2004), 'Monte Carlo maximum likelihood in model-based geostatistics', *Journal of Computational and Graphical Statistics* **13**, 702–718.
- Christensen, O. F., Diggle, P. J. & Ribeiro, P. J. (2001), Analysing positive-valued spatial data: The transformed Gaussian model, in 'geoENV III - Geostatistics for Environmental Applications, P. Monestiez, D. Allard, and R. Froidevaux (eds)', Kluwer Academic Publishers, Boston, pp. 287–298.
- Christensen, O. & Waagepetersen, R. (2002), 'Bayesian prediction of spatial count data using generalized linear mixed models', *Biometrics* **58**, 280–286.
- Cook, D. O., Kieschnick, R. & McCullough, B. D. (2008), 'Regression analysis of proportions in finance with self selection', *Journal of Empirical Finance* **15**, 860–867.
- Cook, R. D. (1977), 'Detection of influential observations in linear regression', *Technometrics* **19**, 15–18.
- Cook, R. D. (1986), 'Assessment of local influence (with discussion)', *Journal of the Royal Statistical Society* **48**, 133–169.
- Cook, R. D. & Weisberg, S. (1983), 'Diagnostics for heteroscedasticity in regression', *Biometrika* **70**, 1–10.
- Cox, C. (1996), 'Nonlinear quasi-likelihood models: applications to continuous proportions', *Computational Statistics & Data Analysis* **21**, 449–461.
- Cressie, N. (1993), *Statistics for Spatial Data*, Revised Edition. John Wiley & Sons Inc., New York.
- 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.
- Cribari-Neto, F. & Vasconcellos, K. (2002), 'Nearly unbiased maximum likelihood estimation for the beta distribution', *Journal of Statistical Computation and Simulation* **72**, 107–118.
- Cribari-Neto, F. & Zeileis, A. (2010), 'Beta regression in R', *Journal of Statistical Software* **34**(2), 1–24.
- 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. & Fortiana, J. (1993), 'Aplicaciones de las distancias en estadística', *Qüestió* **17**, 39–74.
- Cuadras, C. M. (1989), Distance analysis in discrimination and classification using both continuous and categorical variables, in 'Recent Developments in Statistical Data Analysis and Inference', (Y. Dodge ed.). Elsevier Science Publisher, North-Holland, Amsterdam, pp. 459–474.
- Cysneiros, F. J. A., Paula, G. A. & Galea, M. (2007), 'Heteroscedastic symmetrical linear models', *Statistics & Probability Letters* **77**, 1084–1090.
- Davidian, M. (2005), *Applied Longitudinal Data Analysis*, Chapman and Hall, North Carolina State University, North Carolina.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- Dielman, T. (1983), 'Pooled cross-sectional and time series data: A survey of current statistical methodology', *The American Statistician* **37**, 111–122.
- 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.
- Diggle, P., Heagerty, P., Liang, K. Y. & Zeger, S. L. (2002), *Analysis of Longitudinal Data*, Oxford University Press, New York.
- Diggle, P. J. & Ribeiro, P. J. (2007), *Model-Based Geostatistics*, Springer Series in Statistics, New York.
- Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J. H., Boussinesq, M. & Molyneuz, D. H. (2007), 'Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty', *Annals of Tropical Medicine & Parasitology* **101**(6), 499–509.
- Diggle, P., Tawn, A. & Moyeed, R. (1998), 'Model based geostatistics', *Applied Statistics* **49**, 299–350.
- Dobson, A. J. (2002), *An introduction to generalized linear models*, 2nd ed. Chapman Hall, New York.

- Dudley, S. (2004), *Walking Ghosts: Murder and Guerrilla Politics in Colombia*, Routledge.
- Elhorst, J. P. (2003), 'Specification and estimation of spatial panel data models', *International Regional Science Review* **26**(3), 244–268.
- Elhorst, J. P. (2005), 'Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels', *Geographical Analysis* **37**(1), 85–106.
- Elhorst, J. P. (2008), 'A spatiotemporal analysis of aggregate labour force behaviour by sex and age across the european union', *Journal of Geographical Systems* **10**(2), 167–190.
- Espinheira, P. L., Ferrari, S. L. P. & Cribari-Neto, F. (2008a), 'Influence diagnostics in beta regression', *Computational Statistics & Data Analysis* **52**, 4417–4431.
- Espinheira, P. L., Ferrari, S. L. P. & Cribari-Neto, F. (2008b), 'On beta regression residuals', *Journal of Applied Statistics* **35**(4), 407–419.
- Esteve, A., Boj, E. & Fortiana, J. (2009), 'Interaction terms in distance-based regression', *Communications in Statistics - Theory and Methods* **38**(19), 3498–3509.
- Ferrari, S. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **31**, 799–815.
- Ferrari, S. L. P., Espinheira, P. L. & Cribari-Neto, F. (2011), 'Diagnostic tools in beta regression with varying dispersion', *Statistica Neerlandica* **65**.
- Fingleton, B. (2008), 'A generalized method of moments estimator for a spatial panel model with an endogenous spatial lag and spatial moving average errors', *Spatial Economic Analysis* **3**(1), 27–44.
- Fisher, M. M. & Getis, A. (2010), *Handbook of Applied Spatial Analysis*, Springer, Heidelberg.
- Fotheringham, A. S., Brunson, C. F. & Charlton, M. E. (1998), 'Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis', *Environment and Planning A* **30**, 1905–1927.
- Fotheringham, A. S. & Zhan, F. (1996), 'A comparison of three exploratory methods for cluster detection in spatial point patterns', *Geographical Analysis* **28**(3), 200–218.
- 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.
- Geyer, C. J. (1994), 'On the convergence of Monte Carlo maximum likelihood calculations', *Journal of the Royal Statistical Society, Series B* **56**, 261–274.
- Geyer, C. J. & Thompson, E. A. (1992), 'Constrained Monte Carlo maximum likelihood for dependent data (with discussion)', *Journal of the Royal Statistical Society, Series B* **54**, 657–699.
- Godwin, R. J. & Miller, P. C. H. (2003), 'A review of the technologies for mapping within-field variability', *Biosystems Engineering* **84**(4), 393–407.
- Goovaerts, P. (1998), 'Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties', *Biology and Fertility of Soils* **27**(4), 315–334.
- 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**, 857–874.
- Graybill, F. A. (1983), *Matrices with applications in statistics*, 2nd, edition, Wadsworth, Belmont, California.
- 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.
- Hengl, T., Heuvelink, G. 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), 265–277.
- Higdon, D., Swall, J. & Kern, J. (1999), Non-stationary spatial modeling, in 'Bayesian Statistics', Oxford University Press, Oxford, pp. 761–768.

- Hinde, J. & Demétrio, C. G. B. (1998), 'Overdispersion: Models and estimation', *Computational Statistics & Data Analysis* **27**(2), 151–170.
- Holland, D., Saltzman, N., Cox, L. H. & Nychka, D. (1999), Spatial prediction of sulfur dioxide in the eastern United States, *in* 'GeoENV II - Geostatistics for Environmental Applications', Kluwer Academic Publ., Dordrecht, pp. 65–76.
- Højbjerg, M. (2003), 'Profile likelihood in directed graphical models from BUGS output', *Statistics and Computing* **13**, 57–66.
- Hsiao, C. (1985), 'Benefits and limitations of panel data', *Econometric Reviews* **4**, 121–174.
- Hsiao, C. (2003), *Analysis of Panel Data*, 2nd edn, Cambridge University Press, Cambridge.
- Isaaks, E. & Srisvastava, R. (1989), *An introduction to applied geostatistics*, Oxford Univ. Press, New York.
- Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, John Wiley & Sons, New York.
- Journel, A. G. & Huijbregts, C. J. (1978), *Mining geoestistics*, Academic Press, New York.
- Jowaheer, V. & Sutradhar, B. (2002), 'Analysis of longitudinal count data with overdispersion', *Biometrika* **73**, 389–399.
- Kamgno, J., Bouchite, B., Baldet, T., Folefack, G., Godin, C. & Boussinesq, M. (1997), 'Study of the distribution of human filariasis in West province of Cameroon', *Bulletin de la Société de Pathologie Exotique* **90**, 327–330.
- Kieschnick, R. & McCullough, B. (2003), 'Regression analysis of variates observed on (0, 1): percentages, proportions, and fractions', *Statistics Model* **3**, 193–213.
- Krysicki, W. (1999), 'On some new properties of the beta distribution', *Statistics and Probability Letters* **42**, 131–137.
- Laird, J. H. & Ware, N. M. (1982), 'Random-effects models for longitudinal data', *Biometrics* **38**, 963–974.
- Lange, K. (1995a), 'A gradient algorithm locally equivalent to the EM algorithm', *Journal of the Royal Statistical Society, Series B* **57**, 425–437.
- Lange, K. (1995b), 'A quasi-newton acceleration of the EM algorithm', *Statistica Sinica* **5**, 1–18.



- Lee, L. F. (2004), ‘Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models’, *Econometrica* **72**, 1899–1925.
- Lee, Y. J. & Nelder, J. A. (2002), ‘Analysis of ulcer data using hierarchical generalized linear models’, *Journal of Quality Technology* **21**, 191–202.
- Liang, K. Y. & Zeger, S. L. (1986), ‘Longitudinal data analysis using generalized linear models’, *Biometrika* **73**, 13–22.
- Manski, C. F. (1993), ‘Identification of endogenous social effects: the reflection problem’, *Review of Economic Studies* **60**, 531–542.
- Mardia, K. V., Kent, J. T. & Bibby, J. M. (2002), *Multivariate Analysis*, Academic Press, Inc, London.
- Mardia, K. V. & Marshall, R. L. (1984), ‘Maximum likelihood estimation of models for residual covariance in spatial statistics’, *Biometrika* **71**, 135–146.
- Marschak, J. (1939), ‘On combining market and budget data in demand studies: A suggestion’, *Econometrica* **7**, 332–335.
- Martínez, F. (2008), Modelización de la Función de Covarianza en Procesos Espacio-Temporales: Análisis y Aplicaciones., PhD thesis, Universidad de Valencia-España.
- 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.
- McCullagh, P. (2008), ‘Sampling bias and logistic models (with discussion)’, *Journal of the Royal Statistical Society, B* **70**, 643–677.
- McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman Hall, London.
- McCulloch, C. E. & Searle, S. R. (2001), *Generalized linear and mixed models*, John Wiley & Sons, New York.
- McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models*, second edn, John Wiley & Sons, New Jersey.
- Molenberghs, G. & Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, Springer, New York.

- Morningstar Inc., Kinnel, R. & Berry, S. (2008), *Morningstar Funds 500: 2008*, John Wiley & Sons, New Jersey.
- Myers, R. H., Montgomery, D. C. & Vinning, G. G. (2002), *Generalized Linear Models. With Applications in Engineering and the Sciences*, John Wiley & Sons, New York.
- Nelder, J. A. & Wedderburn, R. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society, Series A* **135**, 370–384.
- Neter, J., Kutner, M. H., Nachtsheim, C. J. & Wasserman, W. (1996), 'A tutorial on generalized linear models', *Applied Linear Statistical Models*.
- Nocedal, J. & Wright, S. J. (1999), *Numerical Optimization*, Springer-Verlag, New York.
- Nychka, D. & Saltzman, N. (1998), Design of air quality networks, in 'Case Studies in Environmental Statistics', Springer Verlag, New York, pp. 51–76.
- Ospina, R., Cribari-Neto, F. & Vasconcellos, K. L. P. (2006), 'Improved point and interval estimation for a beta regression model', *Computational Statistics & Data Analysis* **51**, 960–981.
- Pace, R. K., Barry, R. & Sirmans, C. F. (1998), 'Spatial statistics and real estate', *Journal of Real Estate Finance and Economics* **17**, 5–13.
- Paelinck, J. H. P. & Klaassen, L. H. (1979), *Spatial Econometrics*, Saxon House, Farnborough.
- Paez, M., Gamerman, 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.
- Pan, W. (2001), 'Akaike's information criterion in generalized estimating equations', *Biometrics* **57**, 120–125.
- Paolina, P. (2001), 'Maximum likelihood estimation of models with beta-distributed dependent variables', *Political Analysis* **9**, 325–346.
- Papke, L. & Wooldridge, J. (1996), 'Econometric methods for fractional response variables with an application to 401(K) plan participation rates', *Journal of Applied Econometrics* **11**, 619–632.

- Pauwels, J. M., Van-Ranst, E., Verloo, M. & Mvondo-Ze, A. (1992), *Manuel de Laboratoire de Pédologie. Méthodes d'Analyses de Sols et de Plantes, Equipement, Gestion de stocks de Verrerie et de Produits chimiques*, number 28, Publications Agricoles, Bruxelles.
- Perčec-Tadić, M. (2010), 'Gridded croatian climatology for 1961-1990', *Theoretical and Applied Climatology* pp. 1434–4483.
- Prater, N. H. (1956), 'Estimate gasoline yields from crudes', *Petroleum Refiner* **35**, 236–238.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992), *Numerical Recipes in C: The Art of Scientific Computing*, Second ed. Cambridge University Press, New York.
- R Development Core Team (2013), *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 know variance', *Ecology* **68**(3), 744–748.
- Rocke, D. M. (1993), 'On the beta transformation family', *Technometrics* **35**(1), 72–81.
- Royle, J. A. & Wikle, C. K. (2005), 'Efficient statistical mapping of avian count data', *Environmental and Ecological Statistics* **12**(2), 225–243.
- Samper, F. & Carrera, J. (1993), *Geoestadística. Aplicaciones a la hidrogeología subterránea*, Technical report, Centro Internacional de Métodos Numéricos en Ingeniería, UPC Barcelona.
- Sampson, P. D. & Guttorp, P. (1992), 'Nonparametric estimation of nonstationary spatial covariance structure', *Journal of American Statistical Association* **87**, 108–119.
- Schall, R. (1991), 'Estimation in generalized linear models with random effects', *Biometrika* **78**(4), 719–727.
- Searle, S. R., Casella, G. & McCulloch, C. E. (1987), *Variance Components*, John Wiley & Sons, New York.
- Shumway, R. H. & Stoffer, D. S. (2000), *Time Series Analysis and Its Applications*, Springer-Verlag, New York.
- Simas, A. B., Barreto-Souza, W. & Rocha, A. V. (2010), 'Improved estimators for a general class of beta regression models', *Computational Statistics & Data Analysis* **54**(2), 348–366.

- Sipri (2012), Military expenditure by country, in constant (2011) us, 1988-2012. Stockholm International Peace Research Institute.
- Smith, D. M. & Ridout, M. S. (2003), 'Optimal designs for criteria involving log(potency) in comparative binary bioassays', *Journal Statistics Planning Inference* **113**, 617–632.
- Smithson, M. & Verkuilen, J. (2006), 'A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables', *Psychological Methods* **11**(1), 54–71.
- Stein, A. & Ettema, C. (2003), 'An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons', *Agriculture, Ecosystems & Environment* **94**(1), 31–47.
- Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer-Verlag, New York.
- Steinsland, I. (2007), 'Parallel exact sampling and evaluation of Gaussian Markov random fields', *Computational statistics & Data Analysis* **51**, 2969–2981.
- Stiratelli, R., Laird, N. & Ware, J. H. (1984), 'Random-effects models for serial observations with binary response', *Biometrics* **40**(4), 961–971.
- Takougang, I., Meremikwu, M., Wanji, S., Yenshu, E. V., Aripko, B., Lam-lenn, S. B., Eka, B. L., Enyong, P., Meli, J., Kale, O. & Remme, J. H. (2002), 'Rapid assessment method for prevalence and intensity of Loa loa infection', *Bulletin of the World Health Organization* **80**, 852–858.
- Thomson, M. C., Obsomer, V., Kamgno, J., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Remme, J. H., Molyneux, D. H. & Boussinesq, M. (2004), 'Mapping the distribution of Loa loa in Cameroon in support of the African Programme for Onchocerciasis Control', *Filaria Journal* pp. 3–7.
- Thomson, M., Connor, S., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M. & Greenwood, B. (1999), 'Predicting malaria infection in Gambian children from satellite data and bednet use surveys: the importance of spatial correlation in the interpretation of results', *American Journal of Tropical Medicine and Hygiene* **61**, 2–8.
- Vasconcellos, K. L. P. & Cribari-Neto, F. (2005), 'Improved maximum likelihood estimation in a new class of beta regression models', *Brazilian Journal of Probability and Statistics* **19**, 13–31.
- Wackernagel, H. (2003), *Multivariate Geostatistics: An introduction with applications*, Third Completely Revised Edition. Springer-Verlag, New York.

- Wanji, S., Tendongfor, N., Esum, M., Atanga, S. N. & Enyong, P. (2003), 'Heterogeneity in the prevalence and intensity of loiasis in five contrasting bioecological zones in Cameroon', *Transactions of the Royal Society of Tropical Medicine and Hygiene* **97**, 182–187.
- Yavuz, H. & Erdoğan, S. (2012), 'Spatial analysis of monthly and annual precipitation trends in Turkey', *Water Resources Management* **26**, 609–621.
- Yemefack, M., Rossiter, D. G. & Njomgang, R. (2005), 'Multi-scale characterization of soil variability within an agricultural landscape mosaic system in southern cameroon', *Geoderma* **125**(1-2), 117–143.
- Zhang, H. (2002), 'On estimation and prediction for spatial generalized linear mixed models', *Biometrics* **58**, 129–136.
- Zimmerman, D. L. & Zimmerman, M. B. (1991), 'A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors', **23**, 77–91.